Early Warning of Impending Oil Crises Using the Predictive Power of Online News Stories

Felix Wex  
University of Freiburg  
felix.wex@is.uni-freiburg.de

Natascha Widder  
University of Freiburg  
widder@informatik.uni-freiburg.de

Michael Liebmann  
University of Freiburg  
michael.liebmann@is.uni-freiburg.de

Dirk Neumann  
University of Freiburg  
dirk.neumann@is.uni-freiburg.de

Abstract

Extreme events (such as natural disasters, political upheaval, economic crises) typically have a strong impact on crude oil markets and related price fluctuations and may eventually emerge to global oil crises. This study attempts to early detect such events based on the predictive power of online news messages. Text mining algorithms are used to turn unstructured news into actionable information and to determine which news can be regarded as relevant for the oil market. Over 45 million news messages have been examined. A decision support system is constructed which uses an indicator metric to set off an alarm based on information gathered from current and historic news stories. Regression analyses statistically attest the predictive power of online news messages and thus demonstrate the potential of the early warning system. The effect on the price of crude oil is statistically significant.

1. Introduction

The importance of crude oil as a limited natural resource, and as a booster but simultaneously a hazard for the economies, is beyond controversy. An ever-increasing demand and an expected natural limitation to oil production fields impede this issue [6]. Yet, what has been questioned throughout decades is the issue about what factors do actually influence the oil price determination process which is dictated on the commodity trading floor. Due to the global influence of economies and governments on oil, possible factors range from flexible supply quotas (set by OPEC), national oil reserves to oil demand forecasts [1]. Political unrest and instability in OPEC countries deteriorate this matter. On the other hand, economic crises may also lead to momentous reactions by politicians and economics. Other causes, which play a major role when it comes to determining the oil price, are inevitably uncertain consequences of natural disasters which are either directly connected to the oil production process or to other oil-relevant issues. Left-out in this listing are abnormal speculations that can never be explained by the single occurrence of any of the above factors but which are blamed especially for rapid price increases. Particularly since 2004, the price of crude oil and commodities in general has risen steadily – yet with a high volatility – until suddenly and rapidly declining in 2008. One might attribute this to macroeconomic developments and the global economy which evidently flourished in the years before 2008, leaving various economies in a severe recession afterwards.

We take the above as our motivation to descriptively investigate the short-term developments (fluctuations) of the oil price in the last years (Figure 1) during and right in the aftermath of such events. We are especially interested in its volatile behavior. For example, high oil price fluctuations can be observed after Hurricane Katrina at the end of August 2005, after the Lehman Brothers bankruptcy in September 2008, and after tensions in the Gaza strip at the very beginning of 2009.

![Figure 1. Daily closing crude oil prices (WTI, US$/BBL, Jan 2004-Dec 2010).](image)

Interestingly, no remarkable fluctuations in the oil price time series can be identified after the Deepwater
Horizon oil spill in April 2010. The reasons for this may not surprise: (i) the period of the spill was reasonably long and (ii) at the time of the accident, the platform was not fully integrated in the oil producing process and was thus not directly involved in the global oil supply.

Next to the price of crude oil, we draw our attention to the online news media for the same period of time (in detail: Thomson Reuters corpus). We argue that these online news stories possess – even though it may be hidden – information to leverage advanced computational tools (such as machine learning) to reveal trends and correlations in regard to the oil domain [18]. We argue that it is imperative not only for commodity traders but also for governments to respond rapidly to emerging oil crises.

Some 45 million news stories have been examined over a period of eight years (2003-2010). Figure 2 shows the volume of the yearly level of news stories postings. Similarly to the oil price, published news messages seem to constantly rise until 2008 when a saturation level is slowly reached with approximately six million news stories per annum. Looking at the original data sample, less news stories are published on weekends. In the year 2007 for example, “only” some 146,000 messages have been published overall on Saturdays and some 176,000 on Sundays, whereas messages on weekdays accumulated over the whole year to 1.1-1.3 million messages (Ø21,000 to Ø25,000 messages per day).

![Figure 2. Volume of the complete Reuters news data set per year.](image-url)

In this study, we attempt to link the oil price with the predictive power of news stories. In order to do so, we are about to investigate a) whether online news media do project the public reaction on factors relevant for the oil price and b) whether news may serve as an indicator for such effects. We make use of the statement of [2] that news stories are published first before the market reacts, that is, we also hypothesize that oil price changes are mirrored in an (anomalous) release of news messages before the price fluctuation actually occurs. The final research objective is thus to make use of this behavior and to create an early warning system which is able to identify oil critical events underlined by the predictive power of online news messages. Accordingly, the research question reads as follows: are online news stories apt to identify possible short-term oil crises? In the follow-up, we refer to an event as being “oil critical” if an anomalous return on the price of crude oil is noticed in its direct aftermath.

