CRUDE OIL PRICE FORECASTING

1.Introduction

As one of the most important commodities in the global market, crude oil plays an essential role in society, including the economic, political and technological dimensions. Forecasting crude oil prices and the volatility of this market facilitates governmental policy making, as well as generating wealth and reducing financial losses across the business sector. However, the non-linearity of crude oil markets creates difficulties in predicting market movements. The sources of crude oil price forecasting errors include complex supply–demand structures as Numerous unpredictable factors which disrupt the equilibrium of markets (Déesa, Karadelogloua, Kaufmann, & Sáncheza, 2007).

The extensive body of past research has focused predominantly on quantitative analytical approaches. This may be due in part to the availability of historic market data and new advances in statistical methods or intelligent modelling. The majority of such studies have relied heavily on the release of official macroeconomic statistics, which are collected, analyzed and aggregated by regulatory bodies. These economic indicators are usually released at regular intervals and are therefore insensitive to real-time economic issues. More importantly, qualitative analyses of politics, catastrophes and emergent events have not been considered within the realm of quantified evidence until quite recently. Improvements in crude oil forecasting accuracy require new data sources, combined with high frequency and lag-free data, in order to develop more sophisticated forecasting technologies.

Advancing internet technologies and new big datasets provide new opportunities for incorporating previously unexplored sources into modern forecasting models. Information on the internet is released more frequently than official statistics. Furthermore, media messages, the number of which is increasing constantly, contain an abundance of untapped qualitative information which may be used to enhance the prediction of crude oil prices. Specifically, online news is a better information source than general discussions in other social media (blogs, forums, etc.), as there is less noise and the context is more convincing. Thus, this study focuses on online news sources as related and influential qualitative data sources.

Incorporating qualitative data into crude oil forecasting seems promising; however, it comes with an increased complexity. Qualitative data are challenging to use primarily because they are in juxtaposition to quantitative data which are gathered and therefore structured around the knowledge of mechanisms. Text mining techniques are useful for identifying opinions and extracting information. This study employs text mining methods of text classification, sentiment analysis and topic models for transforming the unstructured text into a representative format that is structured and can be processed by the machine.

Recent years have seen the development of deep learning techniques, which have gained the attention of the academic community due to successful applications within the artificial intelligence industry. Even though deep learning techniques are applied commonly in the field of computer vision (Krizhevsky, Sutskever, & Hinton, 2012), speech recognition (Hinton et al., 2012) or computational linguistics, the literature is remarkably silent about the financial domain. This study provides a feasible approach for the application of deep learning techniques to financial markets. Specifically, we use a CNN model to learn hidden patterns embedded in crude oil news, and to predict the price movements (up or down) of the crude oil market. Here, we use the term ‘‘text features’’ to indicate quantitative variables extracted from news texts, including sentiment features, topic features and Convolutional Neural Network (CNN) features, in accordance with the specific method of text mining.

Furthermore, this study uses the Latent Dirichlet Allocation (LDA) topic model to identify latent topics of online news. After an initial investigation of our news data, we find that crude oil news can generally be classified under a few topics, such as ‘‘oil inventory’’, ‘‘stock market’’, ‘‘foreign exchange market’’, ‘‘political events’’, etc. Each of these topics is related to a factor that affects the oil price. Different topics in a single transaction day are likely to have different impacts on the oil price. It is necessary to subgroup text features according to different topics for each transaction day. This study further proposes a feature grouping method based on LDA topic models, which generates sentimenttopic features and CNN-topic features for oil price forecasting.

Although qualitative information on the market is essential, statistical data on financial markets still provide explicit information on fundamental factors of supply, demand or speculations. Thus, the qualitative information embedded in news text should be integrated with statistical data in order to achieve the best performance for oil price forecasting. Our study utilizes multiple data sources, combining both news text and financial data.

Overall, this research takes a novel approach to the prediction of the daily oil price based on financial text mining. News was collected via the ‘‘Crude oil news’’ column of Investing.com, a leading global financial portal, from September 15th, 2009, to July 20th, 2014. We extract the text features of this online news using sentiment analysis and a CNN model. The outputs of the CNN model and the sentiment scores are then grouped based on the topic identification of each transaction day. Following lag order selection and feature selection, topic-grouped text features and financial data are incorporated into the forecasting models.

According to our empirical results, the text features extracted by our study contain predictive information for oil price forecasting. The LDA model shows that the novel information contained in news text relates mainly to political events, social concerns and natural catastrophes. On the other hand, significant improvements in accuracy can be achieved when such news text information is combined with financial market information in machine learning based forecasting models. Specifically, the percentage forecast improvement relative to the support vector regression (SVR) model of the random forest model using combined features is almost double that of the model using financial features, for both MAE and RMSE.

This study contributes to the literature by extracting text features from online news automatically using deep learning algorithms, and illustrating the explanatory power of text features in crude oil price forecasting. Specifically, we propose a text-based topic-sentiment synthesis approach to time series construction based on the CNN model, sentiment analysis and topic identification.

This approach is capable of supplementing statistical data with additional information about unexpected political or social events and real-time fundamental factors. By doing so, the approach improves crude oil price forecasting performances significantly.

The rest of the paper is organized as follows: Section 2 provides a brief literature review of market prediction combined with text mining techniques. Section 3 presents the methodology and system design of this paper, including the CNN model, sentiment analysis, feature grouping based on the LDA model, lag order selection and feature selection. Section 4 presents the data, preprocessing and descriptive statistics. Sections 5 and 6 describe the empirical results of online news text mining and oil price forecasting. Finally, a discussion of our findings and conclusions is given in Section 7.

Literature review

Many researchers have implemented various models for forecasting crude oil prices and their determinants. Most empirical studies on forecasting oil prices rely on econometric models or intelligent algorithms (Abramson & Finizza, 1995; Atalla, Joutz, & Pierru, 2016; Ye, Zyren, & Shore, 2005). Another strand of the literature on oil price forecasting adopts a decomposition-integration method of empirical mode decomposition (EMD) or wavelet analysis. The evidence from using the EMD approach suggests that the major driving force behind crude oil price fluctuations is midterm significant events (Zhang, Yu, Wang, & Lai, 2009). These researchers also find that irregular events increase the variability in the short-term. Yu, Wang, and Lai (2008) propose an EMD-based neural network ensemble learning paradigm for world crude oil spot price forecasting. However, the decomposition-integration method is likely to suffer from poor forecasting performances when unexpected events make an impact on the market.

