Do Investors Read Too Much into News?
How News Sentiment Causes Price Formation

Stefan Feuerriegel
University of Freiburg
Freiburg, Germany
stefan.feuerriegel@is.uni-freiburg.de

Sebastian Felix Heitzmann
University of Freiburg
Freiburg, Germany
sebastian.heitzmann@gmx.de

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

Abstract

It is a well-known fact that financial markets react to information. Even though this relationship seems simple, finding evidence is not easy since information is embedded in textual news releases. Only recent have researchers started to look at the content of news. Interestingly, previous work avoids the inference of a causal relationship between news messages and abnormal returns. In this paper, we concentrate on the oil market from which we identify a (strong) instrument, namely, the number of terroristic attacks, and can reasonably account for the endogeneity problem associated with news releases. In addition, we study how news releases affect stock prices temporally. Thus, we find that a change in news sentiment entails a large change in oil prices.

1. Introduction

To a large extent, market efficiency relies upon the availability of information. Access to market information is available with ease in electronic markets and, because of the straightforward access, decision makers (i.e. consumers, suppliers and intermediaries) can use more information to make purchases and sales more beneficial (e.g. [1]). The (weak) efficient market hypothesis [2] asserts that financial markets are informationally efficient, in the sense that financial stock prices accurately reflect all public information at all times. Adjustments occur when previously unknown information enters the market, as is the case with news. News is typically embodied in text messages that are interpreted differently by different investors [3], [4]. How markets react to news announcements has been the focus of many research publications (e.g. [5]–[8]) – all findings establish a relationship between the facts embedded in news releases and stock returns. Furthermore, recent studies (see [9]–[12] for the first) investigate the content of news releases and reveal that even the content (i.e. the tone of news) affects prices at a statistically significant level.

While previous research succeeded in establishing a link between news tone and prices, authors argue that "the existing literature on financial text does not actually determine the causal link between tone and returns" [13]. Thus, previous work has not been able to determine whether the correlations between stock prices and news tone are simply spurious, hiding the presence of a confounding factor that may establish the causal link. Even in the Big Data era, any reasonable attempt to explain stock prices on the basis of news tone or sentiment must first address the endogeneity problem [14], to ensure that the news tone is not exogenous. If news tone is not purged from endogeneity, it will produce coefficients that are inconsistent and, thus, unreliable. Any policy implications stemming from such studies are suspicious [15], [16]. The reason as to papers fail to account for endogeneity originates from the fact that randomized experiments are infeasible to conduct for practical reasons. This paper, consequently, attempts to address the endogeneity problem by unveiling a causal relationship between news and prices using an instrumental variable approach incorporating non-experimental data. Having inferred the causal transmission channel between news and prices, we can subsequently exploit this channel by deriving how prices change over time depending upon the content of released news.

In our research design, we adopt the oil market as a reference market since the oil market has drawn significant attention in the globalized world. With manufacturing businesses becoming highly dependent on oil, demand for oil has literally exploded, which is reflected by the vast increase in traded oil futures, a total of 138.5 million contracts, each accounting for...
1000 barrels in 2007 [17]. As a consequence, the oil market has attracted many investors and is thus subject to extensive news coverage.

The remainder of this paper is structured as follows. In Section 2, we review previous research on oil price models. To investigate news content, we describe our news corpus and the so-called sentiment analysis to study news tone in Section 3. In order to link the two, Section 4 uses an instrumental variable regression to establish causality. Finally, Section 5 measures to what extent tone in oil news influences oil prices.

2. Related work

This section presents the related literature on econometric models to describe oil price movements. All in all, the following references provide evidence for the causal link between news sentiment and oil prices as both a novel and relevant research question to the Information Systems community.

2.1. Modeling oil prices

To fully capture the concepts of oil price models, we start by presenting the different angles of existing modeling approaches. These models can be distinguished by the type of econometric model and their underlying fundamentals as follows:

- When it comes to distinguish econometric models, we frequently encounter two predominant families of methodological approaches. These include (linear) regression models, such as in [18]. Another common choice are the vector autoregressive models or vector error correction models, as in [19]–[21].
- When investigating models for describing oil prices, we can essentially subdivide approaches according to their used fundamentals into two major streams: first, there are those models (e.g. [20], [21]) that draw on both demand and supply as the main drivers. Second, oil prices can also be modeled by a combination of stock market data [22] and macroeconomic fundamentals [19], [23], [24].

