Using Four Different Online Media Sources to Forecast Crude Oil Price

Mohammed Elshendy
Department of Enterprise Engineering, University of Rome Tor Vergata, Italy.
Andrea Fronzetti Colladon
Department of Enterprise Engineering, University of Rome Tor Vergata, Italy.
Elisa Battistoni
Department of Enterprise Engineering, University of Rome Tor Vergata, Italy.
Peter A. Gloor
MIT Center for Collective Intelligence, Massachusetts Institute of Technology, US.

Abstract

This study looks for signals of economic awareness on online social media and tests their significance in economic predictions. The study analyzes, over a period of two years, the relationship between West Texas Intermediate crude oil price daily movements and multiple predictors extracted from Twitter, Google Trends, Wikipedia, and the Global Data on Events, Language, and Tone database (GDELT). Semantic analysis is applied to study the sentiment, emotionality, and complexity of the language used. ARIMAX models are used to make predictions and to confirm variable values. Results show that the combined analysis of the four media platforms carries valuable information in financial forecasting. Twitter language complexity, GDELT number of articles and Wikipedia page reads have the highest predictive power. The study also gives evidence of the different speeds at which collective awareness is reflected on different media. In comparison to previous works, more media sources and more dimensions of the interaction and of the language used are combined in a joint analysis.

Keywords: Market Forecast; Oil Price; Semantic Analysis; Twitter; GDELT; Google Trends; Wikipedia.

1. Introduction

The art of prediction has a long history. People have always tried to design schemes, implement hypotheses and sometimes create legends to forecast certain outcomes of physical and natural events. In financial market studies the main aspect of making correct predictions depends on identifying the most relevant and critical predictors. Some of these are sometimes difficult to identify or measure, are constantly changing, or may not have been completely explored.

Surowiecki's (2005) seminal book "The Wisdom of Crowds" claims that large groups of individuals are better at making decisions about uncertain events than experts. The wisdom of crowds is meant to be the public opinion or mood. In terms of gathering, quantifying and qualifying wisdom of crowds, business executives and Academia have found an instrument in the social media.

Apparently, social media has exploded as a category of online discourse where people create, share, bookmark and network contents at a prodigious rate (Asur and Huberman 2010). According to “We Are Social” – a comprehensive study of digital, social and mobile usage around the world (http://wearesocial.com/uk/special-reports/digital-in-2016) – there were 3.419 billion internet users at the start of 2016, representing almost 46% of the Earth’s population. Out of these, there were 2.307 billion people – 31% of the global population – who were active users of social media platforms. It is therefore reasonable to say that social media represents a revolutionary trend, being always more of interest to companies operating in online space or in any space (Kaplan and Haenlein 2010).

This paper tests the relevance of selected online open data sources to forecast financial events. It analyzes in depth the sentimental and daily traffic activity of four different social media platforms and tools, in order to predict the West Texas Intermediate (WTI) Crude Oil Price movements. In this context, information from the following online platforms were integrated: (a) the microblogging and media-sharing platform – Twitter; (b) the user generated web encyclopedia – Wikipedia; (c) the online traffic analysis platform – Google Trends; and (d) the world news database of the GDELT Project. The choice of predicting the WTI Crude Oil Price movements is due to its chaotic behavior, which follows a nonlinear dynamic deterministic process – as proved by Moshiri and Foroutan (2006). Thus, understanding the dynamics of Crude Oil Prices, either spot or futures, represents one of the most striking challenges to the forecasting abilities of private and public institutions worldwide (Elekdag et al. 2008).

The main contribution of this study is to propose a new model which can be integrated with the existing crude oil price forecasting techniques. Authors’ findings can help decision makers – either firms, private investors, or individuals – when choosing to buy or sell the crude oil, or when making other investments on the stock market. The price of crude oil is not just an asset itself: forecasting its movements can be extremely useful as they are connected to other stock prices and have a direct impact on several goods and services, such as transportation costs. Simultaneously, it affects export and import costs, which are part of the Gross Domestic Product. An additional contribution is in the fact that this study combines social data from four different sources. Previous works, on the other hand, studied a maximum of two platforms to build forecasting models and usually considered each of them individually (e.g., Choi and Varian 2009; Zhang et al. 2012). By contrast, this paper implements multiple variable/platform models and explores their effects in time. The aim is to compare different types of available online social platforms, either separated or combined, in order to analyze different online publics, ranging from individuals to governments and news outlets. Finally, the study allows a comparison of the different foresighting abilities of each platform, in terms of how many days ahead a platform can predict a price movement before it happens, i.e. a measure of the speed at which the collective mind is reflected by that specific platform.

1.1. Choice of Social media platforms

The choice of the four media platforms included in the analysis was driven by their size – in terms of number of users and produced contents (the bigger the data source, the larger the information that can be analyzed) –, by their reputation, and by the possibility to get free access to the data. Each specific choice is better supported in the following.

Twitter is a worldwide popular platform, which offers a social networking and microblogging service. It enables its users to update their status in tweets, to follow people they are interested in, and to communicate with them directly. In this way, Twitter offers us a novel way of capturing the collective mind up to the last minute (Zhang et al. 2012). Twitter users include, but are not limited to, ordinary individuals, commodity traders, politicians, companies, activists, and major news outlets. Twitter describes itself as "a real-time information network that connects you to the latest information about what you find interesting". It is a social awareness system, which delivers a fragmented mix of information, enlightenment, entertainment, and engagement from a range of sources (Hermida et al. 2014). In 2012, Twitter had more than 100 million registered users posting more than 340 million tweets per day [2], updated to be 310 million monthly active users on March 2016 [1]. Thus, Twitter popularity has drawn more and more researchers' attention from different disciplines, to understand its usage and community structure, influence of users and information propagation, and its prediction power and potential application to other areas (Zhang et al. 2011).

In addition to Twitter, Wikipedia, the well-known online encyclopedia, has been included in this research. Currently, Wikipedia is the largest knowledge repository on the web. Wikipedia is available in dozens of languages, and its English version is the largest of all with more than 400 million words in over one million articles (Gabrilovich and Markovitch 2007). It is also densely structured: its articles have in total hundreds of millions of links. These connections link the topics being discussed, and provide an environment which fosters serendipitous gathering of information (Milne and Witten 2008). This huge amount of information and links provides a real opportunity to help unfold the world history and explore upcoming events. Different from Twitter, the quality of information in Wikipedia is controlled and consequently higher. Within the Wikipedia research community, findings are constantly published and verified, and the reputation of an author grows with the reputation of his contributions (Staub and Hodel 2016). The distribution of users' reputation in Wikipedia shows that saboteurs and inexpert users are quite a minority compared to high reputation users (Javanmardi et al. 2009).

Google is the first search engine in the world, which makes it one of the most reliable resources for investigating web search queries. Since 2004, Google has been providing three data sources that can be useful for social science: Google Trends, Google Correlate, and Google Consumer Surveys (Stephens-Davidowitz and Varian 2014). Google Trends is commonly used in "now-casting" or in the prediction of the present, the very near future, and the very near past (Banbura et al. 2010). In this study, the search queries were set to be the same as the page titles of Wikipedia, mainly for two reasons: (1) to show

how different publics can produce different predictive activity using the same keywords on two different platforms; (2) to show the response time difference between the two platforms.