In our principal understanding, the early warning system, which is about to be proposed, may serve and act as a generic framework which is not bound to the oil domain only. Current requirement analyses reveal the potentials for the real estate domain to identify e.g. subprime crises. Also, other commodity domains are conceivable. Following a sustainable research strand would allow to complete the emergency response depicted by van de Walle and Turoff (2007) [22].

The remainder of this paper unfolds as follows: Section 2 presents related work and common approaches when it comes to econometrically predicting the oil price and text mining approaches related to oil price forecast. The system design and text mining methodologies used are described in section 3. Section 4 introduces the theoretical framework for an oil crisis early warning system and the underlying indicator metric. The results are evaluated with regression analyses thereafter. The paper closes with final remarks and a conclusion.

2. Related Work

This section addresses two literature streams to identify the state-of-the-art in text mining methods used for oil relevant predictions. This section also states the research gap which is investigated in this study. The first stream examines novel (statistical) prediction methods, the second is especially devoted to former and current IS research.

The first strand of researchers approach oil related issues by exploring oil production and demand ratios, shock indications and especially market value prediction methods. Advances in oil price predictions are mainly developed by the econometric research community [9, 20, 23] especially with GARCH, EGARCH, and AR-models. Most of which try to make credible forecast statements by estimating returns on the price of crude oil or the oil price volatility [14, 16], or by complying with oil supply and demand frameworks [5]. Some also take into account the dependence to other economic factors [3, 11].
However, Zhang et al. (2009) [27] have previously proven that the applicability of most oil price prediction methods did not improve the random walk model in practice.

The second strand of researchers approach issues related to credible forecasts about the oil market by text mining. In general, text mining is used to extract information and knowledge from unstructured textual data [21] using – besides others – techniques from statistics, information retrieval, and machine learning. We use the term Text Mining as a proxy for all synonyms (such as knowledge discovery in text, intelligent text analysis, etc.). Unveiling hidden information (e.g. pattern evidence in structures) from large amounts of textual documents is the main objective behind this theory. Huge amount of application areas exist to make use of machine learning approaches to identify trends (e.g. epidemics, disease outspreads, uproars). The United Nations Global Pulse report (2012) [18] lists some of them.

Even though information system methodology, i.e. text mining, is common when it comes to financial (stock) market predictions based on analyses of financial (ad-hoc) messages [10, 12, 17], the literature is remarkably silent about the oil domain. Market value and oil price prediction methods by IS sparely make use of machine learning and artificial intelligence, but if so, then especially in terms of support vector machines [23], genetic algorithms and/or neural networks [4, 8, 20, 26]. Yu et al. (2005) [25] are only some of the few who use findings from text mining and rough set theory to predict tendencies of oil price changes.

We hypothesize that the information system discipline may contribute much more to the discussion about oil relevant prediction methods by providing a decision support platform that is apt to identify days when oil critical events occur. We argue that this becomes possible by means of a) system-based prediction methods, b) early warning systems, and c) knowledge discovery within large databases (feature classification, sentiment analyses, and others). Since we assert that especially online news stories have predictive power when it comes to publicly announcing (negative) consequences about oil relevant events, we identify text mining research for the oil domain as an understudied, yet highly relevant field.

We subsequently address this research gap, which is identified through our literature analysis. We elaborate a decision support module by assisting to identify the events/days by online news which potentially will have a significant impact on the oil-price based on absolute numbers of oil relevant news stories, basic sentiment analyses, and historic oil prices. We commence by presenting the data and methodology used.

3. Data and Methodology

The system architecture of our early warning system consists of two main modules: a text mining module and an analytical component (see Figure 3).

The system itself makes use of two streams of primary data which focus on an eight years period (01/2003-12/2010; 2,921 days). As statistical data source of oil price market data we query Thomson Reuters Datastream to retrieve daily closing crude oil prices to serve as a substitute for the oil market (ICIS Pricing, WTI, US$/Barrel). The time series retrieved includes 2,088 data entries excluding weekends.

We subsequently address this research gap, which is identified through our literature analysis. We elaborate a decision support module by assisting to identify the events/days by online news which potentially will have a significant impact on the oil-price based on absolute numbers of oil relevant news stories, basic sentiment analyses, and historic oil prices. We commence by presenting the data and methodology used.

3. Data and Methodology

The system architecture of our early warning system consists of two main modules: a text mining module and an analytical component (see Figure 3).

