CS555 Term Project

Sentiment Analysis of GDELT Dataset For Forecasting Crude Oil Price

Xiang Li

Grant Charlton

Department of Computer Science

Colorado State University

November 28, 2017

Introduction

Crude oil is a commodity that is used in nearly all places in the world. In 2015, there were 220 countries or regions that reported oil consumption numbers, according to the US Energy Information Administration [1]. The number of countries, worldwide, registered with the US Department of State, is 195, as of January 2017 [2]. This shows that oil is a currently a desirable resource to any nation in the world.

In recent decades, crude oil supplies and reserves have fluctuated due to many reasons; some which are to be expected, such as hurricanes and other natural disasters, and others that were not foreseen by many experts, such as the continued rise in crude oil production in the US. Demand for crude oil also fluctuates, and in the years leading up to the 2008 economic crisis, for example, the demand rose during a period of sustained economic growth, and then fell along with the economic turndown [1].

Due to the far-reaching implications and high demand for it as a finite resource, the idea of being able to predict the oil price accurately is certainly enticing. Knowing the future price of crude oil is of much interest, financially speaking, to many investors, from private individuals to large firms. Due to its nature as a commodity that is consistently in widespread high demand in many industries, knowing the future price of oil is also valuable in terms of being a key indicator of the health of many other aspects of the global economy. It has a direct effect on the cost of transportation, imports, exports, and has a far-reaching impact on many stock prices and even whole industries [3].

Many factors influence the price of crude oil, and predicting its trends accurately has proven to be one of the most elusive goals for forecasters worldwide [3]. Certain patterns were known to be evident in historical oil price trends, providing indications that war, and political instability may have been behind periods of crude oil price volatility [4]. Despite much research and some past success, accurately predicting the crude oil price continues to be a challenge.

Among the factors that influence crude oil supply and demand, and therefore price, are major geopolitical events and civil unrest [5], [6]. Appeal, Refuse, Protest, and military-related activity are just some of the actions that can lead to social unrest [7], and subsequently have a negative effect on the capability of a nation to reliably produce or move oil supplies. Tracking and analyzing the impact of these types of geopolitical events used to require manual collection of data from various information sources around the world. It could not be collected, digested, or analyzed quickly. That situation is changing, with more free, open datasets becoming publicly available all the time.

The GDELT (Global Data on Events, Location, and Tone) project is a free, open database that is a publicly available dataset. It has metadata for all broadcast, web, and print news events published since 1979, and is updated every 15 minutes. It gives programmatic access to nearly everything that was happening around the world for approximately the last 14,000 days (and counting). The increasing abundance and availability of this data online provides potential indicators of changes in the future oil price that have not been fully explored [3].

This paper intends to investigate the possibility of a correlation between the event data available in GDELT datasets, and a subsequent predictable change in the West Texas Intermediate (WTI) crude oil price.

Problem Characterization

There is a plethora of (since 1979) historical data available on GDELT, most of which is well populated. Each GDELT event record has 58 fields, containing various kinds of information and metadata to describe that event. The information that was used to calculate and generate our prediction model was obtained data from the SQLDate, EventRootCode, EventCode, AvgTone, NumSources, and GoldsteinScale fields for each respective GLOBALEVENTID.

Likewise, there is corresponding crude oil price history available for the WTI price index that goes back decades. This is the price index data that was used to train our model, and to test the prediction accuracy. The WTI price was chosen due to its unpredictable nature, and the nonlinear deterministic process underlying the data series, which could allow opportunities for accurate short term predictions [3], [8].

Looking more closely at the GDELT event records, it can be seen that the events are all assigned an event code, which corresponds to a particular code in the Conflict and Mediation Event Observations (CAMEO) code reference [9]. Each event code can indicate a specific type of event or occurrence, from fairly broad base or root categories such as “Appeal”, “Demand”, or “Protest”, to more specific codes to indicate, for example, economic, military, or political connotations. There are certain event patterns that have been shown to indicate a progression towards a situation that qualifies as political or social unrest [7]. The events with codes or event root codes that belong to these patterns are targeted during our analysis and the rest of the events will be filtered from the GDELT sample data used.