The Global Data on Events, Location and Tone (GDELT) is the last considered platform. The GDELT Project is an open source repository of news articles, which is continuously updated and made available to researchers through an application program interface. Initially, the GDELT Project was a coded dataset of 200 million geo-located events; now, it has been updated to be 400 million events spanning over more than 12,900 days [3]. The dataset includes more than 300 different types of events: therefore it reveals all that has happened in place and time, since 1979. Kwak and An (2016) referred to it as a tale of the world. The GDELT Project is described as an initiative to construct a catalog of human societal-scale behavior and beliefs across all countries of the world. It connects every person, organization, location, count, theme, news source, and event across the planet into a single massive network. The database relies on tens of thousands of broadcasts, print and online news sources from every corner of the globe (Leetaru and Schrodt 2013). Including GDELT in the analysis is important to control for world events related to the WTI Crude Oil Price and to have a proxy for media activities on newspapers.

2. Literature Review: Predicting the Oil Price

Oil price forecasting techniques can be classified into two main categories: econometric/financial methods and computational models. Econometric models are usually based on a quantitative approach, combining oil price historical data with analytical models. Scholars make frequent use of ARIMA-, ARCH/GARCH-, and Markov Switching- techniques (e.g., Behmiri and Pires Manso 2013); in several cases these models do not include external predictors and just rely on the historical trends of the input variable. On the other hand, computational models are mostly based on the inclusion of explanatory variables, either oil price or economic related, in order to explain the relationship between spot and future prices. They make a frequent use of Artificial Neural Networks (ANN), Support Vector Machine (SVM) and text mining techniques. Recently, the use of computational models is more frequent due to the major advances of computer technologies and the large pools of data available online (Fronzetti Colladon and Remondi 2017; De Mauro et al. 2016). Replicating these models can however be more complicated, since their setup is mostly data dependent and the contribution of single features is sometimes more difficult to isolate. Since one of the scopes of this research is to show the predictive potential of four different media platforms, the choice has been to use ARIMAX models which can include external independent variables and, at the same time, give evidence of the contribute of each of them. In addition, the findings described in the paper can be used as the base for future predictive models which use a machine learning approach.

Cheong (2009) proposed an ARCH model to forecast the time varying volatility of crude oil prices on the short, medium and long period for the WTI and Europe Brent benchmarks. Similarly, Wei et al. (2010) proposed nine linear and nonlinear generalized GARCH models to capture the volatility features for the same benchmarks. Results showed the superiority of nonlinear models to linear ones especially on longer horizons.

Fernandez (2007) proposed a non-seasonal ARIMA model for the time series analysis of daily prices of the Dubai crude oil and natural gas. Their models were based on performing sample autocorrelation and partial autocorrelation tests. Results showed the superiority of the model in making short-term predictions even when compared to ANN and SVM models; by contrast, SVM models proved to be the best when making prediction in the long term. Torban (2010) compared the prediction performance of the same three models, applying the analysis to quarterly data of free on board (FOB) crude oil prices. Partially contrasting Fernandez (2007), results showed a worse performance of the SVM approach.

Shambora and Rossiter (2007) forecasted crude oil price movements building an ANN model and analyzing the traded futures contracts on the period between April 1991 and December 1997. In addition, they applied a random walk model and a simple moving average model. Empirical results showed that the ANN model is more accurate than the others. However, the overall performance of the three models suggested that traded futures contracts are not an efficient predictor.

Yu et al. (2005) proposed a Refined Text Mining approach for crude oil price forecasting. The model was based on the refining of a dataset obtained extracting unstructured text documents, related to the crude oil price query, from Google results. They compared the forecasting ability of their model with regression, random walk, and ARIMA models. They concluded that random walk and regression models performed worst when compared to ARIMA and refined text mining. Wang et al. (2004) proposed a novel nonlinear integrated model called TEI@I to predict WTI Crude Oil Prices: they integrated ARIMA and a back propagation neutral networks. In this way, they were able to capture both linearity and non-linearity characteristics of the time series. In addition, they investigated the existence of irregular activities during the studied period by using web-based text mining techniques, reaching good overall performance results.

Similarly, this study combines ARIMAX time series models with text mining. Therefore, an integrated model combining both econometric and computational techniques is proposed. The historical data of crude oil price are used to explain spot prices, in addition to the inclusion of 10 external independent variables. These variables are non-economic; however, they track the online discourse about crude oil price and include world broadcasts, print, and web news. To the authors' knowledge, there are no previous studies trying to forecast the oil price from the same sources, taken all together. However, some of the variables used here were also considered in other financial predictions: they are discussed in the coming section.

2.1. Hypotheses formulation

In an early effort in using Twitter for financial predictions, Zhang et al. (2011) collected six months of Twitter feeds and measured the collective hope and fear on each day through sentiment analysis. The study showed a significant correlation between crowd's emotions on Twitter and the stock market movements in the next days. Similarly, Bollen et al. (2011b), Mittal and Goel (2012), Chen and Lazer (2011), Sprenger et al. (2014) applied further studies on stock market movements using volume and sentiment of Twitter feeds, showing that Twitter feeds are effective indicators of real world performance. In addition to Twitter feeds and sentiment, Rao and Srivastava (2013) included the search volume index from Google Insights of Search (currently known as Google Trends), in an effort to model and predict oil, gold, and forex market indices. Results showed that their proposed model outperformed earlier works in terms of accuracy and forecasting errors, even if taking into consideration only Twitter sentiment and feeds. Additionally, Twitter prediction ability has been studied in several fields and has shown a good degree of success, e.g. political elections (Tumasjan et al. 2010; Chen et al. 2012), box office (Arias et al. 2013; Baek et al. 2014), or sales (Culotta 2013; Rui et al. 2013). These studies showed significant correlations of Twitter feeds (i.e. number of tweets) and Twitter sentiment with their respective dependent variables. In the financial field, a higher volume of tweets has been usually associated with a state of concern, which can create a large discussion, usually immediately after a financial event. Thus, it more frequently leads to price drops. Oh et al. (2011) showed how the Twitter traffic dramatically increased after 2008 Mumbai terrorist attacks, which produced major loss in stock returns (Sajid Nazir et. al. 2014). On the other hand, a positive sentiment of the language used represents a state of optimism, which can positively reflect on prices (Bollen et al. 2011a). Accordingly, the following hypotheses are formulated:

H1a: The daily count of tweets is negatively associated with oil price movements.

H1b: The more positive the sentiment of the language of the tweets, the higher the oil price.

Twitter analysis was extended to consider other two dimensions of the language used, which are usually less explored. First, language emotionality, which is calculated as the deviation from neutral sentiment; second, the language complexity, which assigns a complexity level to each tweet by measuring how much each term is commonly used in the overall discourse (Brönnimann 2013). Forbergskog (2013) investigated the leading and lagged relationships between positive and negative emotions on the Twitter feeds regarding the S&P 500 index over 84 trading days. Porshnev et al. (2014) investigated the possibility of analyzing the emoticons in the tweets with the aim of forecasting DJIA and S&P500 stock market indices. Both researches concluded that an increase of emotionality in tweets – conceived in the present study as a larger variation between positive and negative sentiment – leads to a reduction of stock market performance and price drops. On the other hand, there seem to be no available studies linking language complexity to financial prediction. Authors hypothesize that a higher complexity is associated with price drops, as the language becomes more heterogeneous and less shared, which can be connected to new radical events or to the intervention of specialized Twitter actors – experts, professional traders or journalists – commenting on such, mostly negative, events (Zappavigna 2012).