The system itself makes use of two streams of primary data which focus on an eight years period (01/2003-12/2010; 2,921 days). As statistical data source of oil price market data we query Thomson Reuters Datastream to retrieve daily closing crude oil prices to serve as a substitute for the oil market (ICIS Pricing, WTI, US$/Barrel). The time series retrieved includes 2,088 data entries excluding weekends.

The system architecture of our early warning system consists of two main modules: a text mining module and an analytical component (see Figure 3).

The system itself makes use of two streams of primary data which focus on an eight years period (01/2003-12/2010; 2,921 days). As statistical data source of oil price market data we query Thomson Reuters Datastream to retrieve daily closing crude oil prices to serve as a substitute for the oil market (ICIS Pricing, WTI, US$/Barrel). The time series retrieved includes 2,088 data entries excluding weekends.

The system architecture of our early warning system consists of two main modules: a text mining module and an analytical component (see Figure 3).

The system itself makes use of two streams of primary data which focus on an eight years period (01/2003-12/2010; 2,921 days). As statistical data source of oil price market data we query Thomson Reuters Datastream to retrieve daily closing crude oil prices to serve as a substitute for the oil market (ICIS Pricing, WTI, US$/Barrel). The time series retrieved includes 2,088 data entries excluding weekends.

The system architecture of our early warning system consists of two main modules: a text mining module and an analytical component (see Figure 3).

The system itself makes use of two streams of primary data which focus on an eight years period (01/2003-12/2010; 2,921 days). As statistical data source of oil price market data we query Thomson Reuters Datastream to retrieve daily closing crude oil prices to serve as a substitute for the oil market (ICIS Pricing, WTI, US$/Barrel). The time series retrieved includes 2,088 data entries excluding weekends.

The system architecture of our early warning system consists of two main modules: a text mining module and an analytical component (see Figure 3).

The system itself makes use of two streams of primary data which focus on an eight years period (01/2003-12/2010; 2,921 days). As statistical data source of oil price market data we query Thomson Reuters Datastream to retrieve daily closing crude oil prices to serve as a substitute for the oil market (ICIS Pricing, WTI, US$/Barrel). The time series retrieved includes 2,088 data entries excluding weekends.

The system architecture of our early warning system consists of two main modules: a text mining module and an analytical component (see Figure 3).

The system itself makes use of two streams of primary data which focus on an eight years period (01/2003-12/2010; 2,921 days). As statistical data source of oil price market data we query Thomson Reuters Datastream to retrieve daily closing crude oil prices to serve as a substitute for the oil market (ICIS Pricing, WTI, US$/Barrel). The time series retrieved includes 2,088 data entries excluding weekends.

Figure 3. System design.
which will allow for several hundred pages of plain text. News stories may contain up to 1,000 words and are generally formulated concisely. Some news appears twice due to story updates or appendices. To understand the structure of a Reuters news story, Table 1 provides an example with the most relevant entries. For the selected period, 45.58 million news stories were gathered making up for about 700MB to 1.2GB of text messages per month.

We classify messages by their relevancy in regard to the oil domain. This can be regarded as a first pre-processing step of sentiment analysis [15]. In order to further pre-process the textual data, we make use of methods from text mining. First, we select only headers and bodies of news stories which possess an English language tag. This is due to a massive amount of noise caused by (different) messages in 19 different languages. Other story types such as adjudications, interviews, sports, or raw stock price reports are ignored. We omit reports that are published during weekends because we cannot clearly assign weekend news to either Friday or Monday news and because oil pricing data is not available on weekends.

Table 1. Sample of a typical Reuters news story.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Entry</th>
</tr>
</thead>
<tbody>
<tr>
<td>Date</td>
<td>2008-12-01 05:35:09</td>
</tr>
<tr>
<td>Headline</td>
<td>Energy shares slammed by OPEC disappointment</td>
</tr>
<tr>
<td>Story Text</td>
<td>HOUSTON, Dec 1 (Reuters) - North American energy stocks of all types fell sharply on Monday as investors were disappointed that OPEC had deferred a decision on cutting crude oil production. [...]</td>
</tr>
<tr>
<td>Language</td>
<td>EN</td>
</tr>
<tr>
<td>Codes</td>
<td>[Several story classifiers (topic, product, item, and instrument codes)]</td>
</tr>
</tbody>
</table>

Second, we create a classifier that extracts only the news which is regarded as oil relevant in our sense. Thus, we conduct several tasks: most importantly, we filter all messages referring to the commodity crude oil. Verification is one of the main challenges in the classification process of unstructured documents (also see [15]). Due to the sheer mass of information, a manual checking of the correct class of a message is impossible. Thus, we resolve this challenge by creating multiple classes that map all oil relevant news stories to at least one class.