Each event in the GDELT dataset also has metadata associated with it, and this metadata can be analyzed and used to determine the amplitude of a positive or negative impact on the general sentiment of society. Evaluating the frequency and size of the impact of these events can help to find relationships between the data and predictable trends in the crude oil price.

One piece of metadata that our model takes into account, and is available for each event is the field named AvgTone. The "tone" of an event in the GDELT dataset is a measure of the average positive or negative impact for all mentions and across all sources containing mentions of that event. For this field, each event is assigned a value between + and – 100, representing its average impact on sentiment for society. However, the value for AvgTone does not take into account the number of mentions, number of sources, or number of articles that the event appears in. This leaves room for our software to determine just how important those pieces of the metadata are, and how heavily they should influence the prediction model.

Using the CAMEO event code, combined with each event's value for tone, and weighting the events based on other metadata, a positive or negative value for each event can be calculated. We chose to obtain events for all country codes, since the forces of supply and more specifically demand are spread among all countries of the world. The number of sources for each event describes how widespread the mentions of this event are, so that value also contributes to our calculation for each event's impact on sentiment.

The last field used in our calculations is known as the Goldstein Scale. "Each CAMEO event code is assigned a numeric score from -10 to +10, capturing the theoretical potential impact that type of event will have on the stability of a country." [10] The Goldstein Scale is based on the event type, and does not distinguish between the sizes and scales of different events, nor does it account for other specifics of what took place during that particular event. It is helpful to consider it in our prediction model, since it measures the potential impact of an event on a country's stability, rather than the positive or negative tone that is being measured in the AvgTone field [10].

For this paper, the WTI crude oil price index is used for determining correctness when training and testing the model. We expect there to be some lag between the event dates and their effect on the oil prices. The timespan of the delay will be observed for our prediction model. Studies with a few similarities have found it to be most accurate on a 2 or 3 day delay, depending on circumstances [3].

Dominant Approaches to the Problem

There are a couple different aspects to this problem that have been pursued both separately and together by previous research work. The first is the prediction of social unrest trends based on social, economic, and political event analysis, using a database such as the global event data from GDELT [5]. The second is trying to associate the social unrest predictions to changes in the crude oil price.

The research that has used event data available in the GDELT dataset to predict social unrest, or other similar social trends, has used some similar measures as our model, taking into account fields like AvgTone and GoldsteinScale score for each event. Some researchers have attempted to use some additional values such as NumArticles, daily event counts and the intensity of events [3], [7], [11]. There have also been local and regional analyses done using event data from GDELT and the field data that is related to location by latitude/longitude, or by using the country code where the event is located to determine an expected impact for each event [5].

The effort to predict oil price by observing some measure of social unrest has been an ongoing process since before GDELT and other event databases existed [4]. There is still no consensus as to the best way to accomplish this. Many prediction models have been tested, and most often there is some success to report in the findings, but generally the success has been accompanied by some unexpected findings, or other indications of flawed hypotheses [3]. It is possible that the baseline data used for comparison is different for each research team, making prediction accuracy difficult to assess. In the various prior works, ideas for prediction models have been applied using wide ranging variables, another factor making it difficult to make a strict comparison between models. It has also been seen that different event datasets, modeled with different delays in the expected reaction of the oil market price, can be different from one study to the next, possibly due to the same factors [3].

Research combining GDELT sentiment analysis and crude oil price prediction has been done in a variety of ways as well. Trend prediction in these cases takes into account event data, and uses different models such as Artificial Neural Networks, Logistic Regression, Support Vector Machine, Decision Tree, and LASSO based logistic regression with varying results. Measuring the accuracy of the forecasts was done by comparing the predictions to the WTI oil price, and measuring mean squared error, mean absolute error, and root mean square error. Among these techniques, Artificial Neural Networks had the best results based on all three measures when evaluated using the models implemented in one study done by Moshiri and Foroutan [8]. In another study using a combination of data sources to test event data from multiple media platforms together, results showed high correlations with oil price movements. It was found that among the media platforms it evaluated, each was more accurate at a different lag, possibly indicating the speed at which sentiment is reflected on the different platforms. In this measure, it was found that GDELT had its best results at a 2-day lag [3].