H1c: High emotionality values of twitter feeds are negatively associated with oil price movements.

H1d: High complexity values of twitter feeds are negatively associated with oil price movements.

Used broadly in the economic and financial fields, Google Trends showed wide tracking and real time surveillance abilities. For instance, it was used for now-casting of macroeconomic content and probabilities (Koop and Onorante 2013; Giannone et al. 2008), or of unemployment rates (Pavlicek and Kristoufek 2015; Vicente et al. 2015; Chadwick and Sengül 2015). Wu and Brynjolfsson (2013) conducted an explanatory study on how Google search engine data provides an accurate and simple way to predict housing prices and sales activities. Results showed that the housing search index is strongly predictive of future housing markets and sales. Fantazzini and Toktamysova (2015) studied the effect of adding Google Trends to other economic variables in forecasting German car sales. Results showed that new models statistically outperformed old models (not considering Google Trends) for most of the automobile manufacturers and forecast horizons. Choi and Varian (2009a; 2009b; 2012) described how to use Google Trends data to predict several economic metrics, including unemployment rates, automobile demand, and vacation destinations. In their 2012 report, they presented short-term forecasts of multiple economic indicators, showing that the inclusion of Google Trends in the analysis could improve model outcomes by 5% to 20%. Their examples showed a positive association of the volume of search queries with the financial and economic indicators. Moreover, Scott and Varian (2013) predicted gun sales at the national level by analyzing 100 different queries related to the subject. Consistently, this study includes Google Trends search activity in the forecasting of oil price movements, considering two search queries directly related to the oil price movements ("Price of Oil" and "OPEC"). Other queries were also tested with no better associations. The expectation is to find a positive link between the search activity and the oil price.

> *H2: The query counts on Google Trends for terms "Price of Oil" and "OPEC" are positively associated with price movements. More counts means higher prices.*

Recently, Wikipedia have attracted a big attention from scholars working in several fields. For instance, Gloor et al. (2015) analyzed Wikipedia networks of world leaders and introduced a dynamic temporal map of the most influential people of all time. Yasseri and Bright (2015) explored the use of Wikipedia page view data in electoral predictions, for the elections of the members of the European Parliament in 2009 and 2014.

Mestyán et al. (2013) proposed a predictive model for the financial success of movies (before their release) based on the activity level associated to the movie in Wikipedia. Predictive power was strong a few days before the release. Wei and Wang (2016) proved the existence of a link between the number of page views of a company in Wikipedia and its subsequent performance in the stock market. Moat et al. (2013) investigated whether the historical data of Wikipedia page views may contain signs of stock market movements. They considered the page views and edits for all the companies listed in the Dow Jones index, stressing the importance of investigating the activity on Wikipedia while making financial predictions, and finding a negative association of page views with stock prices. With regard to the oil price, similarly to Twitter, a large number of page views could reflect a larger state of concern coming after significant financial events. Accordingly, the following hypothesis is formulated:

H3: The page views count of the specialized Wikipedia articles "Price of Oil" and "OPEC" are negatively associated to price movements.

As regards the more recent GDELT project, scholars have not yet fully explored the link between its data and the trends on financial markets. Up to date, research efforts are mostly focused on political conflicts and demonstrations. Yonamine (2013) used the dataset to predict future levels of violence in Afghanistan districts. Kwak and An (2014) studied GDELT for understanding news geography and major determinants of global news coverage of disasters. Phua et al. (2014) conducted visual and predictive analysis of Singaporean news coverage. They compared Singaporean news articles from 1979 to 2013 and Wikipedia timelines of Singaporean history. Results confirmed the quality and the potential for making predictions based on news articles in GDELT. Bodas-Sagi and Labeaga (2016) used GDELT to analyze the public opinion about the energy policy of the Spanish Government. They carried out the analysis twice, the first time collecting all the data relating to the query "Fuel Price", and then adding only governmental data relating to the same query. Both cases showed a negative association between GDELT activity and Spanish Government energy policy. In this study, a similar trend is expected when querying the database for "WTI Crude Oil Price" related news. Indeed, there is a possible connection between the news activity related to the energy policies and the crude oil price. Moreover, the expectation of a negative association between the oil price and the GDELT activity is consistent with the hypotheses formulated for Twitter and Wikipedia. By contrast, the number of organizations mentioned with regard to the WTI Oil Price (in the news on GDELT) is expected to have a positive association with the price movements: enterprises are often linked to new projects, positive events or business agreements.

H4a: A higher number of articles on GDELT for the search query "WTI Crude Oil Price" is associated with oil price drops.

H4b: The number of organizations names mentioned on GDELT with regard to the search query "WTI Crude Oil Price" is positively related to the oil price movements.

3. Data collection and description of variables

The daily spot price of the WTI Crude Oil has been analyzed from April 1, 2013 to April 1, 2015. The choice of the WTI spot price has been made because of its availability and its common usage in the literature of oil price movements modeling and predictions. Price data have been obtained from the U.S. Energy Information Administration. In addition, ten independent variables representing the online traffic of the four social media platforms – Twitter, Wikipedia, Google Trends and the GDELT Project – have been collected. The aim has been to study the "oil price" search term from different resources to target different publics.

Twitter data have been collected by using the social network analysis software Condor, fetching all tweets that contained the search term "crude oil price". Collected tweets have been then filtered, excluding all non-relating posts – those referring, for instance, to cooking or lubricating oils. Analyzing the tweets, four Twitter variables have been determined, i.e. Number of tweets per day, Sentiment, Complexity and Emotionality.

For Wikipedia, the daily views count of the pages relating to Crude Oil Price, such as "Benchmark (crude oil)", "World oil market chronology", "2000s energy crisis", and others, has been collected. For the final study, two pages have been considered – "Price of oil" and "OPEC" – as their traffic statistics were the most significant in terms of correlation with the dependent variable.

An analogous procedure has been applied to gather Google Trends data. This data is available directly on the Google website, which allows filtering for location, time range, and keywords. The location has been set to "worldwide", whilst the queried keywords have been "Price of oil" and "OPEC". Other possible keyword combinations have been tested, obtaining less significant results.

GDELT data have been obtained by crawling the Global Knowledge Graph (GKG) dataset available on the project website. The daily number of newspaper articles covering out "WTI Crude Oil Price" search criterion and the count of all organizations names or advisory councils mentioned all over the world during the period of the study have been extracted. A more detailed description of all the independent variables is shown in Table 1.