Out of above-mentioned references, only one [21] integrates a news sentiment variable. As it is the publication closest to our research, we follow this approach closely. In fact, it is based on an earlier piece of [20] that appeared in the American Economic Review. The authors utilize a structural vector autoregressive model in order to estimate factors driving oil prices. This approach is based on monthly data that models both the demand and supply side. More precisely, an index of real economic activity describes demand while the percentage change in global crude oil production reflects the supply side.

2.2. Research framework

All in all, this prompts the question of how news drives oil prices. While many publications model oil prices via fundamental variables, we are aware of only a few references which incorporate a news variable. To extend previous research, we investigate (see Fig. 1) the following research questions:

1) Is there a causal relationship between tone in oil news and stock market reactions? To answer this question, we set up an instrumental variable regression using the number of terrorist attacks as the instrumental variable. This proves causation between news sentiment and abnormal returns of crude oil.

2) To what extent are daily crude oil prices driven by news tone over time? We utilize a vector autoregressive model to analyze how the sentiment of oil news influences actual oil prices on a daily basis. Having established causation with abnormal oil returns, we can now quantify the temporal influence on real oil prices. This is in contrast to previous research that studies weekly data [21] or abnormal returns [25], [26] as limitations.

3. Background

This section introduces the background information on both datasets and the sentiment analysis that is
used throughout this paper. Thus, we describe how we construct our news corpus and then transform this running text into machine-readable tokens to measure news sentiment.

3.1. Data sources

This section provides details on the news corpus and the fundamental variables that specify our oil price model. In addition, we describe our instrumental variable that is later used to infer causality.

3.1.1. News corpus. Our news corpus originates from the Thomson Reuters News Archive for Machine Readable News\(^1\). All announcements provided by Reuters are extracted from between January 6, 2004 until May 31, 2012. The announcements come along with additional labels that indicate their content. Based upon these labels, the news corpus is filtered such that we extract announcements that focus on the oil market\(^2\). All in all, the set of filter criteria gives a total of 307,430 announcements related to crude oil.

3.1.2. Fundamental variables. News sentiment can affect daily WTI crude oil prices \(P(t)\). Apparently, WTI crude oil prices are a common choice (e.g. \cite{18}, \cite{19}, \cite{27}) in research that studies oil markets since it acts as the benchmark U.S. price. Besides that, we integrate several control variables, consistent with previous research \cite{20}, \cite{21}. These are as follows: U.S. interest rate \(r(t)\), U.S. Dollar/Euro exchange rate, level \(IM(t)\) of oil imports (in million barrel), open interest \(OI(t)\) in crude oil futures (in million), gold price \(G(t)\) (London, afternoon fixing) and, additionally included, the S&P 500 index \(SP_{500}(t)\). All data spanning business days from January 6, 2004 until May 31, 2012 originates from Thomson Reuters Datastream.

The relationship of standardized news sentiment \(S^*(t)\) and crude oil returns is depicted in Fig. 2.

1. We choose Reuters news deliberately because of four reasons: (1) Reuters conveys, in particular, news about commodity markets. (2) Reuters news is third-party content and thus has a certain level of objectivity. (3) As opposed to newspapers, news agencies feature a shorter time lag and lack of perturbations by editing.

2. This is achieved by applying a set of filter criteria \cite{25}, \cite{26}: (1) The language must be English. (2) The event type is Story Take Overwrite to guarantee that we do not yield an alert but the actual message. (3) Special types of announcements, such as alerts or personal opinions, might have limited relevance to changes in the oil market and we want to exclude these. Thus, we omit announcements that contain specific words (advisory, chronology, corrected, feature, diary, instant view, analyst view, newsmaker, corrected, refile, rpt, schedule, table, service, alert, wrapup, imbalance, update) in their headline. (4) We use topic code CRU to filter announcements that deal with crude oil. (5) We exclude announcements addressing changes in prices to avoid simultaneity. (6) In order to remove white noise, we require announcements to contain at least 50 words.

This diagram features a so-called LOWESS trend line, which is a locally weighted scatterplot smoothing calculated via local regressions. From this LOWESS trend line (smoothing parameter \(f\) set to 2/3), we can identify a visible relationship between the sentiment variable and stock market returns.

3.1.3. Instrumental variable. In our instrumental variable regression, we later refer to the Global Terrorism Database \cite{28} to measure the number of terror attacks. Here, we use the number of incidents that occurred in the top 10 crude oil producing countries\(^3\), as well as Iraq.

