# Assimilation of Oil News into Prices

Tim Loughran
Mendoza College of Business
University of Notre Dame
Notre Dame, IN 46556-5646
574.631.8432 *voice*
Loughran.9@nd.edu

Bill McDonald
Mendoza College of Business
University of Notre Dame
Notre Dame, IN 46556-5646
574.631.5137 *voice*
mcdonald.1@nd.edu

Ioannis Pragidis
Democritus University of Thrace
Department of Economics
University Campus
P.C 69100 Greece
306.946.905.951 *voice*
gpragkid@econ.duth.gr

September 24, 2018

## Abstract

Do investors quickly and rationally react to the content of oil-related news articles revealing supply and demand information? Our paper creates a keyword list of 130 oil-related words and modifiers that enable investors/researchers to measure the information content of oil stories. We find significant short-term overreaction to the text of Dow Jones oil-related news articles. Phrases like *output cut, production cut, shortage,* and *demand up* in lagged news articles are associated with lower oil prices the following trading day. The evidence is consistent with the notion that oil traders overreact to the content of widely-read news articles.

Key words: Textual analysis; news stories; oil prices; overreaction; tone.

We thank Paul Gao, Rita Gunn, Thomas Healy, Michael Hemler, Peter Kelly, Yuan Zhang, and seminar participants at Vanderbilt University, Baruch College, Laval University, and Australian National University for helpful comments.

“The one thing we know for sure about the price of oil is that we can’t predict the price of oil.” Spoken by Paul Appleby, Head of Energy Economics at BP in a 2016 interview[1]

Annual global consumption of oil is measured in trillions of dollars and production quickly responds to changes in market prices. Thus, it is clearly important to know whether these prices are efficient. Since a main source of information concerning oil supply and demand are news articles; media stories are a good starting point in understanding the assimilation of information into prices.

How do investors react to the information content of oil-related news articles? If an article highlights oil oversupply or overproduction, do investors rationally react to the story’s content? As a first step, how would an investor or researcher gauge the information content of the news story? To measure article information content, one would need to tabulate the number of phrases or words that suggest an increase or decrease in oil prices. Clearly, if the oil news article frequently discusses terms such as *glut, oversupply, overproduction,* or *surplus,* the price of oil should be expected to fall after assimilation of the news article. Following the seminal work of Tetlock (2007), one might expect to see a linkage between news article tone/content and the market’s subsequent reaction.

To measure the tone of oil-related news stories, we create a list of 130 oil keywords by reading hundreds of articles available on the Dow Jones Energy Service (DJES) news database. All of our keywords should be expected to affect oil prices. The five most frequently occurring keywords in DJES articles are *recovery, problems, attacks, oversupply,* and *hurricane*. Four of the five listed tokens typically should be associated with higher oil prices while *oversupply* should be linked with lower oil prices. Not all of the keywords have a straight-forward linkage

[1] See https://www.offshoreenergytoday.com/oil-prices-at-100-or-at-20-thats-a-normal-range-bp-head-of-energy-economics-says/.

with oil prices. Many of the keywords are driven by their modifier. That is, keywords like *exports, inventories, output,* and *rigs* should signal lower oil prices if preceded by positive modifiers such as *increase*. Conversely, if a negative modifier (i.e., *decrease*) precedes the keywords listed above, oil prices should tend to rise.

The 130 oil keywords are placed into three categories: (1) 59 words that should be associated with an increase in oil prices; (2) 19 words that should be linked to a decrease in oil prices; and (3) 52 words whose effect is dependent on their modifier. We also create a list of 291 positive and 536 negative modifiers by examining all words that appear four words before and four words following the 52 keywords whose effect is driven by a modifier. The five most common keyword and modifier combinations in our DJES news article sample are *output cut; production cut; demand strong; production increase;* and *output increase.*