One common weakness of all of the approaches discussed above is that the future trends of oil price are deduced based on historical statistical data. The potential information embedded in unstructured big data such as textual data provides a novel data source for oil price forecasting. Recently, a number of studies have made significant contribution within the field of online text mining for market predictions. Among them, sentiment analysis is applied widely for processing unstructured text. For example, Das and Chen (2007) develop a methodology for extracting small investor sentiment from a stock message board. Tetlock (2007) quantitatively measures the interactions between the media and the stock market using daily content from a popular *Wall Street Journal* column and finds that media pessimism has predictive power for stock market prices.

In addition to sentiment analysis, researchers have noticed that online discussions tend to express opinions based on specific topics, and therefore topic models are employed to identify topics that may have effects on market prices. For example, Nguyen, Shirai, and Velcin (2015) attempt to extract topics and the sentiment of the market simultaneously based on a joint sentiment/topic model (JST). As a deep learning technique, the convolutional neural network (CNN) has been used widely in the fields of image classification, speech recognition and sentence modelling (Abdel-Hamid, Mohamed, Jiang, & Penn, 2012; Krizhevsky et al., 2012). In applications related to this paper, CNN model has also been used for sentiment analysis in several studies, such as those by Dos Santos and Gatti (2014), Poria, Cambria, and Gelbukh (2015), and Severyn and Moschitti (2015).

So far, though, related applications in crude oil markets are still limited. Wex, Widder, Liebmann, and Neumann (2013) and Yu, Wang, and Lai (2005) are a few of the researchers who have incorporated text information on financial news into crude oil forecasting. However, none have considered news sentiment or topic effects, which our study measures using sentiment scores and feature grouping method. Furthermore, this study forms the first attempt to apply the cutting-edge techniques of deep learning models to crude oil forecasting.

1.1Methodology

a. *System design*

The primary objective of the system proposed in this study is to mine the qualitative information embedded within financial news, and then incorporate this data into oil price forecasting models. Although sentimental tendencies and price movement information embedded in new text have been proved to have strong correlations with market price changes, both sentiment and text features' impacts on market prices depend on the topic context. For instance, optimistic tunes in a demand trend and supply situation could have the opposite influence on market price change. Thus, our approach designs a feature grouping method based on the LDA model in order to distinguish the news topic contexts. Fig. 1 illustrates the major phases in the system flow.

There are six major phases in the proposed system, namely data retrieval, data preprocessing, new headlines text mining, lag order and feature selection, oil price forecasting, and evaluation. During the first and second phases, news headlines, price data and financial market data are collected separately and preprocessed. In the next phase, text mining, unstructured text documents are transformed into structured time series snippets using CNN classification and sentiment analysis. Then, these features are grouped around topic themes. Finally, both qualitative information from news headlines and financial market data are incorporated into price forecasting models.

1.2 Retrieval and preprocessing

Three datasets are collected independently: news headlines, crude oil price data and financial market data. For the news headlines dataset, we collect all available news data from the ''Crude oil news'' section of a leading global financial portal ''Investing.com''. With over 380 million monthly page-views, it is an optimal one-stop source for investors worldwide.[1] Our study uses news headlines instead of full articles because headlines are easier to retrieve and are mostly summaries of the content that contain sufficient key information. Moreover, short headlines contain less repetition and fewer irrelevant words than the document itself (Nassirtoussi, Aghabozorgi, Wah, & Ngo, 2015).

We also retrieve crude oil price data and financial market data over the same period. For crude oil price data, we use the daily West Texas Intermediate (WTI) crude oil spot prices as predictive variables. For financial data, we choose three indicators that influence crude oil prices: daily WTI futures contract prices traded on the New York Mercantile Exchange (NYMEX), US Dollar index (USDX) and Dow Jones Industrial Average (DJIA). For futures, we select contracts for words are removed. Stop words are some of the most commonly used, short functioning words; for example, ''the'', ''is'', ''at'', ''which'', and ''on'' are all considered stop words in computer search engines. Stop words, punctuation and non-alpha words contain no useful information, and hence are filtered out prior to thematic language processing.

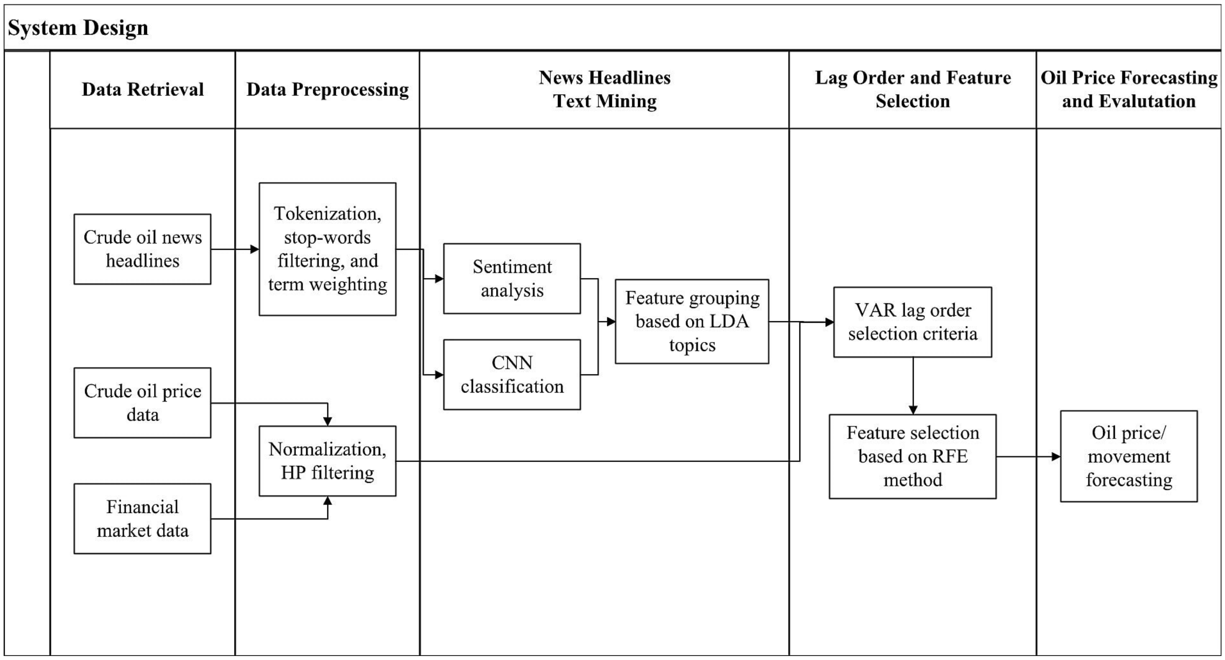

System Design
Data Retrieval
Data Preprocessing
News Headlines Text Mining
Lag Order and Feature Selection
Oil Price Forecasting and Evalutation
Crude oil news headlines
Crude oil price data
Financial market data
Tokenization, stop-words filtering, and term weighting
Normalization, HP filtering
Sentiment analysis
CNN classification
Feature grouping based on LDA topics
VAR lag order selection criteria
Feature selection based on RFE method
Oil price/ movement forecasting

Fig. 1. System design of this paper.