3.2. Analyzing news sentiment

Methods that use the textual representation of documents to measure the positivity and negativity of the content are referred to as opinion mining or sentiment analysis. In fact, sentiment analysis can be utilized to extract subjective information from text sources, as well as to measure how market participants perceive and react to news. One uses the observed stock price reactions following a news announcement to validate the accuracy of the sentiment analysis routines. Based upon sentiment measures, the relationship between news content and its effect on stock markets can be studied.

3.2.1. Preprocessing. Before performing the actual sentiment analysis, there are several preprocessing steps as follows.

- **Tokenization.** Corpus entries are split into single words named *tokens*.

3. The top 10 crude oil producing countries \cite{29} are Russia, Saudi Arabia, United States, Iran, China, Mexico, Canada, United Arab Emirates, Venezuela and Kuwait as of 2007. The year 2007 is roughly in the middle of our news corpus spanning 2004–2012.
4806

- **Negations.** Negations invert the meaning of words and sentences [30], [31]. When encountering the word *no*, each of the subsequent three words (i.e., the object) are counted as words from the opposite dictionary. When other negating terms are encountered (*rather, hardly, couldn’t, wasn’t, didn’t, wouldn’t, shouldn’t, weren’t, don’t, doesn’t, haven’t, hasn’t, won’t, hadn’t, never*), the meaning of all succeeding words is inverted.

- **Stop word removal.** Words without deeper meaning, e.g., articles, are removed [32].

- **Stemming.** Stemming refers to the process of reducing inflected words to their stem [33]. Here, we use the so-called Porter stemming algorithm.

### 3.2.2. Net-Optimism sentiment metric

Having completed the preprocessing, one can continue to analyze news sentiment. As shown in a recent study [25] on the robustness of sentiment analysis, the correlation between news sentiment and abnormal returns in oil markets varies across different sentiment metrics. A sentiment approach that results in a reliable correlation is the Net-Optimism metric [34]. Net-Optimism, along with Henry’s Finance-Specific Dictionary [12], achieves the highest robustness and, consequently, we rely upon this approach in the following evaluation.

The Net-Optimism metric $S(t)$ is applied to all news originating from a trading day $t$. It is calculated as the difference between the number of positive $W_{pos}(A)$ and negative $W_{neg}(A)$ words divided by the total number of words $W_{tot}(A)$ in each announcement $A$. Thus, Net-Optimism is defined by

$$S(t) = \frac{\sum_A W_{pos}(A) - W_{neg}(A)}{\sum_A W_{tot}(A)} \in [-1, +1].$$

(1)

To make calculations and later comparisons easier, we also introduce a standardized sentiment

$$S^*(t) = \frac{S(t) - \mu}{\sigma} \in (-\infty, +\infty),$$

(2)

scaled to feature a zero mean and a standard deviation of one.

### 4. Causal influence of news sentiment on abnormal returns of crude oil

While previous research [13] frequently assumes causality between news content and market reactions, we aim to explicitly realize this causal inference. Nowadays, causality has "grown from a nebulous concept into a mathematical theory with significant applications in the fields of statistics, artificial intelligence, economics" [35] and many more. In order to establish causality, one must translate observations into cause-and-effect relationships. Thus, this section introduces the concept of an instrumental variable (IV) regression and then investigates how news tone causally influences abnormal returns of crude oil. As suitable instruments are likely to influence both news sentiment and absolute oil prices, we present – as a remedy – the concept of abnormal returns in the next section.

#### 4.1. Event studies and abnormal returns

Event studies use financial market data to inspect changes in financial values due to a specific event and measure its impact. Information Systems research frequently exploits event study methodology and has turned it into both an effective and widespread approach [36]. The result appears in the form of abnormal returns.

For each event of interest, one predicts a normal return in the absence of the event $X_t$ and then estimates the difference between the actual and normal return which is defined as the abnormal return [37]. In our research, the event of interest consists of all daily oil-related announcements from the news corpus. We set the period during which the oil price is examined (i.e., the event window) to a single day of the announcement stream since we are provided with daily financial market data. The abnormal return is defined by

$$AR(\tau) = R(\tau) - E(R(\tau) \mid \neg X_{\tau})$$

(3)

where $R(\tau)$ and $E(R(\tau) \mid \neg X_{\tau})$ are the actual and normal returns in period $\tau$ respectively. The abnormal return equals the actual return (during the event window) minus the expected return that is based on the defined pre-event time. Thus, the abnormal return gives credit to the extracted effect of the events and measures its impact.