Using 41,432 different Dow Jones Energy Service oil-related news articles during the 2000-2016 time-period, we create a proxy for each day's collective tone relating to supply and demand factors. We tabulate how many times a word or phrase appears in an article that implies an increase in oil prices (e.g., *cutback, sabotage,* and *demand strong*). We also count the number of words or phrases that imply a decrease in oil prices (e.g., *glut, recession,* and *production increase*). Each day, we tabulate a content measure of the oil news stories, the *% Tone Index*. The *% Tone Index* for each article equals (number of oil price increasing words or phrases - the number of oil price decreasing words or phrases) / (number of words in the article). We standardize by the number of words in an article because of vastly different story lengths. The *% Tone Index* for each day is the average tone across all DJES oil-related articles appearing on a given day. Over the sample period, the typical trading day has a positive *% Tone Index* value; signifying more words or phrases in stories suggesting higher oil prices.

The daily oil news article tone is calculated for articles reported in the DJES sample from one minute after midnight (12:01 AM) to 2:15 PM each day. The daily settlement price for the New York Mercantile Exchange (NYMEX) Cushing, OK Crude Oil Future Contract 1 is determined by the volume-weighted average price of all trades in the outright contract that are executed between 2:28 PM and 2:30 PM. All times are denominated as Eastern Standard Time (EST). Thus, the lagged news content is from the prior trading day (i.e., right after midnight to 2:15 PM on day t-1). This allows a potential trading strategy to be implemented using prior-day oil news stories that should have been incorporated into oil prices.

As a verification of the 130 oil keywords, the contemporaneous coefficient on the *% Tone Index* is consistently positive in sign and statistically significant in ordinary least squares (OLS) and generalized autoregressive conditional heteroskedasticity (GARCH) regressions with oil price changes as the dependent variable. Thus, oil news articles during the same trading day containing words like *production cut, recovery, shortage,* and *storm* are associated with higher contemporaneous oil prices. This is true with or without sentiment and control variables being present in the regression. Our keywords indeed capture the implied content of the oil news stories.

We find, however, significant overreaction by investors to the lagged news stories. The higher is the count of words or phrases in articles from the prior day that suggest increased oil prices, the lower is oil futures prices on the following day. Oil traders appear to be overreacting to the information content of oil supply and demand information in lagged DJES articles. This overreaction to the content of oil news stories is similar to the evidence presented in Ahern and Sosyura (2015) in the context of merger rumors. As noted by Singleton (2014), if investors have different interpretations of public information, this could cause oil prices to move away from fundamental values. In his Presidential address, Duffie (2010) notes that prices can have a reversal pattern following the release of new information because of various attention costs to trade as well as slow moving capital.

Much of the prior literature includes the sentiment of the news articles or columns (see Tetlock (2007), Tetlock, Saar-Tsechansky, and Macskassy (2008), Gurun and Butler (2012), and Liu and McConnell (2013)). The daily percentage of negative words in news articles is included in the regressions as one of our control variables. The variable *% Negative* is the fraction of words in oil new articles that are on the Loughran and McDonald (2011) negative word list. We use the updated version of their word list, containing 2,355 negative words. The most frequently occurring Loughran and McDonald (2011) negative words in our sample of oil news articles are *cut, decline, concerns, against*, and *losses*. Thus, we control for the general sentiment of the oil news article in the regressions.

Consistent with a slow diffusion of non-oil specific news into prices, lagged *% Negative* is significantly related to subsequent oil prices in the OLS regressions. The more negative the generic sentiment contained in the lagged daily news article, the lower are future oil prices. This underreaction evidence is in line with the finding of Tetlock et al. (2008) in the context of firm specific news articles and individual stock returns. The negative relation between oil-specific news and subsequent oil prices, in contrast to a positive relation for the generic tone of the news, suggests a complex relation between oil prices and the information environment.

Our paper contributes to the literature on the assimilation of macroeconomic information into asset prices. Instead of focusing on the diffusion of idiosyncratic information into individual stock prices as in much of the prior literature, we examine how systematic information is incorporated into oil prices.