Lizardo & Mollick, 2010; Sadorsky, 1999).[2]

Before further text mining is undertaken, the raw text data are preprocessed by tokenization and stop words, punctuation and non-alpha

This study employs a technique that is common in text vector transformation, ‘‘bag-of-words’’, in which each document is represented by a vector. Each element of a vector corresponds to a word and its frequency within a document. The length of the vector is equal to the number of distinct dictionary words (verbs or nouns) in the corresponding news headline in our dataset. This study also includes a weighting scheme of Term Frequency– Inverse Document Frequency (TF-IDF). TF-IDF reflects the importance of a specific word in a document with respect to a collection of documents, or corpus. The TF-IDF value increases proportionally with the occurrence of a word within a document, and decreases with the number of documents containing this word across the entire corpus.

Specifically, TF-IDF is defined as:

$$\text{tf-idf}\,(t, d, D) = f_{t,d} \cdot \log \frac{N}{n_t}, \qquad (1)$$

where $f_{t,d}$ is the raw frequency of term t in document d, N is the total number of documents in the corpus, and n_t is the total number of documents containing at least one occurrence of term *t*.

Meanwhile, crude oil price data and financial market data are normalized into the same range [0, 1], and this series is then smoothed using the Hodrick-Prescott (HP) filter (Hodrick & Prescott, 1997) to obtain a smoothedcurve representation of a time series. This step creates a more sensitive analysis that is designed for long-term analytics rather than for short-term fluctuations. Following normalization and HP filtering, historical oil price data and financial data are inputted to forecasting models.

1. headlines text mining

The analysis of news headlines and text mining incorporates four sub-phases. For each news headline, the CNN model is used to predict the next day's oil price movement (up or down). Next, we calculate two sentiment scores (a polarity score and a subjectivity score) for each news headline. Next, for each news headline, we use the LDA model to classify the topic to which the news belongs. Finally, we compute average sentiment scores and CNN classifications for each topic on the same transaction day, to form topic-based features. Each phase is described in more detail in the remainder of this section.

1.1.1 model

This study employs a CNN interface from the Python library, *Tensorflow*, for implementing the prescribed network architecture of Zhang and Wallace's (2017) CNN model. This approach is developed specifically for the sentiment analysis of short text.[3] The advantage of the CNN classifier is based on multi-layer networks and convolution architecture. Since convolutional layers of CNNs apply a convolution operation to the input matrix, they are able to compose different semantic fragments of sentences and learn the interactions between composed fragments, thereby fully exploiting inter-modal semantic relations of crude oil news.

Price movements either up or down within the transaction day are used as the output of the CNN classification. Specifically, the price movement M_t is defined as:

$$\{0, \quad pt < pt-1$$

$$Mt = 1, \quad p_t \geq pt-1, \qquad (2)$$

where p_t denotes the oil price at the end of day t.

In this way, the CNN model is trained to learn hidden patterns embedded within news headlines that affect the oil price. The CNN architecture requires parsimonious hyperparameter settings, specifying input word vector representations, filter region size(s), the number of feature maps, activation function(s), a pooling strategy, and regularization terms (dropout/l2). In a previous experiment, we found the best values of the hyperparameters using a grid search method. The results of this hyperparameter experiment are given in Appendix A.

1.1.2 analysis

This study also employs a Python library, *Textblob*, which provides a simple API for diving into common natural language processing tasks, such as sentiment analysis, classification, translation, etc. The sentiment module of *Textblob* returns two sentiment scores for each document, namely polarity and subjectivity scores. Polarity scores are values within the range [−1.0, 1.0], where below zero is considered negative news and above zero positive news. Likewise, subjectivity scores are values within the range [0.0, 1.0], where 0.0 is very objective and 1.0 is very subjective.

i. *Feature grouping based on the LDA topic model*

We apply the Latent Dirichlet Allocation (LDA) modelling approach of Blei, Ng, and Jordan (2003) in order to identify latent topics embedded within the news headlines corpus. We adopt a Python library, *gensim*, for implementing the LDA model. The LDA model is based on the assumption that each document is a mixture of various topics, and each topic has a corresponding probability distribution for different words. Each document d is viewed as a multinomial distribution $\theta(d)$ over T topics, and each topic z_j, $j = 1 \ldots T$, is assumed to have a multinomial distribution $\Phi(j)$ over the set of words W. In order to uncover both the topics that are present and the distribution of these topics in each document from a corpus of documents, D, estimates of θ and Φ are required. In general in this model, estimating various distributions for topics and word probabilities across each topic for exact inferences is intractable. This research implements a common approach, Gibbs sampling, for approximate inference (Blei et al., 2003; Griffiths & Steyvers, 2004). The Kullback–Leibler (KL) divergence approach was used to determine the topic number T. For further details, see Lin, Yeh, and Chen

(2011).

We also incorporate the dynamic topic model (DTM), as described by Blei and Lafferty (2006), in order to consider possible time-varying effects of online news. The DTM uses state space models on natural parameters from multinomial distributions to represent topics. Variational approximations based on Kalman filters are employed in the DTM in order to carry out approximate posterior inference over latent topics. We estimate DTM with all news collected between September 15th, 2009, and July 20th, 2014. Each year is regarded as a time slice. The results of the DTM and the LDA model are then compared in order to investigate the time-varying effects of news texts in this dataset.

After estimations have been calculated using the LDA model, each news headline is classified according to the topic that has the largest probability within the distribution $\theta(d)$. That is, document d is assigned to topic z_j if $p(\theta, z_j) = \max\{p(\theta, z_i), i = 1 \ldots T\}$.