[Table 1 about here]

4. Methodology

As discussed in Section 2, a wide body of literature has studied the oil price prediction using various methods, including, but not limited to, time series analysis, artificial neural networks, support vector machines, empirical mode decomposition, and wavelet transforms, in addition to the traditional linear and non-linear regressions, error corrections and econometric methods. Many of these studies present some limitations. For instance, they often consider weekly or monthly data in time series forecasts (e.g., Mohammadi and Su 2010; Ekinci and Erdal 2015), thus shrinking the dataset size, and not allowing a day by day prediction. Moreover, taking 300 observations on a monthly basis means going back 25 years, which means including in the study a different or an irrelevant economic situation comparing to the present one. Furthermore, it limits the ability of dealing with daily lead/lag. In addition, several studies base their forecasts on a set of observations on a single variable (e.g. Ahmed and Shabri 2014). Therefore, the output series is described in terms of past values of the input series. Forecasts from such models are therefore only extrapolations of the observed series (Harvey 1990).

In this context, the choice has been to implement both ARIMA and ARIMAX time series forecasting models, for two main reasons. First, when compared to other techniques – such as multiple linear regression models – time series models are usually superior in making daily forecasts, as they are built considering the autocorrelation function (ACF) and the partial autocorrelation function (PACF). Accordingly, including a differencing order, these models can achieve data stationary and uncover the presence of unit roots and trends (As'ad, 2012). Second, the ARIMA model produces forecasts based on past values in the time series (AR terms) and on the errors made by previous predictions (MA terms). Those parameters allow the model to auto-adjust efficiently upon sudden changes, resulting in higher accuracy. The ARIMAX procedure takes the ARIMA model one-step ahead, by including external factors or independent variables into the model (Andrews et al. 2013).

One concern has been dealing with the time series gaps. Missing values during weekends and public holidays have been excluded, while missing data during working days due to shutdowns or unknown reasons have been interpolated, in order to create a five days per week observation time. The lags were limited to a maximum of three days: indeed, the missing data – either in the Crude Oil Price during weekends and holidays or in the platform variables due to failures – returns inaccurate prediction results on higher lags and causes computational failures in financial econometric models. The data set has been trained on the period from April 1, 2013 to March 14, 2015 as sample data. The period from March 15, 2015 to April 1, 2015 has been used to test the models on real predictions, i.e. has been considered as the out-of-sample.

To evaluate and compare the accuracy of forecasts, the Root Mean Squared Error (RMSE) and the Mean Absolute Percentage Error (MAPE) of the predictive models have been determined.

5. Results

This section discusses the prediction ability of each platform in terms of how many days the WTI Crude Oil Price takes to react to the platform activity. In statistical words, the lagging of independent variables is held to predict what will happen in time t, based on the knowledge available at the time (t–n), where n is the number of lags. For instance, to predict the price value on Friday, first, second, and third order lags have been created. The first lag dataset uses the data available on Thursday, the second lag dataset refers to data available on Wednesday, and the third lag dataset refers to data available on Tuesday.

To further test the robustness of predictors, the study includes two control variables: the WTI Crude Oil Price of the previous day, and the NASDAQ100 index.

5.1 Correlation analysis

Pearson correlation analysis was used to test the association between the social awareness variables and the WTI Crude Oil Price on the first, second and third lags.

[Table 2 about here]

From the results in Table 2 we can draw the following considerations.

(1) Twitter variables, except for Emotionality, are significantly associated with the price. Sentiment and Complexity show a better correlation on the first lag, whilst the Number of tweets correlates the best at a three-day lag.

(2) Google Trends and Wikipedia variables are also highly correlated with the WTI Crude Oil Price, at all lags. Similarly, both platforms are consistent in showing that the search query "Price of Oil" returns higher value of correlation on the first lag, whilst the search query "OPEC" is higher on second and third lags. Google Trends variables correlate positively, whilst Wikipedia variables correlate negatively.

(3) The GDELT count of number of articles shows a much higher correlation to the WTI Crude Oil Price than the count of the number of organizations. The number of articles shows the best correlation on a two-day lag, whilst the number of organizations is correlating the best on a three-day lag.

(4) All the predictive variables are significantly correlated with the NASDAQ index – sometimes switching the direction of the association, when compared with the Crude Oil Price.

(5) Predictors from different social media sources are frequently correlated with one-another, indicating that real world events related to the oil price movements spread over different social media platforms simultaneously.

These findings initially support all the research hypotheses.

5.2 Granger causality tests and ARIMA/ARIMAX models

To check the predictive power of each of the independent variables, Granger causality analysis (Granger 1969) was implemented. Granger causality is a statistical test, which reveals if one time series provides useful information which help in forecasting another time series (Table 3). The number of observations is 372 and it refers to the number of days included in the observation period. Per each day, thousands of entries coming from the different platforms have been analyzed.

[Table 3 about here]

Similarly to ARIMA/ARIMAX, determining Granger causality requires the time series to be stationary. Since most of financial time series are subject to trends, seasonality and periodic variations, the stationarity hypothesis is probably rejected. In order to check the series stationarity, the augmented Dicky Fuller unit root test has been applied to all of the predictors, which have shown a non-stationary behavior and have required a first order differencing to achieve stationarity.

Table 3 shows a significant Granger correlation of Crude Oil Price with Twitter Emotionality and Complexity, Google Trends "price of oil" queries count, Wikipedia "price of oil" and "OPEC" page views count, and GDELT number of articles. Moreover, the lag orders reveal a lot about each platform and its users. Twitter Complexity is the most predictive on a one-day lag and it correlates negatively to the WTI price movements. Therefore, Twitter seems faster in reflecting the collective mind when compared to the other platforms, thus confirming the previous findings of Zhang and colleagues (2011). A less shared language – probably more technical and mostly used by specialized traders – can predict a price drop one day before it happens. Similarly, Twitter Emotionality – originally not correlated with the WTI Crude Oil Price – shows the largest predictive power at a one-day lag, after the differencing.

On the other hand, Google Trends and Wikipedia "Price of Oil" seem to require more time to be updated and are mostly effective at a three-day lag. From this perspective, the research findings are in contrast with other studies: some scholars claim that Google Trends is the best media source for almost real time forecasting (e.g. Giannone et. al. 2008; Chadwick and Sengü 2015). The article title "OPEC" – which is a more specialized term in respect to "price of oil" – is only predictive for Wikipedia, at a one-day lag. It could happen that some traders consult the OPEC page on Wikipedia before choosing whether to buy or not. Finally, the GDELT count of number of articles shows prediction ability only on a two-day lag.

Starting from Table 3 outcomes, multivariate ARIMAX models have been trained. Having considered the ACF, PACF, AIC and BIC tests of the WTI Crude Oil Price time series, the ARIMA(2,1,4) model has revealed to offer the best parameters. Model outcomes have been evaluated twice. Once, by comparing the AIC and BIC measures for in-sample fitted values. The second time, by means of RMSE and of MAPE, based on the out-of-sample prediction. Table 4 shows the final results.

[Table 4 about here]

Models are presented in consistent variables blocks. The split of Twitter variables (Models 2 and 3) has been necessary due to collinearity problems.

The ARIMAX models outperform the ARIMA models regardless of the platform used in prediction. In terms of RMSE and MAPE, the best ARIMA model shows forecast errors of 13.878 and 22.579, respectively; in comparison, the best ARIMAX model shows errors of forecast of 1.683 and 2.498, respectively.