## 1. Literature Review

Assimilation of information contained in news articles into security prices has been a well-researched area in finance and accounting. Much of the premise of market efficiency dwells on how quickly market participants incorporate information into security prices. A slow diffusion of public information into security prices would imply a market inefficiency while an immediate incorporation would indicate a well-functioning financial market. As noted by Duffie (2010), price reversals following a supply or demand shock could be caused by a combination of slow moving capital and investor inattention. It often takes time for investors to digest new information. There are a number of papers focusing on the role the media plays in channeling information into prices.

For example, Ahern and Sosyura (2015) analyze how investors react to merger rumors appearing in the media. They find empirical evidence consistent with overreaction to published newspaper rumors by relatively unsophisticated traders. That is, investors push up the target stock price too much following the initial press report; investors overestimate the probability of the takeover rumor being true. Ahern and Sosyura (2015) find that sensational news articles about merger rumors actually affect the prices of publicly-traded equity.

The path-breaking paper by Tetlock (2007) examines how the tone of an influential newspaper column can affect subsequent market level stock returns. He finds that pessimistic tone in the *Wall Street Journal's* "Abreast of the Market" is linked with next day market returns. More pessimistic column tone is associated with an 8.1 basis point lower return on the Dow Jones Industrial Average the following trading day. Importantly, Tetlock (2007) is probably the first paper to document "that news media content can predict movements in broad indicators of stock market activity" (page 1140).

A slight underreaction to negative tone contained in firm specific news articles for S&P 500 companies is found by Tetlock et al. (2008). They find that lagged articles with a higher frequency of negative words is associated with significantly lower stock returns for the firm on the following day. As is similar with other textual analysis papers, the overall economic effect is somewhat muted. The three authors find that a one-standard deviation increase in negative words is linked with 3.20 basis points lower abnormal returns on the following day.

Solomon, Soltes, and Sosyura (2014) examine the impact on investor’s allocation decisions by media coverage of mutual fund holdings. They find that media attention helps direct investor flows into mutual funds. Mutual funds holding past winners, as highlighted by the media, experience significantly higher investor flows than comparable funds holding fewer visible past winners.

Can media coverage affect key managerial decisions? Liu and McConnell (2013) find that firm corporate capital allocation decisions are affected by the tone of media coverage for the proposed investment. They study the firm’s decision to complete or abandon a proposed large corporate acquisition. As the tone of the media’s stories becomes more negative, managers are more likely to walk away from the planned acquisition.

Engelberg and Parsons (2011) and Gurun and Butler (2012) analyze the impact of local media on trading behavior of local investors. Illustrating the importance of the news media, Peress (2014) reports a dramatic decline in trading volume for stocks during a national newspaper strike in several different nations. Clearly, media story content affects investor’s trading behavior and directs their attention to specific information.

Social media postings have been shown to affect subsequent trading volume and stock prices. Antweiler and Frank (2004) find that a positive shock to Internet stock message boards’ postings predicts negative stock returns on the next day. In addition, the two authors report that both the level of message posting and disagreement among the stock postings are linked with subsequent trading volume. Similarly, Das and Chen (2007) find an association between stock message board postings sentiment and stock index levels. Measuring the sentiment of Seeking Alpha articles and commentary, Chen, De, Hu, and Hwang (2014) find a slow assimilation of content into stock prices. The more negative the tone in the articles and commentaries for a particular stock, the lower are subsequent abnormal returns.

Because of the importance of oil for the economy, a number of papers have linked oil prices with the performance of the overall economy. The classic paper by Hamilton (1983) finds that oil shocks are a contributing factor to U.S. recessions prior to 1972. Of the eight post-World War II U.S. recessions, seven were preceded by dramatic increases in oil prices. Hamilton (1996) extends his original sample period and continues to find a strong linkage between oil shocks and U.S. recessions. During our sample period, the pattern documented by Hamilton (1983) is also present. In the months preceding the Great Recession of 2008, oil prices per barrel jumped from $65.08 on June 1, 2007 to $145.08 on July 11, 2008 (an increase of more than 100%) before the economy went into a severe tailspin.