Lastly, for each topic during each transaction day, we compute average scores for the results of the sentiment analysis and CNN classifications, to form sentiment-topic features and CNN-topic features.

These topic features are defined as follows:

Polarity-topic score of topic *k* on day *t* :

$\sum$

Pk,tPi, (3) *nk,t*

Subjectivity-topic score of topic *k* on day *t* :

$\sum$

Sk,tSi, (4) *nk,t*

CNN-topic score of topic *k* on day *t* :

$\sum$

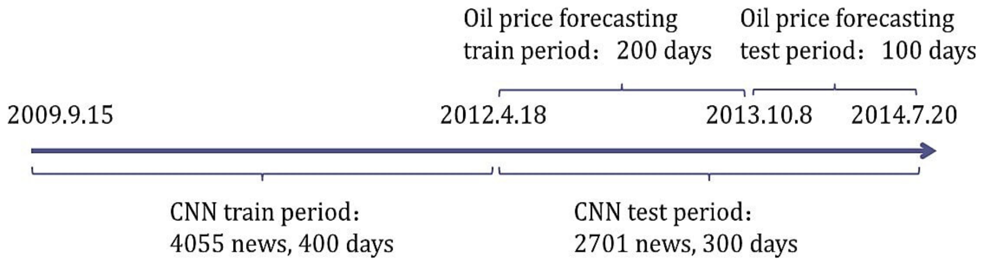
Ck,tCi, (5) *nk,t*

where $n_{k,t}$ is the total number of news headlines from topic k on day t; P_i and S_i are the polarity and subjectivity scores of news headline i belonging to topic k on day t, respectively; and C_i is the CNN classification of price movement M_t, defined in Eq. (2), of the news headline i belonging to topic k on day t.

b. *Oil price forecasting*

During this phase, both topic-based text features (sentiment scores and CNN classifications) and financial market indicators are inputted into oil forecasting .

Fig.



2. Training and testing periods of the CNN model and the oil price forecasting models.

We start by selecting the lag order using the VAR lag order selection criteria, then conduct feature selection across all features including all lag terms of the features. Finally, selected features are inputted into three forecasting models in order to evaluate and contrast their algorithmic performances.

We select the lag order of the input features by implementing the VAR lag order selection criteria using Eviews (Ivanov & Kilian, 2005). This method constructs a VAR model with an input variable and then selects the lag order using five predefined criteria, namely the sequentially modified LR test statistic (LR), the final prediction error (FPE), and the Akaike (AIC), Schwarz (SC), and HannanQuinn (HQ) information criteria.

Furthermore, we perform feature selection based on the recursive feature elimination (RFE) method using the R software *caret*. This algorithm implements a forecasting model (here, random forest) and performs a backwards selection of the inputted features based on a feature importance ranking: the features are ranked and the less important are eliminated sequentially prior to modelling. This algorithm finds a subset of features that can be used to produce the highest accuracy (Genuer, Poggi, & TuleauMalot, 2010).

This study includes the random forest and support vector regression (SVR) models as oil price forecasting models, and uses a linear regression model with ordinary least squares (OLS) as a benchmark. The random forest constructs a multitude of decision trees and deals with problems such as overfitting and small numbers of input variables (Breiman, 2001). In addition, the SVM/SVR model has been used widely, and has performed well compared with other forecasting models in the majority of empirical studies (Fung, Xu Yu, & Wai, 2003; Schumaker & Chen, 2009; Soni, van Eck, & Kaymak, 2007). The model specifications of the random forest and SVR models are determined using a grid search method.

2.Data, preprocessing and descriptive statistics

In this study, three datasets are used as inputs, namely a news headlines dataset, a historic oil price dataset and a financial market dataset. We collect 6756 news headlines from individual trading days, released in the ‘‘Crude oil news’’ column from September 15th, 2009 to July 20th, 2014. Furthermore, the daily WTI spot price, WTI futures price, US Dollar index (USDX) and Dow Jones Industrial Average (DJIA) from the same period are retrieved from the WIND database, and include 700 trading days. Collecting articles from the ‘‘Crude oil news’’ section of the website, the news count on non-trading days of oil market is only a small portion of the total dataset (280 news articles out of 7036).

The dataset is partitioned into two sets of training and testing periods, as shown in Fig. 2. For the CNN model, the training set is 2009.9.15–2012.4.18, and includes 4055 news articles and 400 daily records. The comparative test set is 2012.4.19–2014.7.20, and includes 2701 articles and 300 daily records. Since the CNN classifications are used as input variables for oil price forecasting, the training and testing periods of the price forecasting model are settled using the CNN test period. That is, the training set for the oil price forecasting model is 2012.4.19–2013.10.8, including 200 daily records, while the testing set is 2013.10.9– 2014.7.20, including 100 daily records. We use rolling windows for estimating and further testing the oil price forecasting models.

After the data preprocessing that is described in Section 3.2, descriptive statistics are reported in Table 1. All of the datasets exhibit excess kurtosis, which suggests that the computed datasets do not follow a normal distribution. This observation is confirmed using the Jarque–Bera test, which rejects the null hypothesis of a normal distribution at significance levels of 5% and 1%. The augmented Dickey– Fuller (ADF) unit-root test devised by Dickey and Fuller (1979) is applied and clearly demonstrates the stationarity of all datasets except DJIA.

3 Mining of online news

3.1 CNN text classification and sentiment analysis

Hyperparameters are set for the CNN model based on a hyperparameter experiment, described in Appendix A. The performance of the CNN model on the test set is described in Table 2.

The accuracy of the CNN model is 61%, which is not as high as might be expected. We also use rolling test windows for every single news headline as standard time series forecasting does, and achieve an inferior accuracy of 59.8%. One possible reason could be that online news may not provide enough information about the factors that affect the oil price. Online news tends to focus mainly on a few breaking events in the market, while any changes in supply, demand or related financial markets will lead to oil price movements. However, online news is unlikely to and need not cover all changes in these indicators. As a result, CNN classifications alone may not provide enough predictive information for oil price forecasting. Thus, other sources of financial market data must be incorporated into oil price forecasting models in order to improve their performances.

Table 1

Descriptive statistics of the oil price, the Dow Jones Industrial Average, the US dollar index and the oil futures price.