In terms of out-of-sample evaluation, Google Trends emerges as the best time series predictor for the WTI Crude Oil Price, followed by Twitter. This result is consistent with other studies about crowds intelligence (e.g. Gloor and Cooper 2007; Al-Rifaie et al. 2010). Finally, Model 8 shows that combining Twitter Complexity, Wikipedia "price of oil" page views count and GDELT number of articles returns the best results for both in-sample and out-of-sample predictions. This final finding could suggest that complex language and statistics produced by experts of journalism or Wikipedia readers are the best candidates for predicting a complex issue like the movements of WTI Crude Oil Price. Model 8 graphical performance is presented in Figure 1.

[Figure 1 about here]

6. Discussion and conclusions

An accurate prediction of the crude oil price can be extremely important, since this price has direct effects on several goods and products – in terms of transportation and production costs, for instance – and its movements affect the stock markets. Oil prices are not only driven by economic variables, but they are also affected by key events that can create a state of awareness, such as political events, military conflicts, severe climate abnormalities and even big accidents (Fan et al. 2008). Global and local awareness is reflected on social media and can be measured tracking indicators like the sentiment of the language used or the frequency of interaction (Gloor et al. 2016). In this study, signals of global awareness are collected from four different media platforms (Twitter, Wikipedia, Google Trends and GDELT) and translated into 10 independent variables, with the intent of making more accurate predictions of the WTI Crude Oil Price movements. The main contribution of this approach is to test more media platforms together – when compared to past research (e.g. Choudhury et al. 2008; Asur and Huberman 2010; Granka 2013) – and to assess the predicting power of new variables.

The results coming from the models show that the social parameters – extracted from the above-mentioned platforms – have remarkably high correlations with the oil price movements. Consistently, significant granger-cause relations were identified for six out of ten predictors.

Finally, the study proves that a combined analysis of Twitter, Wikipedia and GDELT (see Table 4, Model 8) can lead to forecasts for crude oil prices with a reasonably high level of accuracy – thus supporting the advantage of integrating multiple data sources, instead of relying on a single media platform. In addition, it has been proved that real world events (related to oil price movements) reflect on different media platforms at different "speeds": Twitter is most informative at a one-day lag, GDELT and Wikipedia at a two-day lag and Google Trends at a three-day lag.

7. Research limitations and future work

This case study has several limitations, which the authors are willing to address with future research. The data collected are limited only to the feeds containing the keywords, defined as search terms. Therefore, some other feeds relating to the same subject but using other keywords or expressions might have been missed. For Twitter, the main limitation can be found in its structure. Powerful individuals – who have a big number of followers, a big retweet traffic on their tweets, or are mentioned in a big number of tweets – can lead others to engage in a certain act (Cha et al. 2010). Therefore, not all users should have the same weight in calculating either sentiment or activity metrics, and for future research the authors recommend applying different weights to different users. With regard to Wikipedia, the variables have been limited to the page views activity. Future works could consider several other variables, such as the number of users who have contributed to the page, the number of page edits, the time span between different versions of an article, the length and sentiment of each edit, and so on. Concerning Google Trends, the search engine itself is not able to represent all online activities related to the Oil Price. Probably, some traders, hedgers, and speculators skip the search engine, directly referring to specialized platforms. Lastly, regarding the GDELT Project, this study measures the volume of news articles based on keyword search and the count of organizations mentioned in that news. Future works could consider, for instance, the number of mentions, which counts the number of information sources containing mentions of the event.

Generally speaking, several extensions to this work are possible, while studying the signals of collective awareness coming from the internet world. The approach and the variables proposed in this study could be easily integrated in other existing models, to improve their predictive power. With a step in the direction of reading the collective mind, authors maintain the importance of extracting relevant knowledge from a large set of public available data sources.

8. References

Ahmed RA, Shabri AB (2014) Daily crude oil price forecasting model using arima, generalized autoregressive conditional heteroscedastic and support vector machines. American Journal of Applied Sciences 11(3):425-432.

Al-Rifaie MM, Bishop JM, Caines S (2012) Creativity and autonomy in swarm intelligence systems. Cognitive Computation 4(3):320-331.

Arias M, Arratia A, Xuriguera R (2013) Forecasting with twitter data. ACM Transactions on Intelligent Systems and Technology 5(1).

Andrews BH, Dean MD, Swain R, Cole C (2013) Building ARIMA and ARIMAX Models for Predicitng Long-Term Disability Benefit Application Rates in the Public/Private Sectors. Society of Actuaries Health Section. https://www.soa.org/Research/Research-Projects/Disability/research-2013-arima-arimax-ben-appl-rates.aspx.

As’ad M (2012) Finding the Best ARIMA Model to Forecast Daily Peak Electricity Demand. Proceedings of the Fifth Annual ASEARC Conference - Looking to the future - Programme and Proceedings, 2 - 3 February 2012, University of Wollongong.

Asur S, Huberman BA (2010) Predicting the future with social media. Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on, Canada, IEEE, 492-499, arXiv:1003.5699v1.

Baek H, Ahn J, Oh S (2014) Impact of Tweets on Box Office Revenue: Focusing on When Tweets are Written. ETRI Journal 36(4):581-590.

Banbura M, Giannone D, Reichlin L (2010) Nowcasting. CEPR Discussion Papers 7883, C.E.P.R. Discussion Papers, Available at http://ideas.repec.org/p/cpr/ceprdp/7883.html.

Behmiri NB, Pires Manso JR (2013) Crude oil price forecasting techniques: a comprehensive review of literature. CAIA Alternative Investment Analyst Review 2(3):30-48.

Bodas-Sagi DJ, Labeaga JM (2016) Using GDELT Data to Evaluate the Confidence on the Spanish Government Energy Policy. International Journal of Interactive Multimedia and Artificial Intelligence 3(6):38-43.

Bollen, J, Mao H, Pepe A (2011a) Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. ICWSM 11:450-453.

Bollen J, Mao H, Zeng X (2011b) Twitter mood predicts the stock market. Journal of Computational Science 2(1):1-8.

Box GE, Jenkins GM (1976) Time series analysis: forecasting and control. Holden-Day, San Francisco.

Brönnimann L (2013) Multilanguage sentiment-analysis of Twitter data on the example of Swiss politicians. Available at http://www.twitterpolitiker.ch/Paper_Swiss_Politicians_On_Twitter.pdf.

Brönnimann L (2014) Analyse der Verbreitung von Innovationen in sozialen Netzwerken. University of Applied Sciences and Arts Northwestern Switzerland, Available at http://www.twitterpolitiker.ch/Master_Thesis_Lucas_Broennimann.pdf.

Chadwick MG, Sengül G (2015) Nowcasting the Unemployment Rate in Turkey: Let’s Ask Google. Central Bank Review 15(3):15-40.

Cha M., Haddadi H, Benevenuto F, and Gummadi P (2010) Measuring User Influence in Twitter: The Million Follower Fallacy. Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media The AAAI Press Menlo Park, California 10-17.

Chen L, Wang W, Sheth AP (2012) Are Twitter users equal in predicting elections? A study of user groups in predicting 2012 US Republican Presidential Primaries. Proceedings of International Conference on Social Informatics Springer Berlin Heidelberg 379-392.

Chen R, Lazer M (201) Sentiment analysis of Twitter feeds for the prediction of stock market movement. Cs 229 1-5. ftp://cse.shirazu.ac.ir:5001/Tempbuffer/InGruheKhashen/ChenLazer-SentimentAnalysisOfTwitterFeedsForThePredictionOfStockMarketMovement.pdf.