As a next step, the normal return is estimated during a time interval: the estimation window. We calculate the normal return by the so-called market model [37]. We model the market portfolio using a commodity index, namely, the Dow Jones-UBS Commodity Index [38], along with an event window of 10 trading days prior to the event.

#### 4.2. Instrumental variable regression

The method of instrumental variables (IV) is frequently used to estimate causal relationships [15], [16], [39], [40]. In particular, instruments come into play when controlled experiments are infeasible. Let us denote the linear regression

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i.$$  

(4)
Then, IV regression is a general way of obtaining a consistent estimator of the unknown coefficients in a regression, when the regressor $x_i$ is correlated with the error term $\varepsilon_i$ (called endogenous). The idea is to break $x_i$ into two parts: one correlated with $\varepsilon_i$ (this is the part that causes problems) and one uncorrelated (called exogenous), which comes from an instrumental variable.

A valid instrumental variable $z_i$ must satisfy two conditions [41], namely, relevance and exogeneity:

1) **Relevance.** If an instrument is relevant, then variation in the instrument is related to variation in $x_i$, i.e.,

\[
\text{cor}(z_i, x_i) \neq 0. \tag{5}
\]

2) **Exogeneity.** If an instrument is exogenous, then variation in the instrument is not related to the error term $\varepsilon_i$, i.e.,

\[
\text{cor}(z_i, \varepsilon_i) = 0. \tag{6}
\]

If both assumptions hold, the desired coefficients $\beta_i$ can be retrieved [15] by a two stage least squares estimator (2SLS). The first stage links $x_i$ and $z_i$ by

\[
x_i = \gamma_0 + \gamma_1 z_i + \eta_i, \tag{7}
\]

while testing relevance. Then, the predicted values $\hat{x}_i = \gamma_0 + \gamma_1 z_i$ can be injected into the second stage

\[
y_i = \beta_0 + \beta_1 \hat{x}_i + \varepsilon_i. \tag{8}
\]

### 4.3. Instrument

We choose the number of daily terrorist attacks in the top 10 crude oil producing countries (plus Iraq) as the instrument. According to Fig. 3, terrorist attacks must influence oil news sentiment and this news sentiment must affect abnormal returns of oil. In addition, there must not be a direct effect of terrorist attacks on abnormal returns of oil and terrorist attacks must not correlate with the error term $\varepsilon_i$. These are, more formally, captured by the concept of valid instruments:

1) **Relevance Assumption (IV correlated with Endogenous Variables).**

The instrument is relevant in the sense that the number of terrorist attacks correlates with news sentiment. From a theoretical point of view, this link seems plausible as news covers all oil-related disclosures, including refineries, production quotas and general market settings. That is, a terrorist attack is likely to affect the overall news sentiment. According to the first stage regression in Table 4, we see that the $t$-value for the instrumental variable accounts for 4.265 ($P$-value below 0.001), providing significant evidence that the correlation is non-zero and, thus, that the instrument is relevant.

2) **Exogeneity Assumption (IV Distributed Independent of the Error).**

The instrument is exogenous, i.e. it affects abnormal returns only through news sentiment. Thus, the instrument has no direct effect on the abnormal returns of crude oil, which must be proven by logic arguments. It is likely that terrorist attacks affect the overall news sentiment. The supporting argument originates from the definition of abnormal returns, as these measure any kind of excess return compared to the market movement.

If the correlation $\text{cor}(z_i, x_i)$ is weak, the IV tends to yield a larger bias. The so-called strength of instruments can be directly measured by the $F$-statistic against the null hypothesis that the excluded instrument is irrelevant in the first-stage regression. As a common rule of thumb [42], models with one endogenous regressor should have a test statistic larger than 10 for strong instruments. In our case, the first stage regression is presented in Table 4 and we yield a first stage $F$-statistic of 18.194, which is evidence of a strong instrument.