Following the work of Hamilton (1983), a number of researchers have examined the linkage between oil prices and key economic variables. For example, researchers have linked oil prices with equity prices (Nandha and Faff (2008)); exchange rates (Chen and Chen (2007) and Amano and Van Norden (1998)); international trade (Backus and Crucini (2000)); and interest rates (Cologni and Manera (2008)). Kilian and Vega (2011) find that oil prices, unlike other financial asset prices, do not respond instantaneously to U.S macroeconomic news. In contrast, Elder, Miao, and Ramchander (2013), using intraday high frequency data, find that economic news impacts oil prices, but the impact dissipates relatively quickly. Barrero, Bloom, and Wright (2017) find that volatility in oil prices is linked with short-run uncertainty.

Our paper is not the first to examine the link between investor attention and oil prices. Instead of using media stories concerning oil, a recent paper by Han, Lv, and Yin (2017) uses the Google search volume index (SVI) to measure investor interest in oil prices. However, the list of their 85 search terms (see their Table 1) captures investor interest in much more than just oil. A number of their selected search terms, like *bargain holidays, car donate, charity,* *depreciation*, *social security office, Detroit bankrupt,* and *donation* should not be expected to generally affect oil supply and demand. The premise of their paper, examining the linkage between oil prices and investor attention is very reasonable. Yet, quite a few of their keywords cast too wide a net on what investors are researching.

## 2. Data and sample creation

### 2.1 Dow Jones Energy Service News Articles

The Dow Jones Energy Service (DJES) news database is the source for our oil market news from January, 2000 to September, 2016. The DJES text corpus contains real-time oil market news, commentary, and analysis. The service's subscribers include traders, analysts, and industry professionals. The DJES database includes news articles from the *Wall Street Journal* (*WSJ*), *Barron's,* and *MarketWatch*. The *WSJ* is the most highly circulated business newspaper in the United States.

In order to remove DJES news articles that are not meaningful (i.e., just a few words), not related to oil prices, or are company specific, we use the following filters based on article character count and specific word tokens. First, we exclude all news articles containing fewer than 180 characters. Next, we select only news articles with specific oil words in their headlines. These oil words/acronyms (in either upper or lower case) are *oil, crude, OPEC, Brent, and WTI (West Texas Intermediate).* The relative occurrence of each accepted headline are 52.25% (*oil*), 26.79% (*crude*), 10.26% (*OPEC*), 9.76% (*Brent*), and 0.94% (*WTI*). The DJES database contains a voluminous set of articles reporting oil company press releases. To remove articles related to specific company news, we exclude stories containing the name of oil companies or headlines containing the following words: *Ltd., Co., Inc.,* or *Corp.*

Following Loughran and McDonald (2011), the first step in parsing the DJES articles is matching the tokens with the 2016 updated Loughran-McDonald Master Dictionary (https://sraf.nd.edu/textual-analysis/resources/). As noted in Loughran and McDonald (2016), the researcher needs to specify which collection of characters (i.e., tokens) are identified as words in the analysis. The Loughran-McDonald Master Dictionary removes all numbers, single letters, acronyms, and proper nouns from our analysis. Since our context relates specifically to oil news, we add to the Loughran-McDonald Master Dictionary words/acronyms that frequently appear in oil news articles. Our additions to the LM Master Dictionary are *API (American Petroleum Institution), Arab, Arabia, Brent, EIA (U.S Energy Information Administration), IEA (International Energy Agency), Iran, Iraq, Kuwait, NYMEX, OPEC, Saudi, UAE (United Arab Emirates),* and *WTI (West Texas Intermediate).*

Figure 1 plots the annual number of DJES oil-related articles during the January, 2000 to September, 2016 time period. There is a wide variation in the annual number of oil news articles. Three calendar years (2005, 2006, and 2008) have more than 3,600 articles. Conversely, in the months following the 9/11 terrorist attack there was a significant drop in the number of oil related articles. In the six months following September, 2001, the number of monthly DJES stories meeting our criteria were 385, 119, 62, 38, 53, and 53. Both calendar years 2002 and 2003 had an unusually low number of DJES articles. In Appendix A, we include four DJES new articles with varying content. The articles provide examples of how the keywords typically appear in DJES stories. Our final sample is 41,432 unique DJES news articles relating to the oil market.