Results of this prior research study where multiple event data sources were pulled from have indicated that perhaps Twitter, Google Trends, or Wikipedia page view counts may provide better accuracy when evaluating real-time event data when compared to the GDELT event dataset [3]. Other studies have shown that perhaps the link between social unrest and changes in crude oil price is not as strong as previously thought. The data available for years prior to 2000 shows a much more pronounced reaction by oil price to major social or political unrest events. It

is more difficult to locate a similar trend since 2000, perhaps due to the consistent rise of global demand for crude oil drowning out the signals from social unrest [4].

Methodology

Interacting with a dataset on the scale of the GDELT event database requires more resources than most users have access to. Storing, viewing, and processing data at this size would be impossible if it required users to download the data before interacting with it. For this reason, the GDELT project has made available several tools to query, filter, and analyze the datasets that are available. Some of these tools include visualization, format translation, data export, and querying functionality. One of these tools, Google BigQuery, a database designed for extremely large datasets like GDELT, was used for all of our queries to GDELT [12].

For determining which events were more important to our model, we used the CAMEO code book in combination with evaluating previous research techniques. We chose to focus on events that were related to economic, military, or policy for categories Appeal, Yield, Disapprove, Reject, and Protest, as they have shown to be event types that lead to social unrest [7]. Events that occur in nations with large oil production capacity were not singled out, since it has been shown that a cartel, even one the size of OPEC, does not have control over oil prices [6]. We included events from all country codes, to allow our model to account for global unrest. A nation does not have to be an oil producer for the unrest within its borders to affect oil price [4].

The schema descriptions available for the fields in the GDELT event table led us to incorporate the values from the fields SQLDate, EventCode, EventRootCode, NumSources, GoldsteinScale, and AvgTone into our prediction model [10]. Other fields such as NumArticles and NumMentions were considered, but the relationship between the values for those fields and the AvgTone or GoldsteinScale values were not as strong as NumSources for weighting the impact of individual events in our testing. Other tables in the GDELT project were considered as well, specifically the EventMentions table and the Global Knowledge Graph (GKG). They hold more granular information for each mention of a particular event, but for the purpose of detecting unrest it was decided that they were not able to provide anything additional that was useful above and beyond the data available in the primary full event table that we considered.

Using the Hadoop Distributed File System (HDFS) as our distributed file system, we have pulled data from GDELT using BigQuery and stored it on our system. For training and testing our model against the actual oil price, historical crude oil pricing data was obtained from the St. Louis Federal Reserve Economic Data (FRED) [13]. All data was stored in the local system as

comma separated values (.csv) files. To process and analyze the data, we have used HDFS and Spark, a processing engine that is compatible with Hadoop data [14]. Event data was pulled from GDELT for the last 5 years, and stored locally in our system.

For transforming, analyzing, training, and testing with the data, all code was written in Scala, the natively supported language for Apache Spark, since it offers the most efficient implementations of Spark's API's. Spark has components that allow for ease of use with Hadoop, such as libraries and data structure for querying and pulling data (DataFrames), and libraries for machine learning (Spark MLlib) [14]. Using the criteria for the event types and field information that was desired, a query written using SQL for BigQuery was used for obtaining GDELT event data.

After all our GDELT event data was pulled into HDFS, first we did some pre-processing of the data. During development of the software, we tested many different combinations of query parameters and formulas before arriving at our final data analysis strategy. Some of the early attempts to collect a meaningful score from the GDELT event data and the values in the various fields involved scaling the score for each event differently. First it was scaled directly proportionally with the NumMentions value, and also with a ratio between the number of mentions and the values for the NumSources and NumArticles fields. The data still all needed to be normalized for the ratio to affect the scaling and final score properly. We observed some scoring on sample queries of the GDELT event data. We were seeing results that showed sentiment scores for some events that were not being scored high enough or vice versa, being scored too low.