Cheong CW (2009) Modeling and Forecasting Crude Oil Markets Using ARCH-type Models. Energ Policy 37:2346–2355.

Choi H, Varian H (2009a) Predicting the present with Google Trends. Technical report, Google. http://google.com/googleblogs/pdfs/google_predicting_the_present.pdf.

Choi H, Varian H (2009b) Predicting initial claims for unemployment benefits. http://static.googleusercontent.com/media/research.google.com/it//archive/papers/initialclaimsUS.pdf.

Choi H, Varian H (2012) Predicting the present with Google Trends. Econ Rec 88(s1):2-9.

Choudhury M, Sundaram H, John A, Seligmann D (2008) Can blog communication dynamics be correlated with stock market activity?. Proceedings of the nineteenth ACM conference on Hypertext and hypermedia 55-60.

Culotta A (2013) Lightweight methods to estimate influenza rates and alcohol sales volume from Twitter messages. Lang Resour Eval 47(1):217-238.

De Mauro A, Greco M, & Grimaldi M (2016) A Formal definition of Big Data based on its essential Features. Library Review 65(3):122-135.

Ekinci A, Erdal H (2015) Optimizing the monthly crude oil price forecasting accuracy via bagging ensemble models. Journal of Economics and International Finance 7(5):127-136.

Elekdag S, Lalonde R, Laxton D, Muir D (2008) Oil price movements and the global economy: A model-based assessment. IMF Economic Review 55(2):297-311.

Fan Y, Liang Q, Wei YM (2008) A generalized pattern matching approach for multi-step prediction of crude oil price. Energ Econ 30(3):889-904.

Fantazzini D, Toktamysova Z (2015) Forecasting German car sales using Google data and multivariate models. Int J Prod Econ 170:97-135.

Fernandez V (2007) Forecasting commodity prices by classification methods: The cases of crude oil and natural gas spot price. Banco Central De Chile Conference, July 27, 2007.

Forbergskog JO (2013) Twitter and stock returns. Doctoral dissertation, BI Norwegian Business School.

Fronzetti Colladon A, Remondi E (2017) Using Social Network Analysis to Prevent Money Laundering. Expert Syst Appl 67:49–58.

Gabrilovich E, Markovitch S (2007) Computing semantic relatedness using Wikipedia-based explicit semantic analysis. Proceedings of International Joint Conference on Artificial Intelligence 1606:1611. http://www.aaai.org/Papers/IJCAI/2007/IJCAI07-259.pdf.

Giannone D, Reichlin L, Small D (2008) Nowcasting: The real-time informational content of macroeconomic data. J Monetary Econ 55(4):665-676.

Gloor PA, Cooper S (2007) The new principles of a swarm business. MIT Sloan Manage Rev 48(3):81-+. http://apps.isiknowledge.com/full_record.do?product=WOS&search_mode=GeneralSearch&qid=1&SID=P2j93IKOkg62E4FIII9&page=3&doc=147.

Gloor PA, Fronzetti Colladon A, Miller C Z, Pellegrini R (2016) Measuring the Level of Global Awareness on Social Media. In M. Zylka, H. Führes, A. Fronzetti Colladon, & P. A. Gloor (Eds.), Designing Networks for Innovation and Improvisation 125:139. Cham, Switzerland: Springer International Publishing.

Gloor PA, Marcos J, De Boer PM, Fuehres H, Lo W, Nemoto K (2015) Cultural anthropology through the lens of Wikipedia: Historical leader networks, gender bias, and news-based sentiment. arXiv preprint 1:22.

Go A, Huang L, Bhayani R (2009) Twitter sentiment analysis. Entropy 17.

Granger CW (1969) Investigating causal relations by econometric models and cross-spectral methods. Econometrica: Journal of the Econometric Society 37(3):424-438.

Granka L (2013) Using Online Search Traffic to Predict US Presidential Elections. PS: Political Science and Politics 46(1):271-279. http://journals.cambridge.org/abstract_S1049096513000292\npapers3://publication/uuid/88611D8E-07DD-4480-8EB8-D691AC5D0978.

Harvey AC (1990) Forecasting, structural time series models and the Kalman filter. Cambridge university press. Cambridge, UK.

Hermida A, Lewis SC, Zamith R (2014) Sourcing the arab spring: a case study of Andy Carvin's sources on twitter during the Tunisian and Egyptian revolutions. J Comput-Mediat Comm 19(3):479-499.

Javanmardi S, Ganjisaffar Y, Lopes C, Baldi P (2009) User contribution and trust in Wikipedia. Proceedings of the 5th International Conference on Collaborative Computing: Networking, Applications and Worksharing, IEEE, 1-6.

Kaplan AM, Haenlein M (2010) Users of the world, unite! The challenges and opportunities of Social Media. Bus Horizons 53(1):59-68.

Koop G, Onorante L (2013) Macroeconomic nowcasting using Google probabilities. Working paper, University of Strathclyde, August 2013.

Kouloumpis E, Wilson T, Moore JD (2011) Twitter sentiment analysis: The good the bad and the omg!. Proceedings of the Fifth International AAAI

Conference on Weblogs and Social Media, The AAAI Press, Menlo Park, California 538-541.

Kwak H, An J (2014) Understanding news geography and major determinants of global news coverage of disasters. arXiv preprint arXiv:1410.3710.

Kwak H, An J (2016) Two Tales of the World: Comparison of Widely Used World News Datasets GDELT and EventRegistry. arXiv preprint arXiv:1603.01979.

Leetaru K, Schrodt PA (2013) Gdelt: Global data on events, location, and tone, 1979–2012. Proceedings of ISA Annual Convention (Vol. 2, No. 4), Available at http://data.gdeltproject.org/documentation/ISA.2013.GDELT.pdf.

Mestyán M, Yasseri T, Kertész J (2013) Early prediction of movie box office success based on Wikipedia activity big data. PloS one 8(8) e71226. doi:10.1371/journal.pone.0071226.

Milne D, Witten IH (2008) Learning to link with wikipedia. Proceedings of the 17th ACM conference on Information and knowledge management, ACM, pp. 509-518, Available at http://dl.acm.org/citation.cfm?id=1458082.1458150.

Mittal A, Goel A (2012) Stock prediction using twitter sentiment analysis. Standford University CS229. http://cs229.stanford.edu/proj2011/GoelMittal-StockMarketPredictionUsingTwitterSentimentAnalysis.pdf.

Moat HS, Curme C, Avakian A, Kenett DY, Stanley HE, Preis T (2013) Quantifying Wikipedia usage patterns before stock market moves. Scientific reports 3.

Mohammadi H, Su L (2010) International evidence on crude oil price dynamics: Applications of ARIMA-GARCH models. Energ Econ 32(5):1001-1008.

Moshiri S, Foroutan F (2006) Forecasting nonlinear crude oil futures prices. The Energy Journal 27(4):81-95.

Oh O, Agrawal M, Rao HR (2011) Information control and terrorism: Tracking the Mumbai terrorist attack through twitter. Inform Syst Front 13(1):33-43.

Onn E, Morewedge CK (2009) Emotionality in Text as Predictor of Behavior. Proceedings of the Third International ICWSM Conference, The AAAI Press, Menlo Park, California, pp. 286-287.