	Oil price	DJIA	USDX	Futures price
Mean	0.7973	0.4830	0.4525	0.5440
Median	0.7918	0.5660	0.3515	0.5236
Std. Dev.	0.0742	0.3124	0.2897	0.2531
Skewness	−0.1216	0.0357	0.4841	−0.0993
Kurtosis	2.2183	1.5375	1.7972	2.1968
Jarque–Bera	8.3762**	26.7968***	29.8018***	8.5535**
Augmented Dickey-Fuller	−2.9398**	−0.1486	−2.7134*	−2.85636*

Table 2

CNN classification results.

Accuracy	Precision	Recall	F-measure

0.61	0.6029	0.3069	0.6508

Note: The accuracy, precision, recall and F-measure are defined as follows:

Accuracy = $(TP + TN)/(TP + FP + TN + FN)$

Precision = $TP/(TP + FP)$

Recall = $TP/(TP + FN)$

F-measure = $2 * TP/(2 * TP + FP + FN)$,

where TP is the number of positive observations which are classified as positive; FP is the number of positive observations which are classified as negative; TN is the number of negative observations which are classified as negative; and FN is the number of positive observations which are classified as negative.

The hyperparameters of the CNN model are set as follows: embedding dimension = 128; filter size = 3,4,5; number of filters = 128; drop out probability = 0.5; and l2 regulation = 0.

Table 3

Descriptive statistics of the polarity and subjectivity scores.

	Polarity	Subjectivity
Mean	−0.2052	0.5863
Median	−0.2028	0.5825
Std. Dev.	0.0864	0.0795
Skewness	−0.2289	0.0120
Kurtosis	3.2235	3.4328
Jarque–Bera	29.2139***	21.1505***
Augmented Dickey-Fuller	−5.8135***	−7.3942***

*** Denotes significance at the 1% level.

In addition to the CNN model, we also analyse possible effects of a sentiment tendency expressed by online news. Table 3 presents descriptive statistics of sentiment scores. Similarly to the price and finance datasets, the sentiment scores do not follow a normal distribution, while the ADF test suggests the stationarity of the data sets.

C. *Feature grouping based on LDA topics*

We distinguish the different effects of news topics by analysing sub-grouped text features based on LDA topics. Specifically, the topics embedded in online news are classified based on the LDA model, and each segment of news is assigned to one topic. The CNN classifications, polarity scores and subjectivity scores are then grouped again based on LDA topics.

Accordingly (see Section 3.3), the number of topics is determined by the largest Topic_KL divergence. Table 4 depicts the Topic_KL divergence of topic numbers ranging from 3 to 16. The largest Topic_KL divergence appears for a topic number of 4.

According to the statistical inferences generated using the LDA model, four topics are identified, along with their topic distributions, topic words and word distributions. Table 5 presents the top 20 words with the largest weightings in each topic, where some impact factors of oil price can be extracted. For example, ''natural'' and ''gas'' in Topic 1 represent a closely related commodity market to crude oil, namely the natural gas market. Topic 1 may also reflect some climate factors (''weather'' and ''temperature'') such as catastrophes or abnormal climatic changes. The words ''crude'', ''oil'' and ''supply'' in Topic 2 may suggest fundamental factors of crude oil. In addition, ''euro'' and ''dollar'' in Topic 2 suggest some exchange rate factors. Furthermore, Topic 3 reflects political events, according to specific words such as ''Russia'' and ''Iran''. ''China'' in Topic 3 may suggest that emerging economies in Asia can affect oil prices because of their demand for industrial raw materials (Kilian, 2009). Topic 4 mainly represents words related to events and changes in the stock market, e.g. ''stock'', ''S&P'', ''Dow'' and ''Jones''. These topics, extracted using the LDA model, contain some qualitative information that are difficult for statistical indicators, such as weather and political events, to reflect.

We also perform dynamic topic modelling (DTM), in order to investigate the time-varying effects of online news. The number of topics (four) is also selected based on KLdivergence. The results indicate that the estimated frequencies of most of the topic words remain stable over the six years of the sample period. For example, in Topic 1 of the DTM, the estimated frequencies only changed for a minority of the words and by a narrow margin over the period (''dollar'' from 0.0040 to 0.0050; ''ahead'' from 0.0040 to 0.0050; and ''global'' from 0.0040 to 0.0050; the remainder stayed the same over the six years). The results for the other topics are similar to those for Topic 1, suggesting that the time-varying effect in our news dataset is marginal.[4] For this reason, the feature grouping was based on the results of the LDA model.

After each news headline has been assigned to one of the four topics, we perform feature grouping based on LDA topics. Since Topics 3 and 4 contain only a few news headlines (3.92% for Topic 3 and 6.76% for Topic 4), we only perform feature grouping on Topics 1 and 2. Based on Eqs. (3)–(5), each of the text features is sub-grouped into one of two topic-based features: CNN-Topic1, CNN-Topic2, Polarity-Topic1, Polarity-Topic2, Subjectivity-Topic1, and Subjectivity-Topic2.

Fig. 3 shows a trend chart for each topic-based feature compared with oil prices over the period of the test dataset. The green and red lines represent the text features of Topics 1 and 2 respectively, while the blue lines represent oil prices. All of the topic-based features present similar trends to oil prices, either contemporarily or with a slight lag. We conclude qualitatively that topic-based features help to forecast oil prices. In the next section, we examine the correlations between topic-based features and oil prices empirically, and evaluate the forecasting performances of text features.

2.Crude oil price forecasting

This section performs lag order selection and feature selection in order to achieve high quality combinations for the forecast modelling of input features. Next, we perform a Granger causality test and build a VAR model to justify the explanatory power of text features. Finally, oil price forecasting performances are evaluated using several predefined criteria.

We examine cross correlations between topic-based text features and oil prices by performing Granger causality tests (Granger, 1969), as presented in Appendix B. According to the results, all of the topic-based text features significantly Granger cause the oil price within two days except for Subjectivity-Topic1. Moreover, the oil price significantly Granger causes most of the text features within two days. This suggests that topic-based text features are correlated significantly with the oil price and contain predictive information for oil price forecasting.