Figure 2 reports the number of DJES oil articles meeting our criteria by hour during the day. There is a clear spike in the number of oil news articles at 11:00 AM. London is five hours ahead of New York. Typically, European closing oil prices are based on activity from roughly 4:00 PM to 4:30 PM London time. This effectively marks the end of the day for European oil traders, and this corresponds to the 11:00 AM EST spike in articles. A significant fraction of energy trading is transatlantic and it is difficult for traders to do deals when the other continent is out of the office. As a result, the sweet spot for physical trading activity is typically late morning to evening London time (i.e., early morning to early afternoon U.S. time) when both continents are present in the office.

The unofficial start of business for London physical oil trading is roughly 9 AM. Given the time difference between the continents, this is 4:00 AM New York time. Corresponding to this start time, Figure 2 shows a jump in the number of news articles at 4:00 AM EST. Around 6:00 AM EST, most of the oil trading benches in New York start coming to life and market participants are checking out what information has changed overnight. Even though oil trading occurs 24 hours a day, according to an U.S. oil trader we spoke with, there is still a "premarket" period in the early morning. This is before much of the physical trading picks up, and when overnight information gets discussed and digested. Consistent with this, Figure 2 shows a rise in the number of articles at 6:00 AM. The very early morning hours of 1:00 AM and 2:00 AM, as might be expected, have the lowest number of DJES oil articles.

### 2.2 Oil price data

For the daily oil price data, we use futures oil prices during January 3, 2000 to September 30, 2016. The oil price data is obtained from the U.S Energy Information Administration.[2] The contract we focus on is the NYMEX Cushing, OK Crude Oil Future Contract 1, in U.S. dollars per barrel. The delivery month for Contract 1 is the calendar month following the trade date. The paper uses the futures price because its contracts are more liquid than the spot oil prices. As noted in Alquist, Kilian, and Vigfusson (2013), most financial researchers and central banks use oil futures as the best available forecast of oil spot prices. During our sample, the daily settlement price for the NYMEX Cushing, OK Crude Oil Future Contract 1 is determined by the volume-weighted average price of all trades in the outright

[2] Available at https://www.eia.gov/dnav/pet/pet_pri_fut_s1_d.htm.

contract that are executed between 2:28 PM and 2:30 PM EST, rounded to the nearest tradable tick. All times are Eastern Standard Time (EST).

The time-series pattern of the oil futures contract prices during the sample period is reported in Figure 3. As presented in the figure, oil prices during this period were volatile. The peak for oil was on July 3, 2008 (hitting $145.29 per barrel). Once the U.S. economy sharply contracted during fourth quarter of 2008, the price per barrel quickly fell to only $33.87 on December 19, 2008. The other sharp decline in oil prices occurred in 2014-2015 when Saudi Arabia kept its production steady in face of decreasing Chinese demand, and U.S. and Canadian oil fields increased their production through the use of more effective fracking methods. These factors caused oil to fall from $100.27 on July 30, 2014 to $45.15 by January 26, 2015 (a 55% decline). All daily sample oil prices are converted into a daily nominal continuously compounded return series for crude oil, i.e., $\mathit{Oil\ Price\ Return}_t = 100*\ln(\mathit{Price\ of\ Oil}_t/\mathit{Price\ of\ Oil}_{t-1})$ for $t = 1, 2, \ldots, T$.

### 2.3 Keyword creation methodology

To create our oil keyword list, we examine over 500 randomly selected DJES news articles. In order to create an exhaustive word list, we select all concept words that are likely related to either the supply side or the demand side of the oil market. We rely on economic theory in our selection of the keywords.