In an attempt to isolate the values we obtained from the different sentiment score fields, we made changes to our strategy and scaled the AvgTone and GoldsteinScale values separately. At this point, however, they still required more manipulation, since the GoldsteinScale is a value from -10 to +10, and the AvgTone is a value between -100 and +100. To alleviate this, we normalized the resulting data output for each score, and applied a coefficient to each. This coefficient could be adjusted as needed for either portion of the score. This is what allowed room for our software to fit a model that showed predictable changes in the price of crude oil following certain observations in the event data.

Additional streamlining of the query was done, for efficiency and to maximize effectiveness of our scoring formulas. We noticed that, as we observed more sample queries of the event dataset, the values for fields NumMentions and NumArticles were almost always very nearly equivalent, especially for events with very high numbers of mentions. Events with a lower number of mentions had a greater chance of exhibiting a high ratio of mentions to articles, and that could skew the score for an event with a small number of mentions to be artificially high. For this reason, we removed the NumArticles and NumMentions ratio from our formula.

After carefully evaluating our sample data, we distilled the formula down further, removing the scaling by number of mentions. This was done since the number of sources was emerging as the more accurate representation of the characteristic we wanted to capture. It represented more strictly the degree of distribution for this particular event. The final formula had two resulting scores that could be fed into Spark's MLlib library along with the WTI oil price data, and then the training and testing could begin in Spark.

The values for GoldsteinScale and AvgTone were grouped by date, and then the sum of each was taken. This data needed to be normalized before fitting to our model, so it was scaled by the number of sources for each value divided by the total number of sources for each day. Applying this to the event data yielded a "score" value for each day as follows. A score for the AvgTone and a score for the GoldsteinScale were both calculated.

$$AvgToneScore = Sum(AvgTone * NumSources/TotalSources)$$

$$GoldsteinScaleScore = Sum(GoldsteinScale * NumSources/TotalSources)$$

The Spark MLlib library offers several algorithms for fitting models to data. We have used a Linear Regression to fit a prediction model based on the WTI oil price data and the corresponding event data. Essentially the linear regression was given a set of feature values x and a label y. MLLib accepts multiple values per feature, so for our purposes, our two scores, the AvgTone score and the GoldsteinScale score, were used as the values (x), and our oil price data was used as our label (y).

The data was then converted into DataFrames datasets, a distributed collection of data that is available in Spark, and is similar to RDDs. It can be thought of in the same way as a table in a relational database, organized with named columns [14]. DataFrames can offer a higher degree of efficiency than basic RDDs when the schema of the data being worked with is known ahead of time. Since DataFrames have additional metadata due to its tabular format, iw allows Spark to run certain optimizations on the finalized query. Using DataFrames more highly informs Spark about the compute operations being done, and about the data structure being worked upon [14]. The final data format for each row after joining and filtering all the csv files from all the sources was

$$Data[row] = (oil\ price, [SumAvgToneScore, SumGoldsteinScaleScore])$$

This final data was divided into training and testing data using a random split with 80% dedicated to training, and 20% reserved for testing. A random split was used so that the training and testing datasets would be dispersed evenly over events from the same period of time. The model was trained using those datasets until it converged on stable values for variables m and b in the linear formula $y = mx + b$. After the training process was complete, the testing process