We justify the explanatory power of text features by building a VAR model using text features and financial variables. The impulse response functions of the VAR model are then estimated and the variance decomposition undertaken, to investigate the percentage contribution of the text-based data. The VAR model specifications and empirical results are presented in Appendix B. The results of the impulse response analysis may suggest that information shocks to the natural gas market from online news cause negative changes in oil prices. At stable state, CNN-Topic1 contributes to 5.18% of the oil price fluctuations and CNNTopic2 contributes to 19.35%. Thus, the VAR model results confirm that text features have explanatory power and contain predictive information for the oil price.

d. *Lag order and feature selection*

This study takes the lag effect into account, and performs feature selection prior to oil price forecasting. We implement VAR lag order selection criteria for each of the text features and financial features in turn, and perform feature selection for all lags selected by the VAR lag order selection criteria.

The full results of the lag order selection criteria are presented in Appendix C. Table 6 depicts selected lag orders for all text and financial features, and demonstrates that the best selected lag order for all features is five. This suggests that the text features extracted by this study, as well as the financial features, have an impact on oil markets for a period of five days. The feature selection processes for these features and all lag terms of the features involve a total of 54 features before feature selection.

This study employs the recursive feature elimination (RFE) method for feature selection. The best combination of input features is presented in Table 7, which contains of 27 features (the numbers in brackets are the lag orders of the variables). Fig. 4 displays the root mean squared error (RMSE)[5] of the random forest model for all combinations of input features, and indicates that the combination of 27 features achieves the smallest RMSE.

According to the results of the feature selection, almost all of the financial features, including all five lags – except DJIA (−3) – are selected by the algorithm. Specifically, most lags of the WTI oil futures price are ranked as more important than the other features. These results are consistent with previous research, in that the futures market is closely related to the spot market (Bekiros & Diks, 2008; Kaufmann & Ullman, 2009; Maslyuk & Smyth, 2009). The results also suggest that financial features reflect more predictive information for the oil price than text features, at least for short-term fluctuations.

On the other hand, all lag orders of the two CNN-topic based features are selected by the algorithm, while none of the sentiment-topic based features are selected. This indicates that the sentiment of online news is correlated less closely with oil price change than the CNN extracted text features. Thus, the proposed CNN model is capable of eliciting hidden patterns and predictive information that is embedded in online news. In addition, CNN text features are ranked more highly than some of the financial features, in terms of forecasting power, which also suggests that text features contain new predictive information that financial features do not reflect.

e. *Crude oil price forecasting*

The selected input features derived using the RFE are then used in oil price forecasting models. We evaluate the models' forecasting performances over the test period using two different criteria: the mean absolute error (MAE) and root mean squared error (RMSE), which are defined as follows. The results are presented in Table 8.

n

$\sum$

MAE =(6)

�

�1 n

RMSE , (7)

where n is the number of observations in the test period, y_i is the actual value of the oil price on day i, and $\hat{}y_i$ is the price forecast obtained using the forecasting models. When computing the criteria, we use rolling windows for evaluating the forecasts. At each window, the parameters are updated prior to price forecasting.

We quantify the additional explanatory power of the text features by computing a new criterion, the percentage improvement, which means the percentage improvement in forecasting performance between the combination of the two features and the single group of financial features. Specifically, the percentage improvement is defined as:

Percentage Improvement

(MAE of Financial Features)−(MAE of Combination) = ,

MAE of Financial Features

(8)

Percentage Improvement

(RMSE of Financial Features) − (RMSE of Combination) = .
RMSE of Financial Features

(9)

Table 6

The lag orders selected using the VAR lag order selection criteria for each of the text and financial features.

	CNN-Topic1	CNN-Topic2	Polarity-Topic1	Polarity-Topic2	Subjectivity-Topic1
Lag order selected	5	5	5	5	5
	Subjectivity-Topic2	Futures	USDX	DJIA	
Lag order selected	5	5	5	5	

Input features selected by RFE (ranked by the variable importance[a])

1. Futures 2. Futures (−1) 3. Futures (−2) 4. Futures (−3)
5. USDX 6. Futures (−4) 7. CNN-Topic2 (−4) 8. CNN-Topic2 (−5)
9. USDX (−1) 10. CNN-Topic2 11. CNN-Topic1 (−4) 12. CNN-Topic2 (−3)
13. Futures (−5) 14. USDX (−4) 15. USDX (−2) 16. DJIA
17. USDX (−3) 18. CNN-Topic2 (−1) 19. CNN-Topic1 (−3) 20. CNN-Topic2 (−2)
21. CNN-Topic1 (−5) 22. USDX (−5) 23. DJIA (−1) 24. DJIA (−5) 25.
CNN-Topic1 (−2) 26. DJIA (−4) 27. CNN-Topic1 (−1)

Random forest	0.0785	0.0082	0.0073	**12.32%**
SVR	0.0252	0.0032	0.0030	**6.67%**
Linear regression	0.0854	0.0035	0.0045	−22.22%

Random forest	0.0883	0.0092	0.0088	**4.55%**
SVR	0.0325	0.0041	0.0040	**2.50%**
Linear regression	0.0953	0.0044	0.0056	−21.42%

Fig. 4. Root mean squared error (RMSE) for all combinations of input features.

Technically, the percentage improvement represents the forecasting improvement obtained by adding text features to financial features.

Our results suggest that SVR performs better than either the random forest or the linear regression in most scenarios. Specifically, for the MAE and RMSE criteria, the SVR model using all selected features achieves the best performance among all of the models. Overall, the SVR and random forest models perform better than the benchmark linear regression model in most of the scenarios.

We investigated the relative predictive power of text and financial features by comparing the performances of text features only (10 features), financial features only (17 features), and all selected features (27 features). The results show that the performances in the scenarios where only text features are inputted into the models are much poorer than those for financial features and all selected features. This suggests that text features alone may not provide enough predictive information for oil price forecasting. Furthermore, with the random forest and SVR models, the combination of text features and financial features achieves the best performance, which is slightly better than that for financial features only.

The results of the percentage improvement, defined by Eqs. (8)–(9), are presented in the last column of Table 8. For both MAE and RMSE, the improvement of combined features over financial features alone with the random forest model is almost double that of the SVR model. However, in the case of linear regression, the model using combined features performed less well than the one using single financial features. The forecasting results generally suggest that text and financial features complement each other for oil price forecasting. Significant improvements in accuracy can be achieved when news text information is combined with financial market information in a machine learning based forecasting model.