In accordance to the keywords provided by Thomson Reuters topic codes, we establish several filters (classes) selecting either one of the topic codes “OILS”, “CRU” (crude oil), “HOIL” (heating oil/diesel), “ENR” (energy & resources), “JET” (jet fuel), “NSEA” (North Sea oil), or “OPEC”.¹ Note that stories may possess several topic codes, such that the same story may appear in several classes. Also, we tested all non-redundant news with the product code label “O” (Oil/Energy industry stories). Figure 4 lists the daily news volume (moving average) of some four important news categories (topic codes) per day over the whole eight years period. Two characteristics are surprising: (1) the amount of news stories about a topic code can exceed several hundred messages a day, (2) the numbers of some topic codes seem to assimilate towards a constant threshold, whereas the number of messages that are “ENR”-encoded (stories related to energy and resources) seems to be growing constantly. This may be attributed to several factors, such as an increased awareness level about the topic or a novel publication strategy of Reuters. Coefficients of variations lie between 0.59 and 0.83 in all four classes.

¹For a detailed explanation of these codes please visit https://customers.reuters.com/training/trainingCRMdata/promo_content/ReutersCodes.pdf, last accessed June 07, 2012.

Figure 4. News stories per day according to topic codes (cumulative moving average).

In some cases, neither a product nor a topic code may appropriately reflect the contextual relationship of a message linked to oil. Thus, we establish a manual classifier and consequently query the whole story sample (text bodies only) for a pre-defined bag-of-words by retrieving all messages consisting of at least one word of a distinct keyword list. This keyword list comprises the names of the 20 largest oil and gas companies [13], the names of the three most important crude oil types (WTI, Brent, Dubai-Oman), as well as terms that inflict a relationship to oil (e.g. ‘OPEC’, ‘Fuel’, ‘Butane’, ‘Petrol’, ‘Kerosene’, ‘Diesel’, and more). We then count the number of news stories each
day which are selected. Figure 5 depicts the daily news volume gathered by this bag-of-words query. The grey line depicts discrete daily numbers; the black solid line represents the moving average of news stories queried by the bag-of-words model. A high fluctuation can be noticed, altering between a few to up to 560 observations per day. The moving average levels off at around 100 messages a day. Yet no significant correlation exists, the number of oil relevant messages seems to strongly fluctuate at times when also the volatility of the crude oil price is high (esp. in years 2007-2009, cp. Figure 1 to Figure 5).

Figure 5. Daily news volume of bag-of-words model (continuous, daily) and its moving average.

Since we are interested in absolute numbers of (words of) oil messages per day for the subsequent analyses, we are able to omit other popular text mining procedures such as stemming and stop word removal. In addition, all redundantly classified messages are neglected. However, we count not only the number of oil relevant messages but also, in an ongoing effort, the number of (negative) words in these classified news stories by a given word list. Similar dictionary-based sentiment approaches are common in research [7, 17]. We make use of dictionaries provided by Loughran and McDonald 2011 [7]. All classifications and preprocessing steps are manually implemented using Java. Queries are posed using SAS.

The analytical module is used as a prerequisite for our early warning system. In this module, information from the above text mining is confronted to financial oil data (Datastream) to identify abnormalities. An indicator metric makes use of absolute numbers of oil relevant news messages, as previously classified, on a daily basis, to determine whether an alarm is to be triggered. The metric itself is presented in the next section.

A man-machine interface acts as a possible third component of our system. This module enables an information exchange between our system and the user. Any predefined parameters or information source can be adapted. It is also used by the user to communicate with other components of the system. Dashboard functionalities or pure visual analytics are conceivable as well as tracking possibilities of real-time analyses based on incoming news messages.

### 4. Indicator Metric

We now present an indicator metric that forms the analytical component of the early warning system. This metric is based on data previously described. We assume that an anomalous change in the oil price between two days may serve as a proxy for an emerging, oil critical event and thus as an indicator for a potential crisis.