Pak A, Paroubek P (2010) Twitter as a Corpus for Sentiment Analysis and Opinion Mining. Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10), European Language Resources Association (ELRA) 1320-1326.

Pavlicek J, Kristoufek L (2015) Nowcasting unemployment rates with Google searches: Evidence from the Visegrad Group countries. PloS one 10(5) e0127084.

Phua C, Feng Y, Ji J, Soh T (2014) Visual and Predictive Analytics on Singapore News: Experiments on GDELT, Wikipedia, and STI. arXiv preprint arXiv:1404.1996.

Porshnev A, Redkin I, Karpov N (2014) August. Modelling Movement of Stock Market Indexes with Data from Emoticons of Twitter Users. In Russian Summer School in Information Retrieval 297-306. Springer International Publishing.

Rao T, Srivastava S (2013) Modeling movements in oil, gold, forex and market indices using search volume index and Twitter sentiments. Proceedings of the 5th Annual ACM Web Science Conference, ACM, pp. 336-345.

Rui H, Liu Y, Whinston A (2013) Whose and what chatter matters? The effect of tweets on movie sales.Decis Support Syst 55(4):863-870.

Sajid Nazir M, Younus H, Kaleem A, Anwar Z (2014) Impact of political events on stock market returns: empirical evidence. Pakistan.Journal of Economic and Administrative Sciences 30(1):60-78.

Schoen H, Gayo-Avello D, Takis Metaxas P, Mustafaraj E (2013) The power of prediction with social media. Internet Res 23(5):528-543.

Scott SL, Varian H (2013) Bayesian variable selection for nowcasting economic time series. National Bureau of Economic Research No. w19567.

Shambora W, Rossiter R (2007) Are There Exploitable Inefficiencies in the Futures Market for Oil?. Energ Econ 29:18–27.

Sprenger TO, Tumasjan A, Sandner PG, Welpe IM (2014) Tweets and trades: The information content of stock microblogs. Eur Financ Manag 20(5):926-957.

Staub T, Hodel T (2016) Wikipedia vs. Academia: An Investigation into the Role of the Internet in Education, with a Special Focus on Wikipedia. Universal Journal of Educational Research 4(2):349-354.

Stephens-Davidowitz S, Varian H (2014) A Hands-on Guide to Google Data. Technical Report, Google.

Surowiecki J (2005) The wisdom of crowds. Anchor.

Torban FB (2010) Models for oil price prediction and forecasting. Dissertation, San Diego State University.

Tumasjan A, Sprenger TO, Sandner PG, Welpe IM (2010) Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment. Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media, The AAAI Press, Menlo Park, California, pp. 178-185.

Van der Haak B, Parks M, Castells M (2012) The future of journalism: Networked journalism. International Journal of Communication 6(16):2923-2938.

Vicente MR, López-Menéndez AJ, Pérez R (2015) Forecasting unemployment with internet search data: Does it help to improve predictions when job destruction is skyrocketing?. Technol Forecast Soc 92:132-139.

Wang SY, Yu L, Lai KK (2004) A Novel Hybrid AI System Framework for Crude Oil Price Forecasting. Lect Notes Comput Sc 3327:233– 242.

Wei P, Wang N (2016) Wikipedia and Stock Return: Wikipedia Usage Pattern Helps to Predict the Individual Stock Movement. Proceedings of the 25th International Conference Companion on World Wide Web, pp. 591-594, Available at http://www2016.net/proceedings/companion/p591.pdf.

Wei Y, Wang Y, Huang D (2010) Forecasting Crude Oil Market Volatility: Further Evidence Using GARCH-class Models. Energ Econ 32:1477–1484.

Wu L, Brynjolfsson E (2013) The future of prediction: How Google searches foreshadow housing prices and sales. SSRN 2022293. http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2022293.

Yasseri T, Bright J (2015) Predicting elections from online information flows: towards theoretically informed models. arXiv preprint arXiv:1505.01818.

Yonamine JE (2013) Predicting Future Levels of Violence in Afghanistan Districts Using GDELT. http://data.gdeltproject.org/documentation/Predicting-Future-Levels-of-Violence-in-Afghanistan-Districts-using-GDELT.pdf. Accessed 12 April 2013.

Yu L, Wang SY, Lai KK (2005) A Roughset-refined Text Mining Approach for Crude Oil Market Tendency Forecasting. International Journal of Knowledge and Systems Sciences 2:33–46.

Zappavigna M (2012) Discourse of Twitter and social media: How we use language to create affiliation on the web. A&C Black.

Zhang X, Fuehres H, Gloor PA (2011) Predicting stock market indicators through twitter “I hope it is not as bad as I fear”. Procedia-Social and Behavioral Sciences 26:55-62.

Zhang X, Fuehres H, Gloor PA (2012) Predicting asset value through twitter buzz. In Altman, J., Baumöl, U. and Krämer, B. (Eds.), Advances in Collective Intelligence 2011, Springer Berlin Heidelberg 23-34.

[1] https://about.twitter.com/company

[2] https://blog.twitter.com/2012/twitter-turns-six

[3] https://cloudplatform.googleblog.com/2014/05/worlds-largest-event-dataset-now-publicly-available-in-google-bigquery.html

Platform	***Variable***	***Retrieved By***	***Description***
Twitter	Messages per Day	Social Network Analysis software "Condor"	It is the count of the number of tweets written in the English language, which contains the search query on a given day.
	Sentiment		Sentiment varies in the range [0,1], where a higher score represents a more positive language with respect to the query term. The calculus was made by means of machine learning algorithms included in the software Condor (Brönnimann, 2014).
	Emotionality		Emotionality is calculated as the deviation from neutral sentiment (Brönnimann, 2014).
	Complexity		Text complexity is measured considering the occurrence of single words in the whole text: less used words have higher complexity than the other commonly used words. We then calculate text complexity as suggested by (Brönnimann, 2014), considering the Inverse Document Frequency (IDF), which reflects how important a word is to a document or a corpus.
Wikipedia	Page Visits	Web Crawling	A count, which refers to how many times a certain article has been visited during a specific timeframe. The measure was considered twice for two different articles, "Price of Oil" and "OPEC".
Google Trends	Query Count		A count which shows how often a particular keyword or query is searched in relation to the total search volume in a certain geographical area at a certain time period. The measure was considered twice for two queries, "Price of Oil" and "OPEC".
GDELT	Number of Articles	Query of the Global Knowledge Graph (GKG) Dataset	A numerical quantification, which measures the volume of news coverage nominating the "WTI Crude Oil Price" as the main event in a particular geographical location.
	Number of Organizations		Organizations mentioned in articles matching the searching criteria. It includes multiple organizations names and councils, expressing both non-profit and commercial enterprises.

Table 1: Description of independent variables.