### 4.4. Results

We specify our model — with control variables according to [20], [21] — by

\[
AR_{\text{log}}(t) = \beta_0 + \beta_1 S^*(t) + \beta_2 r_{\text{log}}(t) \\
+ \beta_3 FX_{\text{log}}(t) + \beta_4 IM(t) + \beta_5 OI(t) \tag{9} \\
+ \beta_6 G_{\text{log}}(t) + \beta_7 S_{\text{log}}(t) + \varepsilon(t),
\]
5. Temporal influence of news sentiment on absolute oil prices

While the previous section infers causality, we proceed to analyze how news sentiment influences oil prices temporally. Thus, we can study news reception over time. This research question demands the integration of two model adjustments: on the one hand, we need to switch from an instrumental variable regression to a multivariate time series approach, more precisely, a vector autoregressive model. On the other hand, the vector autoregressive model forces us to replace abnormal returns with absolute oil prices. This is because the IV regression requires abnormal returns as a utility to guarantee a valid instrument, but now our interest shifts to the development of absolute oil prices. Consequently, we can examine the linear interdependencies in the vector autoregressive model, which specify how fundamental variables, news tone and crude oil prices interact.

5.1. Vector autoregressive model

The vector autoregressive (VAR) model is used in econometrics to capture the linear interdependencies

Table 4. Results of first stage regression with standardized news sentiment as the dependent variable using data from January 6, 2004 until May 31, 2012.

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
<th>(5)</th>
<th>(6)</th>
<th>(7)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-0.033</td>
<td>-0.039</td>
<td>-0.052</td>
<td>0.497</td>
<td>-0.470</td>
<td>-0.553</td>
<td>-0.452</td>
</tr>
<tr>
<td>$T(t)$</td>
<td>0.028***</td>
<td>0.028***</td>
<td>0.027***</td>
<td>0.029***</td>
<td>0.029***</td>
<td>0.028***</td>
<td>0.027***</td>
</tr>
<tr>
<td>$r_{log}(t)$</td>
<td>0.001*</td>
<td>0.001*</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001*</td>
<td>0.001</td>
<td>0.001</td>
</tr>
<tr>
<td>$FX_{log}(t)$</td>
<td>0.368***</td>
<td>0.367***</td>
<td>0.367***</td>
<td>0.258***</td>
<td>0.258***</td>
<td>0.299***</td>
<td>0.299***</td>
</tr>
<tr>
<td>$IM(t)$</td>
<td>-0.002</td>
<td>-0.003*</td>
<td>-0.003*</td>
<td>-0.003*</td>
<td>-0.003*</td>
<td>-0.003*</td>
<td>-0.003*</td>
</tr>
<tr>
<td>$OI(t)$</td>
<td>1.908***</td>
<td>1.891***</td>
<td>1.914***</td>
<td>1.891***</td>
<td>1.914***</td>
<td>1.914***</td>
<td>1.914***</td>
</tr>
<tr>
<td>$G_{log}(t)$</td>
<td>0.156***</td>
<td>0.166***</td>
<td>0.166***</td>
<td>0.166***</td>
<td>0.166***</td>
<td>0.166***</td>
<td>0.166***</td>
</tr>
<tr>
<td>$SP_{log}(t)$</td>
<td>0.131***</td>
<td>(8.860)</td>
<td>0.131***</td>
<td>(8.860)</td>
<td>0.131***</td>
<td>(8.860)</td>
<td>0.131***</td>
</tr>
</tbody>
</table>

|        | 0.030 | 0.032 | 0.090 | 0.091 | 0.114 | 0.135 | 0.166 |
| Adj. R² | 5929.179 | 5926.866 | 5796.249 | 5796.442 | 5743.148 | 5693.442 | 5617.791 |
| AIC     | 5991.368 | 5994.708 | 5869.745 | 5875.591 | 5827.950 | 5783.898 | 5713.900 |
| BIC     | 5991.368 | 5994.708 | 5869.745 | 5875.591 | 5827.950 | 5783.898 | 5713.900 |

Stated: OLS Coeff., t-Stat. in Parenthesis; Dummies: Yearly; Obs.: 2108

Significance: *** 0.001, ** 0.01, * 0.05

where $AR_{log}(t)$ is the log-return variant corresponding to the abovementioned abnormal returns. Similarly, $r_{log}(t), FX_{log}(t), G_{log}(t)$ and $SP_{log}(t)$ also denote log-returns. The second stage results of the IV regression are provided in Table 5. Accordingly, one standard deviation increase in the sentiment measure correlates positively with a change in log-returns by an economically significant 46.2%.