In the analysis, we take two factors into account: absolute news messages per day (by category and by bag-of-words query from above) as well as daily WTI (crude oil) prices. We chose grade “WTI” crude oil since this is often used as a benchmark in oil pricing. All (historic) information has been gathered for the years 2003 until the 2010 (2088 weekdays). The metric is to identify days when to expect an extraordinary, absolute return on the (WTI) crude oil price by exceeding a pre-defined threshold (price peaks as well as bottom prices). As one reason, we argue that this may presumably be due to an oil relevant event which provokes an extreme low or high volume of news releases in online media and which is thus critical on the oil price. We refer to \( r(t) \) as the return on the price of crude oil (WTI) for day \( t \):

\[
r(t) = p_{WTI}(t) - p_{WTI}(t-1)
\]

On a daily basis \( (t \in T) \), we define our crisis identification metric as a function of \( x(t) \):

\[
f(x(t)) = \begin{cases} 
1, & \text{if } (x(t) - \bar{x}_d)^2 - s_d^2 > 0; \\
1, & \text{if } (x(t-1) - \bar{x}_d)^2 - s_d^2 > 0; \\
1, & \text{if } (x(t-2) - \bar{x}_d)^2 - s_d^2 > 0; \\
0, & \text{otherwise.}
\end{cases}
\]

\( x(t) \) is the absolute number of news messages which were previously classified as oil relevant and retrieved by the simple bag-of-words query at day \( t \). \( f(x(t)) \) is a function over \( x(t) \) which determines whether day \( t \) is identified as crucial or not. In our sense, \( f(x(t)) = 1 \) will trigger an alarm. \( \bar{x}_d \) and \( s_d^2 \) are common statistical measures (mean and variance of
news message releases within the last \(d\) days up to day \(t\). “\(d\)” is a pre-defined parameter. \(p_{WTI}(t)\) are daily closing prices of the crude oil grade “WTI” at day \(t\). Based on the pricing information we can determine whether the return on the price of crude oil exceeds any kind of threshold at day \(t\), that is whether \(|r(t)| > z\). The excess of the threshold \(z\) also helps us to determine whether the metric \(f(x(t))\) identifies a day as being crucial correctly (true positive) or if it misidentifies the very same.

Before triggering an alarm \(f(x(t)) = 1\), the metric distinguishes three cases: let an alarm be triggered if an abnormal behavior in news stories has been recognized (i) at the observation day \(t\), (ii) a day ago \((t-1)\), or (iii) two days \((t-2)\) before the actual excess of the threshold \(z\) by the price change at day \(t\). This distinction needs to be necessarily accounted for since events are usually published at the time of their occurrence but which will not immediately have an effect on the price but rather a day later or two (latency). In order to compare different settings in the threshold parameter \(z\) used, we introduce a confusion matrix in Table 2 and let \(z\) vary. The accuracy of the indicator metric is depicted by the traditional F-measure (harmonic mean between precision and recall [19]).

Table 2. Crises identification (confusion matrix), \(d=30\).

<table>
<thead>
<tr>
<th>Threshold (z=2)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>201 true positives</td>
<td>114 false negatives</td>
</tr>
<tr>
<td>349 false positives</td>
<td>1424 true negatives</td>
</tr>
<tr>
<td>Precision</td>
<td>.365</td>
</tr>
<tr>
<td>Recall</td>
<td>.638</td>
</tr>
<tr>
<td>F</td>
<td>.464</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Threshold (z=2.5)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>120 t.p.</td>
<td>70 f.n.</td>
</tr>
<tr>
<td>430 f.p.</td>
<td>1468 t.n.</td>
</tr>
<tr>
<td>Precision</td>
<td>.218</td>
</tr>
<tr>
<td>Recall</td>
<td>.632</td>
</tr>
<tr>
<td>F</td>
<td>.324</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Threshold (z=3)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>71 t.p.</td>
<td>51 f.n.</td>
</tr>
<tr>
<td>479 f.p.</td>
<td>1487 t.n.</td>
</tr>
<tr>
<td>Precision</td>
<td>.129</td>
</tr>
<tr>
<td>Recall</td>
<td>.582</td>
</tr>
<tr>
<td>F</td>
<td>.211</td>
</tr>
</tbody>
</table>

If we set \(z = 2.5\), we calculate \(120 + 70 = 190\) days when the oil price indication metric exceeds the threshold. Thus, the perfect indicator metric would need to correctly identify all 190 days as being critical on the oil price.

Altogether, the matrix indicates several conclusions: first, the metric triggers in \(120 + 430 = 550\) of 2088 cases an alarm, yet, 430 of those are false alarms (for \(z = 2.5\)). Even though precision is quite low, we get a decent recall indication. We attribute the low precision to the simple classification process which selected too many news messages as oil relevant (noise). One can assume that various news releases have occurred reporting about several oil-relevant topics, even if they were minor, and they marginally affected the oil price. This concluded in causing false alarms (false positives). We noticed this behavior of news releases during and after the 2004 Atlantic hurricane season and during the period of the Deepwater Horizon oil spill in 2010: the indicator metric set off an alarm but the threshold \(z\) was not exceeded. On the other hand, the metric did not trigger an alarm correctly in between 73%-79% of all cases (for all \(z\)). The indication metric performs best when setting threshold \(z = 2\). We are then able to achieve an F-measure of 0.464. In order to evaluate the indicator metric in regard to its parameter settings, we altered parameter \(d\) \((d = 50, d = 100, \text{and } d = 200)\). In addition, we added more cases in terms of potential days to the metric \((x(t-3)\) and \(x(t+1)\)) that are able to set off an alarm. Both approaches resulted in weaker F-measures than above. Another test was to include additional information from absolute news volume according to single topic codes: adding the information from [ENR, OILS, HOIL, OPEC] encoded messages does not result in a better performance of the indicator metric. We attribute this to the existence of extreme noise with no information attached.