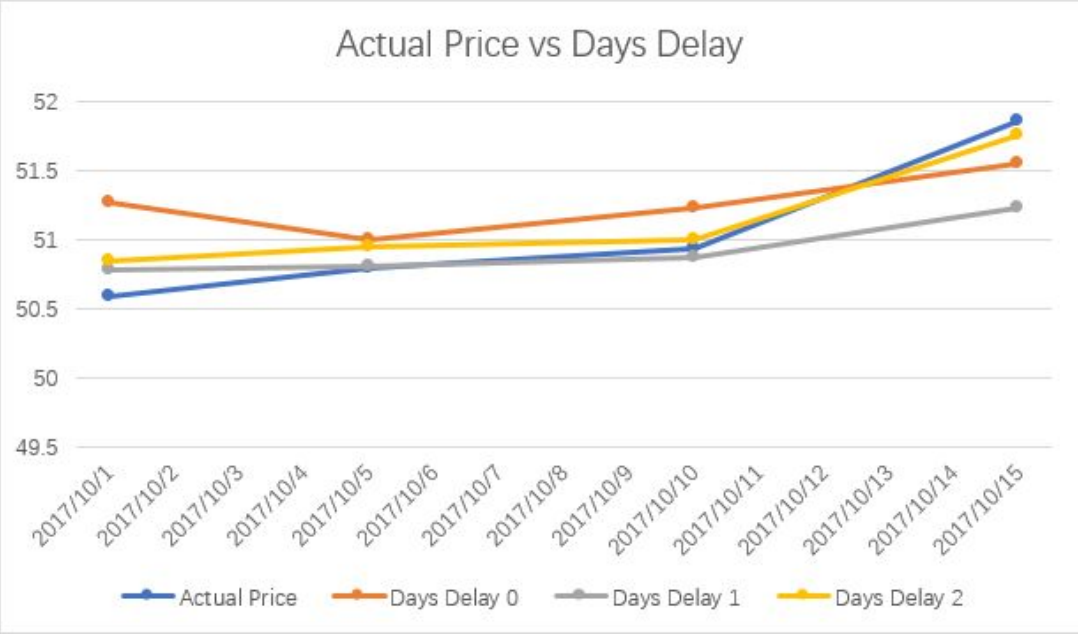
began. Using the linear formula obtained in the training, the prediction model was against the testing portion of the GDELT event data.

The prediction model was then used to generate a predicted value for the oil price on the (previously unseen to the model) testing data. To measure performance in terms of prediction accuracy, the predicted price was compared to the actual price and Root Mean Squared Error (RMSE) was used to evaluate correctness.

Experimental Benchmarks

To assess the performance of our solution and decide which method should be used, we mainly evaluate three parts: days to delay in order to match with the actual oil price, regression model choice and difference measurement formula choice.

After we finished building the system, we simply changed the matched date in join function between oil price date and event date. By keep adding 1 more days in the original data, we got the following graph. Among all three days delay options, Days Delay 2 has the least RMSE which is around 0.6. Therefore, we decided to add two more days to match with the oil price data. This is rational since the oil price will not reflect to the events that happened on the same day.



We also considered different regression models by observing their definition, best used situation and inner methodology. After that, we decided still use the linear regression model since this is

the most appropriate model that fits our data and meets our requirements. Classification models such as logistic regression, decision tree and random forest are better for categorization or predicting the probability. To predict a specific value that has a simple linear relation with the data, linear regression model is the best we can use. We also verified our result by applying other regression models in the program. It turned out the linear regression always has the least RMSE value.

For calculating the difference between predicted and actual value, we got two options at first. R-squared is an intuitive method to evaluate performance. It ranges from zero to one, with zero represents model does not improve prediction at all and one means it is perfect for the prediction. This make the result easy to tell whether it's good or bad. However, it doesn't show how much the difference is. That's why we chose Root Mean Square Error (RMSE) as our main benchmark. The RMSE is the square root of the variance of the residuals. By using this, we can easily see the mean difference value between our model and the actual data and make further optimization.

Choosing a strategy and algorithm for the machine learning library implementation in Spark, with code written in Scala, was a matter of using the Spark framework and testing variations of our queries and scoring formulas. Spark offers a suite of machine learning algorithms, including several classifier algorithms, logistic regression, linear regression, decision trees, and tree ensembles [14]. While testing and training our model throughout the development work on this project, the results we found to be most accurate as our project evolved led us to use a linear regression algorithm using Spark MLlib. With the linear regression choice, we obtained an RMSE of 0.79 over our testing event data.

Insights Gleaned

At the outset of this project, the GDELT dataset was intended to be used in a different way, by analyzing data from a few different fields such as Location, Actor code, and ArticleUrl, and by filtering events by country code. We anticipated we would be gathering event data where the location or one of the actors in the event was either an OPEC member nation, or another major oil producing nation. The article URLs would be crawled and analyzed for sentiment using Python software libraries and Google language processing APIs.