2. Conclusions

This study proposes a new text-based crude oil price forecasting method using deep learning techniques, sentiment analysis and topic extraction. Hidden patterns in the way in which news text corresponds to oil price movements can be uncovered using a convolutional neural network (CNN). We have designed two sentiment scores and calculated quantifiable sentiment analysis. A feature grouping method is proposed based on LDA topic modeling in order to recognize the levels of influence from different types of online news topics. Following a further exploration of lag effects and feature selection, this study proposes and implements a synthesis forecasting framework. Our empirical results show that the proposed approach achieves a favourable accuracy for crude oil forecasting.

According to the results of this study, news text does add meaningful information to crude oil price forecasting, and a significant improvement in accuracy can be achieved when such news text information is combined with financial market information. The results of LDA modelling show that crude oil news text contains novel information that statistical data do not possess. This novel information is related mainly to political events, social concerns and natural catastrophes. Further empirical investigations of the Granger test and the VAR model confirm that the text features extracted from online news present significant cross-correlations with the oil price, and contribute a portion of the oil price variance. The explanatory power of text features can be maximized by combining them with financial features, with the use of both text and financial features in random forest and SVR models increasing the forecast accuracy according to both the MAE and RMSE. Specifically, the improvement in the random forest model when using combined features rather than only financial features is almost double the improvement seen in the SVR model.

Overall, the main contribution of this study is the introduction of a novel data-driven machine learning approach which utilizes predictive information from news text that cannot be obtained using statistical data alone. Such an approach can increase the crude oil price forecasting accuracy by considering unexpected political or social events that are reported by news media. At the same time, this approach may also provide real-time high-frequency data on fundamental factors for crude oil forecasting models. Moreover, our findings contribute to the forecasting methodology, in that a text-based topic-sentiment synthesis approach could be used to construct useful time series indicators for forecasting based on unstructured text data. In most cases with these empirical tests, the highlighted forecasting performances can be improved further by utilizing multiple data sources and combining text features with financial data.

There are various limitations and potential extensions of our study. It is likely that the modeling performance of a deep learning model could be improved by extending the size of the news sample or using full articles instead of only headlines. The proposed approach could be extended to access the public opinion of regulatory announcements, press release and breaking news using further sentiment analysis models, so that expectations could be measured directly for economic forecasts. The N-gram model (Brown, Desouza, Mercer, Pietra, & Lai, 1992) also has potential in distinguishing sentiment and topic context from text. A comparison between the N-gram model and our approach will be conducted in our future work. In addition, the prediction of specific influential events like financial crisis could be implemented by combining our approach with event detection and impact evaluation methods. To explore the applicability of this proposed approach in other forecasting areas, it may be worth trying text-mining methods based on other deep learning models, such as long shortterm memory (LSTM), and/or more sophisticated mixedfrequency models, in which both high frequency text features and low frequency market features can be taken into consideration within the same model.

References

Abdel-Hamid, O., Mohamed, A. R., Jiang, H., & Penn, G. (2012). Applying convolutional neural networks concepts to hybrid NN-HMM model for speech recognition. In *Proceedings of 37th IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 4277–4280).

Abramson, B., & Finizza, A. (1995). Probabilistic forecasts from probabilistic models: a case study in the oil market. *International Journal of Forecasting*, *11*(1), 63–72.

Atalla, T., Joutz, F., & Pierru, A. (2016). Does disagreement among oil price forecasters reflect volatility? Evidence from the ECB surveys. *International Journal of Forecasting*, *32*(4), 1178–1192.

Baumeister, C., Guérin, P., & Kilian, L. (2015). Do high-frequency financial data help forecast oil prices? The MIDAS touch at work. *International Journal of Forecasting*, *31*(2), 238–252.

Bekiros, S. D., & Diks, C. G. (2008). The relationship between crude oil spot and futures prices: Cointegration, linear and nonlinear causality. *Energy Economics*, *30*(5), 2673–2685.

Blei, D. M., & Lafferty, J. D. (2006). Dynamic topic models. In *Proceedings of the 23rd international conference on machine learning* (pp. 113–120). ACM.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research (JMLR)*, *3*(1), 993–1022.

Boureau, Y. L., Ponce, J., & LeCun, Y. (2010). A theoretical analysis of feature pooling in visual recognition. In *Proceedings of the 27th international conference on machine learning (ICML-10)* (pp. 111–118).

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32.

Brown, P. F., Desouza, P. V., Mercer, R. L., Pietra, V. J. D., & Lai, J. C. (1992). Class-based n-gram models of natural language. *Computational Linguistics*, *18*(4), 467–479.

Chen, S. S., & Chen, H. C. (2007). Oil prices and real exchange rates. *Energy Economics*, *29*(3), 390–404.

Cifarelli, G., & Paladino, G. (2010). Oil price dynamics and speculation: A multivariate financial approach. *Energy Economics*, *32*(2), 363–372.

Das, S. R., & Chen, M. Y. (2007). Yahoo! for Amazon: Sentiment extraction from small talk on the web. *Management Science*, *53*(9), 1375–1388.

Déesa, S., Karadelogloua, P., Kaufmann, R. K., & Sáncheza, M. (2007). Modelling the world oil market: Assessment of a quarterly econometric model. *Energy Policy*, *35*(1), 178–191.

Dickey, D. A., & Fuller, W. A. (1979). Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association*, *74*(366a), 427–431.

Dos Santos, C. N., & Gatti, M. (2014). Deep convolutional neural networks for sentiment analysis of short texts. In *Proceedings of the 24th international conference on computational linguistics* (pp. 69–78).

Fung, G. P. C., Xu Yu, J., & Wai, L. (2003). Stock prediction: integrating text mining approach using real-time news. In *Proceedings of 2003 IEEE international conference on computational intelligence for financial engineering* (pp. 395–402).

Genuer, R., Poggi, J. M., & Tuleau-Malot, C. (2010). Variable selection using random forests. *Pattern Recognition Letters*, *31*(14), 2225–2236.

Granger, C. W. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Journal of the Econometric Society*, *37*(3), 424–438.

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences*, *101*(suppl 1), 5228–5235.

Hinton, G., Deng, L., Yu, D., Mohamed, A.-R., Jaitly, N., Senior, A., et al. (2012). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Process Magazine*, *29*(6), 82–97.

Hodrick, R., & Prescott, E. C. (1997). Postwar U.S. business cycles: an empirical investigation. *Journal of Money, Credit and Banking*, *29*(1), 1–16.

Ivanov, V., & Kilian, L. (2005). A practitioner's guide to lag order selection for VAR impulse response analysis. *Studies in Nonlinear Dynamics & Econometrics*, *9*(1), 1–34.