5. Empirical Study and Evaluation

The analysis of the indicator metric above is based for the most part on a descriptive evaluation. We now try to prove that specific news stories do possess information relevant for the return on the crude oil price (WTI) and show that the indicator metric from above possesses statistical merit. Four predictive (regression) models are introduced. The information content from our bag-of-words model is taken first before messages of single topic codes are added into the regressions. Messages of several topic codes messages are then appended as separate variables. We refer to the same representation of the oil price change \(r(t)\) as in the previous section and commence with the evaluation in the follow-up:

\[
r(t) = p_{WTI}(t) - p_{WTI}(t - 1)
\]

\(\beta_i\) are the so-called unknown regression coefficients and \(\varepsilon_t\) the error terms (noise) which are all normally
and independently distributed. Previous works [11] applied the Jarque-Bera test on the oil price which showed that the crude oil price change distribution lacks evidence to map a normally distributed series. However, we also made use of a basic GARCH(1,1) model that makes volatility forecasts of the oil price by referring to previous day price changes, and to prove that there is no predictive evidence in the volatility of the oil price:

$$\sigma_t^2 = \beta_1 r_{(t-1)}^2 + \alpha_1 \sigma_{(t-1)}^2 + \vartheta, \quad \vartheta > 0, \alpha_1, \beta_1 \geq 0$$

$$\beta_1 = \gamma_t \sigma_t, \quad \gamma_t \sim NID(0,1)$$

Our statistical computations proved that, given \(r(t-1)\), it is not possible to iteratively derive \(\sigma^2\). Also, no correlation exists though \(r(t-1)\) is statistically significant at the 5% level over the whole observation period (with \(R^2=0.003\)). To test whether news stories possess information content which may be relevant for predicting the oil price, we introduce four different regression models.

**Model 1**

$$r(t) = \beta_0 + \beta_1 x(t) + \epsilon_t, \quad \epsilon_t \sim NID(0, \sigma^2_t)$$

**Model 2**

$$r(t) = \beta_0 + \beta_1 x(t) + \beta_2 r(t-1) + \epsilon_t, \quad \epsilon_t \sim NID(0, \sigma^2_t)$$

**Model 3**

$$r(t) = \beta_0 + \beta_1 y(t) + \epsilon_t, \quad \epsilon_t \sim NID(0, \sigma^2_t)$$

**Model 4**

$$r(t) = \beta_0 + \beta_1 y(t) + \beta_2 j(t) + \beta_3 u(t) + \beta_4 v(t) + \beta_5 q(t) + \beta_6 m(t) + \beta_7 n(t) + \epsilon_t, \quad \epsilon_t \sim NID(0, \sigma^2_t)$$

### Table 3. Independent variables (shaded cells) within the various regression models.

<table>
<thead>
<tr>
<th>Regressors</th>
<th>x(t)</th>
<th>j(t-1)</th>
<th>y(t)</th>
<th>j(t)</th>
<th>u(t)</th>
<th>v(t)</th>
<th>q(t)</th>
<th>m(t)</th>
<th>n(t)</th>
</tr>
</thead>
<tbody>
<tr>
<td>M1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>M2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>M3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>M4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 3 depicts the use of the different independent variables within the various regression models (shaded cells). Model 1 (M1) makes use of information gathered from the number of daily absolute, oil-relevant news messages \(x(t)\) which were previously retrieved by the simple bag-of-words model. Whereas Model 2 (M2) integrates both oil news volume \(x(t)\) as well as previous day price changes \(r(t-1)\). Models 3 (M3) and Model 4 (M4) both make use of other classes of daily news volume: \(y(t)\) denotes the number of news messages related to the topic code “CRU” (crude oil) at day \(t\). Regressions of M3 are calculated solely over \(y(t)\). Additional independent variables (degrees of freedom) are included in M4: \(j(t)\) indicates the number of news related to topic code “OPEC” (u(t) to “ENR”, v(t) to “OILS”, q(t) to “NSEA”, m(t) to “HOIL”, and n(t) to “JET”, respectively).