We found we needed to adjust our perspective after examining the field descriptions listed in the BigQuery schema for the event table, and the values in some sample queries from the database. Each event code falls into a broader category of either Cooperation or Conflict. There are many event codes under each category that could indicate that a particular event has military, political, economic, or social connotations. We actually cared more about particular event codes that were

related to specific incidents which were linked to civil, political, and social unrest [5]. We added filtering for specific event codes, or event base codes to our queries.

It was also seen that there is data in the GDELT event dataset in the AvgTone and GoldsteinScale fields that provide the values that we were interested in for determining sentiment, and impact on society's feelings as a whole [12]. This data allowed us to avoid the time that was expected to be needed to calculate these values by crawling the article URLs.

After looking into the drivers behind the WTI price of crude oil, we chose to remove any filtering from our queries that was based on actor, location, or country code. In the supply and demand relationship for the global crude oil market, the supply side can be affected not only by producers, but by long distance transportation reliability in countries between the source and the destination, as well as nations positioned along major shipping routes. The demand side includes all nations of the world, which informed our decision to include all countries and locations in our queries [1].

While it has been shown in other studies that crude oil prices follow a non-linear deterministic dependence, we found a linear regression algorithm to be best for our test data and our prediction model. This may be due to many factors, such as the differences in the timespans from which the training data was collected for each study [8], or the sources considered in addition to the GDELT event data in some studies. Evaluating data from multiple media platforms is an interesting strategy that has shown promise, and would likely be a good complement to the GDELT data [3].

How Will the Problem Space Transform in the Future

Using GDELT and other datasets of events to predict social unrest and subsequently the effects of its presence or absence will continue to be a challenging, and potentially lucrative area of research. As technology advances, and the availability of internet connectivity becomes increasingly prevalent, there will be vast potential for exploring the large amounts of data that will be generated. Finding a way to sift through it and glean what is meaningful to make predictions may require new tools and ideas for managing such large datasets, and for working with and analyzing the data itself.

The opportunities for governments, investors, and data scientists will attract interest, and likely draw more resources from private and public sectors alike. Investment in these efforts could lead to some new methods for assessing global sentiment that allow more precise estimates of what society as a whole is feeling. Better data will enable better analyses, and the results obtained from prediction models will have higher likelihood of locating trends.

A limited resource in high demand, crude oil will remain a key resource on our planet well into the future. Understanding the global oil price fluctuations and what factors could be at play will be a challenge that is looked at from many points of view as we move into a future that will have ever more data available to digest. The wide variety and increasing number of places that people post information on the internet is creating limitless possible combinations of sources and permutations of data combinations for prediction models. The sheer scale of this will be so enormous that it will require some new ways of handling these datasets as they grow.

Some questions present themselves when considering the future of prediction abilities using extremely large datasets. How will we know which event data sources are the most reliable, as people can change preferences for media platforms over time? What ethical questions will arise if algorithms can predict economic, political, or military behavior and outcomes of events before they happen? Will governments increase regulation on open data sources like GDELT, having a negative effect on the quality of the publicly available datasets? These and other issues will need to be dealt with as the field of study evolves.

Machine learning is likely to make advances that allow more sophisticated analysis, giving us the ability to assess more data points than is currently possible. That, combined with the increasing amount of data available on many media platforms, could make the prediction models even more accurate.

Conclusions

When compared with similar research studies, we have attempted something that others also have, in searching for predictable trends in the oil price with a link to sentiment analysis. In our careful selection of event data from the GDELT project, we were able to find a new slice of the dataset to evaluate. By focusing on the most important parts of the data being explored, we were able to construct our prediction strategy on top of higher quality data. Using a higher number of data points in a machine learning algorithm can have the potential to uncover hidden relationships between seemingly unrelated data, but could also be inefficient in the case that those relationships are not very meaningful, or even non-existent. That could result in wasted processing time, data transfer time, and even money, since the charges are levied based on the amount of data queried. We attempted to use input data that we knew was pertinent to our goals with the thought that it would increase both efficiency and effectiveness.