	Variable	***1***	***2***	***3***	***4***	***5***	***6***	***7***	***8***	***9***	***10***	***11***	***12***	***13***
1	WTI Crude Oil Price (No lag)	1												
2	WTI Crude Oil Price (*t*–1)	.967*	1											
3	Nasdaq 100 (*t*–1)	-.688*	-.579*	1										
4	Twitter Messages (*t*–1)	-.508*	-.479*	.437*	1									
5	Twitter Sentiment (*t*–1)	.503*	.565*	-.391*	-.385*	1								
6	Twitter Emotionality (*t*–1)	-.097	.058	.424*	.161*	-.014	1							
7	Twitter Complexity (*t*–1)	-.280*	-.097	.415*	.305*	-.022	.685*	1						
8	Google Trends “OPEC” Query Count (*t*–1)	.617*	.614*	-.632*	-.286*	.461*	.003	-.081	1					
9	Google Trends “Price of Oil” Query Count (*t*–1)	.359*	.384*	-.325*	-.095	.336*	-.034	-.007	.521*	1				
10	Wikipedia “OPEC” Page Views (*t*–1)	-.562*	-.526*	.346*	.412*	-.387*	.167*	.301*	-.140*	-.090	1			
11	Wikipedia “Price of Oil” Page Views (*t*–1)	-.785*	-.762*	.505*	.498*	-.447*	.040	.197*	-.435*	.005	.688*	1		
12	GDELT Number of Organizations (*t*–1)	.267*	.245*	-.184*	-.134*	.228*	.029	-.188*	.301*	.124*	-.254*	-.258*	1	
13	GDELT Number of Articles (*t*–1)	-.717*	-.688*	.586*	.441*	-.485*	.100	.224*	-.485*	-.225*	.560*	.653*	-.273*	1
1	WTI Crude Oil Price (No lag)	1												
2	WTI Crude Oil Price (*t*–2)	.938*	1											
3	Nasdaq 100 (*t*–2)	-.657*	-.457*	1										
4	Twitter Messages (*t*–2)	-.515*	-.457*	.439*	1									

5	Twitter Sentiment (*t*–2)	.479*	.590*	-.252*	-.347*	1								
6	Twitter Emotionality (*t*–2)	-.090	.165*	.520*	.177*	.147*	1							
7	Twitter Complexity (*t*–2)	-.245*	.044	.517*	.292*	.150*	.784*	1						
8	Google Trends “OPEC” Query Count (*t*–2)	.618*	.613*	-.566*	-.282*	.461*	.044	-.024	1					
9	Google Trends “Price of Oil” Query Count (*t*–2)	.357*	.401*	-.244*	-.084	.364*	.062	.088	.522*	1				
10	Wikipedia “OPEC” Page Views (*t*–2)	-.567*	-.496*	.354*	.416*	-.344*	.187*	.295*	-.133*	-.075	1			
11	Wikipedia “Price of Oil” Page Views (*t*–2)	-.785*	-.740*	.483*	.500*	-.420*	.046	.175*	-.432*	.011	.688*	1		
12	GDELT Number of Organizations (*t*–2)	.274*	.225*	-.194*	-.139*	.198*	-.011	-.191*	.294*	.111*	-.258*	-.259*	1	
13	GDELT Number of Articles (*t*–2)	-.720*	-.664*	.568*	.443*	-.450*	.111*	.211*	-.480*	-.215*	.562*	.653*	-.275*	1
1	WTI Crude Oil Price (No lag)	1												
2	WTI Crude Oil Price (*t*–3)	.910*	1											
3	Nasdaq 100 (*t*–3)	-.630*	-.353*	1										
4	Twitter Messages (*t*–3)	-.521*	-.436*	.443*	1									
5	Twitter Sentiment (*t*–3)	.458*	.613*	-.139*	-.312*	1								
6	Twitter Emotionality (*t*–3)	-.085	.243*	.587*	.193*	.256*	1							
7	Twitter Complexity (*t*–3)	-.222*	.142*	.586*	.291*	.263*	.835*	1						
8	Google Trends “OPEC” Query Count (*t*–3)	.617*	.611*	-.509*	-.278*	.462*	.075	.016	1					
9	Google Trends “Price of Oil” Query Count (*t*–3)	.354*	.417*	-.173*	-.073	.389*	.131*	.156*	.524*	1				
10	Wikipedia “OPEC” Page Views (*t*–3)	-.576*	-.469*	.363*	.420*	-.306*	.205*	.298*	-.127*	-.061	1			

11	Wikipedia “Price of Oil” Page Views (*t*–3)	-.785*	-.720*	.464*	.502*	-.397*	.052	.163*	-.428*	.016	.688*	1		
12	GDELT Number of Organizations (*t*–3)	.279*	.208*	-.205*	-.144*	.172*	-.040	-.197*	.289*	.099	-.262*	-.260*	1	
13	GDELT Number of Articles (*t*–3)	-.716*	-.636*	.551*	.451*	-.418*	.120*	.208*	-.472*	-.199*	.566*	.654*	-.279*	1

*p<0.05; (*t*–1) One-day lag; (*t*–2) Two-day lag; (*t*–3) Three-day lag.

Table 2: Correlation coefficients.

Variable	*First Lag*	*Second Lag*	*Third Lag*
	(χ^2)	(χ^2)	(χ^2)
Nasdaq 100	12.69*	9.605*	8.702*
Twitter Messages	1.016	1.306	1.959
Twitter Sentiment	.339	.296	2.380
Twitter Emotionality	9.920*	8.452*	7.061*
Twitter Complexity	12.32*	1.976	4.606
Google Trends “OPEC” Query Count	.275	1.356	.0315
Google Trends “Price of Oil” Query Count	6.851*	8.824*	12.522*
Wikipedia “OPEC” Page Views	7.010*	14.047*	.5586
Wikipedia “Price of Oil” Page Views	1.204	8.277*	11.153*
GDELT Number of Organizations	.723	.194	.915
GDELT Number of Articles	3.076	4.355*	5.348

N = 372; *$p<0.05$.

Table 3: Granger causality test results

Variable	***ARIMA***	***ARIMAX***						
	Model 1	**Model 2**	**Model 3**	**Model 4**	**Model 5**	**Model 6**	**Model 7**	**Model 8**
Nasdaq 100 (*t*–1)		-.001						-.002
Twitter Messages (*t*–3)			-.001					
Twitter Sentiment (*t*–3)			-1.481					
Twitter Emotionality (*t*–1)			-7.524					
Twitter Complexity (*t*–1)				-.689*				-.810*
Google Trends "OPEC" Query Count (*t*–2)					.002			
Google Trends "Price of Oil" Query Count (*t*–3)					.017*			
Wikipedia "OPEC" Page Views (*t*–2)						-3.36E-05		
Wikipedia "Price of Oil" Page Views (*t*–3)						.001*		.001*
GDELT Number of Organizations (*t*–3)							-.018	
GDELT Number of Articles (*t*–2)							-.004*	-.004*
N	372	372	372	372	372	372	372	372
AIC (in-sample)	1323.306	1274.454	1230.390	1252.675	1226.700	1219.926	1256.149	**1218.428**
BIC (in-sample)	1354.614	1305.431	1268.679	1283.493	1261.160	**1254.387**	1290.819	1260.546
RMSE (out-of-sample)	13.878	1.887	1.810	2.001	1.752	1.970	1.833	**1.683**
MAPE (out-of-sample)	22.579	2.898	2.762	3.044	2.674	3.016	2.794	**2.498**

*p<0.05; (*t*–1) One-day lag; (*t*–2) Two-day lag; (*t*–3) Three-day lag.

Table 4: Evaluation of ARIMA/ARIMAX(2,1,4) model