### Table 4. Results of the regression analyses and statistical significance.

<table>
<thead>
<tr>
<th>Model</th>
<th>(\beta_0)</th>
<th>(\beta_1)</th>
<th>(\beta_2)</th>
<th>(R^2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.375**</td>
<td>-.012***</td>
<td>.004</td>
<td>.012</td>
</tr>
<tr>
<td>2</td>
<td>2.154***</td>
<td>-.009***</td>
<td>.026</td>
<td>.026</td>
</tr>
<tr>
<td>3</td>
<td>3.609***</td>
<td>-.033***</td>
<td>.029</td>
<td>.029</td>
</tr>
<tr>
<td>4</td>
<td>.511*</td>
<td>-.004</td>
<td>.008</td>
<td>.008</td>
</tr>
<tr>
<td>5</td>
<td>2.834***</td>
<td>-.024**</td>
<td>.025</td>
<td>.025</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Model</th>
<th>(\beta_0)</th>
<th>(\beta_1)</th>
<th>(\beta_2)</th>
<th>(R^2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>1.408**</td>
<td>-.012*</td>
<td>.035</td>
<td>.013</td>
</tr>
<tr>
<td>3</td>
<td>2.169**</td>
<td>-.009*</td>
<td>.009</td>
<td>.026</td>
</tr>
<tr>
<td>4</td>
<td>3.621*</td>
<td>-.033*</td>
<td>-.038</td>
<td>.029</td>
</tr>
<tr>
<td>5</td>
<td>2.213*</td>
<td>-.018</td>
<td>.027</td>
<td>.025</td>
</tr>
<tr>
<td>6</td>
<td>2.889*</td>
<td>-.024*</td>
<td>.022</td>
<td>.025</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Model</th>
<th>(\beta_0)</th>
<th>(\beta_1)</th>
<th>(\beta_2)</th>
<th>(R^2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>.855**</td>
<td>-.003**</td>
<td>.021</td>
<td>.021</td>
</tr>
<tr>
<td>4</td>
<td>.464</td>
<td>.001</td>
<td>.015</td>
<td>.015</td>
</tr>
<tr>
<td>5</td>
<td>2.225**</td>
<td>-.008**</td>
<td>.023</td>
<td>.023</td>
</tr>
<tr>
<td>6</td>
<td>.559</td>
<td>-.002</td>
<td>.002</td>
<td>.002</td>
</tr>
<tr>
<td>7</td>
<td>-.598</td>
<td>.002</td>
<td>.008</td>
<td>.008</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Model</th>
<th>(\beta_1)</th>
<th>(\beta_2)</th>
<th>(\beta_3)</th>
<th>(\beta_4)</th>
<th>(R^2) (adj. (R^2))</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>-.006***</td>
<td>-.011</td>
<td>.003*</td>
<td>.004**</td>
<td>.062 (.036)</td>
</tr>
<tr>
<td>5</td>
<td>.001</td>
<td>-.017***</td>
<td>.002*</td>
<td>.001</td>
<td>.069 (.043)</td>
</tr>
<tr>
<td>6</td>
<td>-.018***</td>
<td>.013</td>
<td>.002</td>
<td>.003</td>
<td>.052 (.026)</td>
</tr>
<tr>
<td>7</td>
<td>-.001</td>
<td>.003</td>
<td>.001</td>
<td>.003</td>
<td>.014 (-.018)</td>
</tr>
<tr>
<td>8</td>
<td>.005*</td>
<td>.012*</td>
<td>-.003**</td>
<td>-.004</td>
<td>.049 (.023)</td>
</tr>
</tbody>
</table>

\(*, **, ***\) indicate statistical significance levels 10%, 5%, and 1% respectively.
We performed our evaluation for the sample period using SPSS statistics software. State-of-the-art regressions are used as estimation procedure depending on the type of (AR-) model (calibration period: 2003-2005). Table 4 present the results of the regression analyses of the four models. The table depicts most important results of our study, due to paper length restrictions and reasons of clarity and readability. For each model, coefficients of determination $R^2$, regression coefficients $\beta_i$, and their individual significance levels are presented. In order to make meaningful statements about the explanation content and influence of news stories on the oil price, significance levels of various regression variables are desired to be within the 10% margin.

It is not surprising that results of Model 1 and Model 2 are rather similar compared to each other, especially since $r(t-1)$ turns out to be not significant. Coefficients of determination (adjusted) are fairly low, but amazingly, the independent variable “number of news messages” is significant in most cases. At the beginning of the ongoing research, these weak results of M1 and M2 have been presented by the authors in a research-in-progress study.