To use IV estimation, it must be balanced against the inevitable loss of efficiency for the sake of consistency. Thus, one performs a test of endogeneity where, under the null hypothesis, the specified endogenous regressors can actually be treated as exogenous. Here, we apply the Wu-Hausman test, which estimates the model via both IV and OLS approaches; the resulting coefficients are compared statistically. In our case, the test statistic equals 1.349 with a P-value of 0.246, which is not significant at common significance levels. The IV regression is a way of detecting an otherwise unobservable omitted variable bias by other unknown influences. Apparently, in comparison to the OLS estimation, we see that the influence of a possible omitted variable bias is marginal. Hence, the IV regression is not favored over the OLS estimation; the least-squares estimator will provide reasonable results.

Overall, the instrumental variable approach provides strong evidence of causation. Consequently, this justifies the use of the time series approach in the subsequent section in order to study the effect over time.

5.1. Vector autoregressive model

The vector autoregressive (VAR) model is used in econometrics to capture the linear interdependencies
among multiple time series. A VAR model describes the evolution of $K$ endogenous variables $y_t \in \mathbb{R}^K$ over a sample period as a linear function of their past values. Then the VAR process with $p$ lags, denoted by $\text{VAR}(p)$, is

$$y_t = c + A_1 y_{t-1} + \ldots + A_p y_{t-p} + u_t$$

with intercept $c \in \mathbb{R}^K$, time-invariant matrices $A_i \in \mathbb{R}^{K \times K}$ and error terms $u_t \in \mathbb{R}^K$, with certain mathematical properties given in [43]. In order to estimate the model parameters, the next section has to first check, as a prerequisite, that all variables are of the same order of integration followed by the determination of the optimal lag length $p$.

### 5.2. VAR estimation

Consistent with [20], [21], our vector autoregressive model includes the following variables: news sentiment, WTI crude oil price, U.S. interest rate, U.S. dollar/euro exchange rate, level of oil imports, total open interest in crude oil future contracts and gold price. In addition, we include the S&P 500 index as a benchmark for the development of the U.S. stock market. As noted in [19], it is common to use logarithms of financial prices. Accordingly, we use logarithmic values of oil prices, gold price, exchange rates and the S&P 500 index.

We tailor the news sentiment in two ways. First, we insert the sentiment variable with a lag of one day, such that the news of day $t$ can affect crude oil closing prices of the same day $t$. We do so because news sentiment is measured before the market closes, whereas all other time series are closing values. Second, we construct a sentiment index $S^\sigma(t)$ from a cumulative sum since we want the sentiment values (and not its differences) to impact news.

In a first step, we test for structural breaks in order to ensure that stationarity tests are unbiased, as advocated in [19]. We find evidence of structural breaks, the last on March 1, 2011 and so restrict our time span of interest to the subsequent observations, March 2, 2011 until May 30, 2012.

Next, we test for unit roots using the augmented Dickey-Fuller (ADF) test, as well as the Phillips-Perron (PP) test. Only the level of oil imports seems to be integrated of order zero, while testing the other time series are closing values. Second, we construct a sentiment index $S^\sigma(t)$ from a cumulative sum since we want the sentiment values (and not its differences) to impact news.

Table 5. Second stage results of instrumental variable regression (January 6, 2004 until May 31, 2012).

<table>
<thead>
<tr>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
<th>(5)</th>
<th>(6)</th>
<th>(7)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>0.015</td>
<td>0.024</td>
<td>0.024</td>
<td>−0.774</td>
<td>1.259</td>
<td>1.280</td>
</tr>
<tr>
<td>$S^\sigma(t)$</td>
<td>1.628**</td>
<td>1.643**</td>
<td>1.644**</td>
<td>1.634**</td>
<td>1.531**</td>
<td>1.537*</td>
</tr>
<tr>
<td>$r_{log}(t)$</td>
<td>(−0.001)</td>
<td>(−0.001)</td>
<td>(−0.001)</td>
<td>(−0.001)</td>
<td>(−0.001)</td>
<td>(−0.001)</td>
</tr>
<tr>
<td>$FX_{log}(t)$</td>
<td>0.002</td>
<td>0.000</td>
<td>0.004</td>
<td>0.070</td>
<td>−0.011</td>
<td></td>
</tr>
<tr>
<td>$IM(t)$</td>
<td>(0.009)</td>
<td>(0.209)</td>
<td>(0.390)</td>
<td>(−0.072)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$OI(t)$</td>
<td>(0.003)</td>
<td>0.005</td>
<td>0.004</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$G_{log}(t)$</td>
<td>(−0.036)</td>
<td>−0.005</td>
<td>(−0.046)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$SP_{log}(t)$</td>
<td></td>
<td>0.271**</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Adj. $R^2$ | 0.030   | 0.031   | 0.090   | 0.090   | 0.114   | 0.135   | 0.166   |
AIC | 11625.232 | 11639.921 | 11639.917 | 11632.166 | 11473.197 | 11478.712 | 11329.041 |
BIC | 11693.074 | 11713.417 | 11719.066 | 11716.969 | 11563.653 | 11574.821 | 11430.804 |