Kaufmann, R. K., & Ullman, B. (2009). Oil prices, speculation, and fundamentals: Interpreting causal relations among spot and futures prices. *Energy Economics*, *31*(4), 550–558.

Kilian, L. (2009). Not all oil price shocks are alike: disentangling demand and supply shocks in the crude oil market. *American Economic Review*, *99*(3), 1053–1069.

Krizhevsky, A., Sutskever, I., & Hinton, G. (2012). ImageNet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, & K. Q. Weinberger (Eds.), *Advances in neural information processing systems, Vol. 25* (pp. 1097–1105). Curran Associates, Inc.

Li, X., Shang, W., Wang, S., & Ma, J. (2015). A MIDAS modeling framework for Chinese inflation index forecast incorporating Google search data. *Electronic Commerce Research and Applications*, *14*(2), 112–125.

Lin, S. H., Yeh, Y. M., & Chen, B. (2011). Leveraging Kullback–Leibler divergence measures and information-rich cues for speech summarization. *IEEE Transactions on Audio, Speech and Language Processing*, *19*(4), 871–882.

Lizardo, R. A., & Mollick, A. V. (2010). Oil price fluctuations and US dollar exchange rates. *Energy Economics*, *32*(2), 399–408.

Maslyuk, S., & Smyth, R. (2009). Cointegration between oil spot and future prices of the same and different grades in the presence of structural change. *Energy Policy*, *37*(5), 1687–1693.

Nassirtoussi, A. K., Aghabozorgi, S., Wah, T. Y., & Ngo, D. C. L. (2015). Text mining of news-headlines for FOREX market prediction: A multi-layer dimension reduction algorithm with semantics and sentiment. *Expert Systems with Applications*, *42*(1), 306–324.

Nguyen, T. H., Shirai, K., & Velcin, J. (2015). Sentiment analysis on social media for stock movement prediction. *Expert Systems with Applications*, *42*(24), 9603–9611.

Poria, S., Cambria, E., & Gelbukh, A. F. (2015). Deep convolutional neural network textual features and multiple kernel learning for utterancelevel multimodal sentiment analysis. In *Proceedings of the 2016 conference on empirical methods on natural language processing* (pp. 2539–2544).

Sadorsky, P. (1999). Oil price shocks and stock market activity. *Energy Economics*, *21*(5), 449–469.

Schumaker, R. P., & Chen, H. (2009). Textual analysis of stock market prediction using breaking financial news: The AZFin text system. *ACM Transactions on Information Systems*, *27*(2), 12.

Severyn, A., & Moschitti, A. (2015). Unitn: Training deep convolutional neural network for Twitter sentiment classification. In *Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015)* (pp. 464–469). Denver, Colorado: Association for Computational Linguistics.

Soni, A., van Eck, N. J., & Kaymak, U. (2007). Prediction of stock price movements based on concept map information. In *Proceedings of 2007 IEEE symposium on computational intelligence in multicriteria decision making* (pp. 205–211), Honolulu, HI.

Tetlock, P. C. (2007). Giving content to investor sentiment: the role of media in the stock market. *The Journal of Finance*, *62*(3), 1139–1168.

Wex, F., Widder, N., Liebmann, M., & Neumann, D. (2013). Early warning of impending oil crises using the predictive power of online news stories. In *46th hawaii international conference on system sciences (HICSS)* (pp. 1512–1521). IEEE.

Ye, M., Zyren, J., & Shore, J. (2005). A monthly crude oil spot price forecasting model using relative inventories. *International Journal of Forecasting*, *21*(3), 491–501.

Yu, L., Wang, S., & Lai, K. K. (2005). A rough-set-refined text mining approach for crude oil market tendency forecasting. *International Journal of Knowledge and Systems Science*, *2*(1), 33–46.

Yu, L., Wang, S., & Lai, K. K. (2008). Forecasting crude oil price with an EMDbased neural network ensemble learning paradigm. *Energy Economics*, *30*(5), 2623–2635.

Zhang, Y., & Wallace, B. (2017). A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification. In *8th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (pp. 253–263).

Zhang, X., Yu, L., Wang, S., & Lai, K. K. (2009). Estimating the impact of extreme events on crude oil price: an EMD-based event analysis method. *Energy Economics*, *31*(5), 768–778.

[1] This paper might benefit from extending the analysis to other news sources such as the ‘‘Wall Street Journal’’, ‘‘Financial Times’’ and ‘‘The Economist’’. However, since our news source ‘‘Investing.com’’ is a portal website, there is a large overlap of information between ‘‘Investing.com’’ and other sources. For example, we find that ‘‘Investing.com’’ not only covers most of the important events reported in ‘‘Financial Times’’ and other news sources, but also provides more exclusive information. For example, over the period from September 1st, 2017, to September 30th, 2017, the number of crude oil news stories on ‘‘Investing.com’’ is 265, while the corresponding number in ‘‘Financial Times’’ is 96.

[2] The financial data and internet data are measured accurately and are available in real time, while the lower-frequency macroeconomic data tend to be subject to revisions and are available only with a delay (Baumeister, Guérin, & Kilian, 2015; Li, Shang, Wang, & Ma, 2015). The main purpose of our research is to investigate whether the use of internet data helps to explain the short-term fluctuations in oil prices. As a result, financial data are more suitable for our daily forecasting models than monthly demand or supply variables.

[3] The CNN model begins with a tokenized sentence which is then converted to a sentence matrix, the rows of which are word vector representations of each token. We treat the sentence matrix as an ‘image’, and perform convolution on it via linear filters. The dimensionality of the feature map generated by each filter will vary as a function of the sentence length and the filter region size. Thus, a pooling function is applied to each feature map to induce a fixed-length vector. A common strategy is 1-max pooling (Boureau, Ponce, & LeCun, 2010), which extracts a scalar from each feature map. Together, the outputs generated from the filter maps can be concatenated into a fixed-length, ‘top-level’ feature vector, which is then fed through a softmax function to generate the final classification.

[4] The reason why our news dataset observes little time-varying effect is that we focus on crude oil news rather than general news reports. The news topics and important events concerning with crude oil market are almost stable over the time periods. For example, crude oil news constantly reports on oil supply, demand, global economics, political events or other relevant topics, no matter of the time point.

[5] RMSE is defined in Eq. (8).