Overall, the results for our linear regression prediction model on the data evaluated had an RMSE of 0.79. This is among the most accurate prediction models among similar research studies. For example, in some similar studies where event data was used to measure sentiment for forecasting crude oil prices, there were other, non-linear algorithms such as Artificial Neural Networks (ANN), GARCH, ARMA, ARIMA, and ARIMAX, that were used. The values

obtained for the RMSE in those studies were 2.85, 3.90, 5.41, 13.878, and 1.683, respectively [3], [8].

The fact that these are all higher than the RMSE that was obtained in our research could be due to a few reasons. The studies in question were not identical to the study we conducted, but merely had some similarities in methods and goals, to varying degrees. Some key differences are in the details of the data that was evaluated, and in the data manipulation used to find meaning in that data.

Our study evaluated GDELT data and WTI oil price data for the most recent 5 years, as opposed to some of the comparable studies that used the same database but with a different window of time, or studies that used entirely different datasets along with GDELT. With a linear regression algorithm, we were able to obtain results with lower RMSE than reported by these same other mentioned studies that used nonlinear algorithms. This indicates that perhaps over windows of time that are shorter term, a linear regression algorithm may provide a better fit for the data model we tested, compared to the WTI oil price data.

Bibliography

[1] U.S. Energy Information Administration, "Independent Statistics & Analysis," [Online]. Available: https://www.eia.gov.

[2] US Department of State, "Independent States in the World," Bureau of Intelligence and Research, 20 January 2017. [Online]. Available: https://www.state.gov/s/inr/rls/4250.htm.

[3] M. Elshendy, A. F. Colladon, E. Battistoni and P. A. Gloor, "Using Four Different Online Media Sources to Forecast Crude Oil Price," *Journal of Information Science,* March 2017.

[4] J. Noguera-Santaella, "Geopolitics and the Oil Price," *Elsevier Economic Modeling,* vol. 52, no. Part B, pp. 301-309, 2016.

[5] G. Korkmaz, J. Cadena, C. Kuhlman, A. Marathe, A. Vullikanti and N. Ramakrishnan, "Combining Heterogeneous Data Sources for Civil Unrest Forecasting," in *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, Paris, France, 2015.

[6] M. Radetzki, "Politics - Not OPEC Interventions - Explain Oil's Extraordinary Price History," *Elsevier Energy Policy,* 2012.

[7] F. Qiao, P. Li, X. Zhang, Z. Ding, J. Cheng and H. Wang, "Predicting Social Unrest Events with Hidden Markov Models Using GDELT," *Discrete Dynamics in Nature and Society,* vol. 2017, no. Article ID 8180272, p. 13, 2017.

[8] S. Moshiri and F. Foroutan, "Forecasting Nonlinear Crude Oil Futures Prices," *The Energy Journal,* vol. 27, no. 4, pp. 81-95, 2006.

[9] P. A. Schrodt, "Cameo Code Reference," March 2012. [Online]. Available: http://data.gdeltproject.org/documentation/CAMEO.Manual.1.1b3.pdf. [Accessed November 2017].

[10] Google, "Google BigQuery GDELT Event Data," [Online]. Available: https://bigquery.cloud.google.com/table/gdelt-bq:full.events?tab=schema. [Accessed 28 November 2017].

[11] J. Li, Z. Xu, L. Yu and L. Tang, "Forecasting Oil Price Trends with Sentiment of Online News Articles," *Procedia Computer Science,* pp. 1081-1087, 2016.

[12] GDELT Project, "GDELT Data," [Online]. Available: https://www.gdeltproject.org/data.html. [Accessed 28 November 2017].

[13] St. Louis Federal Reserve Economic Research Data (FRED), "Crude Oil Prices: West Texas Intermediate (WTI)," [Online]. Available: https://fred.stlouisfed.org/series/DCOILWTICO/. [Accessed 28 November 2017].

[14] Apache, "Apache Spark," [Online]. Available: https://spark.apache.org/. [Accessed 28 November 2017].