Stated: OLS Coeff., t-Stat. in Parenthesis; Dummies: Yearly; Obs.: 2108 Significance: *** 0.001, ** 0.01, * 0.05
As specification tests, we perform the Johansen test with trace statistics, but find no evidence of cointegration (test statistic: 116.31; 10% sign. level: 118.99). With no evidence of cointegration and with time series stationary in differences, we consequently estimate a vector autoregressive model in differences, i.e.

\[ \Delta y_t = c + A_1 \Delta y_{t-1} + u_t. \]  

(11)

5.3. Results

The VAR estimates are listed in Table 6. We immediately notice a clearly visible effect of news sentiment on oil prices. The coefficient of how \( \Delta S^* (t) \) (or \( S^* (t) \) respectively) affects prices \( \Delta P (t) \) accounts for 0.012 with a corresponding \( t \)-value of 11.907. This remains significant at all common significance levels and is the highest \( t \)-statistic of all the VAR estimates. This supports the hypothesis stated in Section 3.2: news strongly influences prices. Other than news sentiment, only the lagged value of oil prices has a significant impact.

The dynamic effect of news influence is illustrated by the impulse response plot in Fig. 7. Keeping the above VAR model in mind, we find a matching curve with a considerable positive peak indicating the influence of news on the publishing day. Within a few business days, the response converges to zero. Based on the strong influence of the news on day \( t \), this provides evidence that further lagged values of news sentiment are not necessary. In other words, the market absorbs news within a single day. This indicates market efficiency and gives strong reason for the use of daily rather than monthly data.

6. Conclusion

Although it is a well-known fact that financial markets are very sensitive to the release of news, the way in which this information is received is not at all well studied. Not until recently have researchers started to look at the content of news stories using very simple techniques to determine the news tone. Typically, these papers use the bag-of-words model for text representations; the words are subsequently evaluated using a finance-specific dictionary to compute a news tone metric. Interestingly, what these approaches all have in common is the fact that they do not address the inherent endogeneity problem, arising from the many potential factors of influence, but which are rather omitted from their investigations.

The paper at hand attempts to adequately cope with the endogeneity problem by constructing an instrumental variable for news sentiment. If the instrument is properly chosen, the instrumental variable approach remedies biases that may arise from endogeneity. Due to the nature of econometric analyses of financial news content, it is not surprising that any paper has thus succeeded in defining a meaningful instrument. In our analysis, we concentrate on the oil market, which is particularly useful for our purposes, as the functioning of the oil market is less complicated than the stock market, since the price is a result of the interplay between the demand and supply of oil. We construct an instrument that refers to the number of terroristic effects and which has an impact on abnormal returns. As we have shown in our analysis, the instrument is valid, relevant and also strong. Using our instrumental variable approach, we can correct for endogeneity and, furthermore, can infer a causal link between news tone and abnormal returns.

Next, we studied how news acts on oil prices over time. As terroristic attacks are rather infrequent, we cannot use the instrument in the second part of our analysis. Having learnt that the bias is not too large, we employ a modified vector autoregressive model to represent the impact of news tone on oil prices which estimates the influence of news announcements on the same day rather than with a first lagged value. We provide evidence for an enormous response to news tone occurring on the same day.

This paper opens avenues for future research. First, we have investigated the endogeneity problem associated with news limited to the oil domain. Finding suitable instruments for other domains, such as 10-K filings, would be of interest. Second, the investigation of other theories is also appealing, such as probabilistic approaches [35], in order to study causation.