Results also show that Model 3 performs worse than Model 1 based on $R^2$, except for the year 2006. However, in years 2006 and 2008 we get a 5% level of significance for the independent variable(s). Our underlying tests indicate that messages of topic code “CRU” possess more information than other topic codes taken solely into the regression.

Analyses of Model 4 uncover the potential of achieving a higher $R^2$ by adding the information from the different topic codes to the regression model. Though, this may not surprise since additional parameters trivially possess some information related to the oil price, albeit the respective information content may be small. This is attested by no significance of most independent variables. On the other hand, some information is logically lost due to the degrees of freedom. The adjusted R-squared reveals this.

Comparing the results of Models 1 to 4, we draw several conclusions: (1) Numbers of all oil relevant messages contain some significant information relevant for the oil price, especially after 2006 when $R^2$ tend to become better. (2) The information about the current oil price change $r(t - 1)$ itself is not sufficient to have predictive value on $r(t)$ solely. This can be proved by the Jarque-Bera test and by applying a simple GARCH(1,1) model. Neither improves an additional regressor $r(t)$ the predictability in a model which only integrates news volume. (3) The superiority of our bag-of-words method to classify oil relevant messages attests that the previous indicator metric makes use of significant information which is more useful than information gather from messages of predefined topic codes. (4) Even though the results seem to be weak based on very low values of $R^2$, some of them are yet preliminary and further analysis is promising.

Our intuition is further confirmed by a regression (in progress) of the return on crude oil by the number of negative words in oil relevant messages. This approach may be seen as another step into messages’ sentiment to make meaningful and more exact statements about the direction of oil price changes. In a preliminary regression analysis, we set the number of negative words in all oil relevant messages (bag-of-words) as independent variable and try to predict $r(t)$. For the month January 2008, a coefficient of determination $R^2$ of 0.216 is calculated which inflicts a comparably high impact on the oil price compared to the results which have been achieved in this study.

6. Conclusion and Outlook

For this study, a decision support system has been implemented which refers to hidden, predictive information in online news stories. An indicator metric is used to exploit this information and to identify emerging oil critical events (days) based on excesses of a) the return on the price of crude oil (peaks or bottoms) and b) the number of relevant news stories published. News stories are classified as relevant according to a semi-automatism based on given topic codes and a manual bag-of-words model. The recommendations of this decision support may be used not only by commodity traders to form their buy and sell decisions [24] but also by government to be prepared for upcoming crises. We find that it is imperative to understand the influence of publishing online news stories on the oil price and how this connects to predictive outcomes and propose that findings can be used in financial practice or research. This study tries to prove this bond and attempts to extend potentials of real-time data analyses to global development.

We empirically investigated news stories and showed that they possess hidden information and that this information can be used to expose effects on the oil price. This predictive value, in turn, can also be used to implement a decision support system identifying oil crises. We conclude this by studying news messages over a period of 8 years. Yet, the main contribution of this study is that critical oil price movements can be explained by news stories to some extent.
This paper has yet several shortcomings which are addressed in ongoing research: first, it is essential to revisit the semi-automated classification scheme. Currently, this is done on a rather simple basis which causes too much noise since many messages are selected containing a single word related to oil only. A single word match may be a coincidence and does not indicate a relationship to oil necessarily. Obviously, this would also profit from text mining that analyzes not only the volume but also the content of those news stories in some more detail. Following this approach would clearly indicate the notion of sentiment analyses in future works as well as correctness checks of these (F-measures). Particularly, sentiment analyses allow not only for negative indications of news stories as was tested in this study, but also for neutral and positive “emotions”. Mapping these emotions on speculation behavior could in turn foster the risk assessment process of the introduced oil crises identifier from a financial or political perspective. Methods as of Loughran et al. (2011) and Tetlock et al. (2008) [7, 17] can serve as a benchmark. Another ongoing effort is to enhance the regression analyses (i) by non-linear regression models, (ii) by additional variables such as information from not yet considered text mining streams (i.e. sentiment) or (iii) by macroeconomic attributes (control variables) that do affect the oil domain. On the other hand, the contributions of the text mining methodology may be even further underlined by applying similar methods to other domains (i.e. gold, metal markets) and to validate the findings by an empirical case study with trading agents.

Another extension of this study is to refine the indicator metric. Even though the metric triggers a correct alarm in most cases, false alarms may lead to economical fatal consequences. It is imperative to optimize this behavior. Also, the metric is based on a set of assumptions that need approval. First of all, it needs to be confirmed that oil critical events (days) can be defined solely by an anomalous oil price return. Precise event studies may yield this outcome but other factors play certainly a role. To conclude, the implementation of this study under a real-time environment will prove the merits of the proposed methodologies.

7. References