<table>
<thead>
<tr>
<th>Dependent Variables</th>
<th>ΔS^2(t)</th>
<th>ΔP(t−1)</th>
<th>Δr(t−1)</th>
<th>ΔFX(t−1)</th>
<th>ΔIM(t−1)</th>
<th>ΔOI(t−1)</th>
<th>ΔG(t−1)</th>
<th>ΔSP(t−1)</th>
<th>Const.</th>
</tr>
</thead>
<tbody>
<tr>
<td>ΔS^2(t+1)</td>
<td>0.190***</td>
<td>7.206*</td>
<td>−6.364</td>
<td>8.756</td>
<td>0.012</td>
<td>3.915</td>
<td>−3.796</td>
<td>−4.528</td>
<td>−0.023</td>
</tr>
<tr>
<td>[−S^2(t)]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ΔP(t)</td>
<td>0.012***</td>
<td>−0.131*</td>
<td>0.032</td>
<td>−0.161</td>
<td>0.000</td>
<td>0.077</td>
<td>0.043</td>
<td>0.053</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>(11.907)</td>
<td>(−2.304)</td>
<td>(0.274)</td>
<td>(−1.106)</td>
<td>(0.237)</td>
<td>(−1.637)</td>
<td>(−0.622)</td>
<td>(0.642)</td>
<td>(−0.331)</td>
</tr>
<tr>
<td>Δr(t)</td>
<td>0.000</td>
<td>−0.017</td>
<td>0.005</td>
<td>0.006</td>
<td>0.000</td>
<td>−0.011</td>
<td>−0.019</td>
<td>0.038</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>(0.543)</td>
<td>(−0.601)</td>
<td>(0.085)</td>
<td>(0.091)</td>
<td>(−0.012)</td>
<td>(0.484)</td>
<td>(−0.555)</td>
<td>(0.952)</td>
<td>(−0.118)</td>
</tr>
<tr>
<td>ΔFX(t)</td>
<td>0.003***</td>
<td>−0.004</td>
<td>0.008</td>
<td>−0.106</td>
<td>0.000</td>
<td>0.006</td>
<td>−0.053*</td>
<td>0.097**</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>(6.712)</td>
<td>(−0.209)</td>
<td>(0.188)</td>
<td>(−1.930)</td>
<td>(1.314)</td>
<td>(0.323)</td>
<td>(−2.029)</td>
<td>(3.124)</td>
<td>(−0.912)</td>
</tr>
<tr>
<td>ΔIM(t)</td>
<td>−0.617*</td>
<td>−9.783</td>
<td>6.635</td>
<td>75.599*</td>
<td>−0.008</td>
<td>0.628</td>
<td>−8.704</td>
<td>−8.855</td>
<td>0.167</td>
</tr>
<tr>
<td></td>
<td>(−2.354)</td>
<td>(−0.655)</td>
<td>(0.214)</td>
<td>(1.977)</td>
<td>(−0.147)</td>
<td>(0.051)</td>
<td>(−0.473)</td>
<td>(−0.039)</td>
<td>(0.690)</td>
</tr>
<tr>
<td>ΔOI(t)</td>
<td>0.001</td>
<td>−0.067</td>
<td>−0.031</td>
<td>0.180</td>
<td>0.000</td>
<td>−0.005</td>
<td>−0.172*</td>
<td>−0.042</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>(1.061)</td>
<td>(−0.974)</td>
<td>(−0.219)</td>
<td>(1.022)</td>
<td>(−0.856)</td>
<td>(−0.096)</td>
<td>(−2.037)</td>
<td>(−0.415)</td>
<td>(−0.233)</td>
</tr>
<tr>
<td>ΔG(t)</td>
<td>0.002*</td>
<td>0.104*</td>
<td>0.081</td>
<td>0.687***</td>
<td>0.000</td>
<td>0.039</td>
<td>−0.042</td>
<td>−0.219***</td>
<td>0.001</td>
</tr>
<tr>
<td></td>
<td>(2.178)</td>
<td>(2.439)</td>
<td>(0.906)</td>
<td>(6.275)</td>
<td>(−0.884)</td>
<td>(1.107)</td>
<td>(−0.802)</td>
<td>(−3.513)</td>
<td>(0.965)</td>
</tr>
<tr>
<td>ΔSP(t)</td>
<td>0.007***</td>
<td>−0.009</td>
<td>0.080</td>
<td>−0.029</td>
<td>0.000</td>
<td>−0.047</td>
<td>−0.045</td>
<td>−0.243***</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>(8.843)</td>
<td>(−0.209)</td>
<td>(0.916)</td>
<td>(−2.266)</td>
<td>(−1.230)</td>
<td>(−1.330)</td>
<td>(−0.864)</td>
<td>(−3.944)</td>
<td>(0.227)</td>
</tr>
</tbody>
</table>

Stated: Coefficients, t-Statistics in Parenthesis; Obs.: 316

Significance: **0.001, *0.01, †0.05

### References