Table 1 reports all 130 keywords in the resulting lexicon derived from articles in the Dow Jones Energy Service news database. The first column lists the 59 keywords expected to increase oil prices (examples include *closures, delay, explosion, hurricane, outage,* and *upheaval*). The second column in Table 1 reports the 19 keywords expected to decrease oil prices (e.g., *discoveries, glut,* and *oversupply*). The last column reports the 52 keywords that need to be signed by a modifier (e.g., *cargo, economy, output,* and *production*) to gauge their predicted impact on oil prices.

To create the list of positive and negative modifiers, we programmatically extract all words appearing four words before and four words after the 52 keywords that need to be signed after removing stop words from the text. By carefully examining all related words, we explicitly take into account any inflections or different forms of the words of interest. This paper uses the Loughran and McDonald generic stopword list of 121 words.[3] We remove the word "up" from their generic stopword list since this token is one of our positive modifiers.

We then identify tone modifiers for each of the 52 keywords that need to be signed. The tone modifiers can be either positive or negative. In doing so, we use collocation and identify the verbs, adjectives, and adverbs that surround the words in our newly created market word list. These words will identify, in each case, whether the tone is positive or negative as it relates to oil prices. As an example, a *production fall* typically increases oil prices while a *stockpile surge* usually lowers oil prices. When a tone modifier is not detected in the specified range of plus or minus four words, we ignore the keyword. The 291 positive modifiers (Panel A) and the 536 negative modifiers (Panel B) are reported in Appendix B.

As a prima facie test of reasonableness, it is critically important to present the keywords that strongly impact our analysis. Panel A of Table 2 lists the 30 most frequently occurring keywords in the DJES oil news sample. These 30 keywords account for almost 89% of the cumulative counts. The top eight occurring oil keywords are *recovery, problems, attacks, oversupply, hurricane, glut, concerned,* and *disruptions*.

Panel B of Table 2 lists the 30 most frequently occurring keywords and modifiers. The top five keyword and modifier combinations are *output cut; production cut; demand strong; production increase;* and *output increase*. For Panel B, the reported cumulative percentiles

[3] Available at https://sraf.nd.edu/textual-analysis/resources/.

reflect when the modifier appears before or after the oil keywords. Thus, the combined counts of *cut output* and *output cut* account for 2.68% of all the keyword-modifier combinations. The oil keywords appear to be capturing what was intended.

### 2.4 Control variables and *% Tone Index*

The *% Tone Index* is calculated as: (number of oil price increasing phrases - number of oil price decreasing phrases per document) / (number of words in the article). We divide by the number of words in order to normalize the *% Tone Index,* as the number of words per article varies substantially. We aggregate across all oil news articles per day to create the daily *% Tone Index* value. The daily *% Tone Index* is tabulated during the time from immediately after midnight (12:01 AM) to 2:15 PM (fifteen minutes before the daily settlement price). Our key independent variable in the regressions is the lagged value of the *% Tone Index* calculated prior to the daily settlement of the oil futures contract (i.e., 2:28 PM to 2:30 PM). We are interested in how lagged news articles affect subsequent oil prices.

Our regression control variables are *% Negative*, number of daily oil articles on day t-1, a trade-weighted U.S dollar index, the spot price of gold, 10-year Treasury constant maturity rate, and the VIX index. *% Negative* is the fraction of news oil articles words that are on the updated Loughran and McDonald (LM, 2011) negative word list. Although there are an assortment of word lists created to measure extreme emotion in earnings conference calls (Larcker and Zakolyukina (2012)) or financial constraints (Bodnaruk, Loughran, and McDonald (2015)), their negative word list should capture the general sentiment of the news articles. The LM negative word list has been used to gauge sentiment of a wide range of documents including S&P credit rating action reports, earnings conference calls, public comment letters submitted by banks, annual reports, front-page news articles from the *WSJ*, and Form 8-Ks.[4]

The updated LM negative word list contains 2,355 negative words (obtained from https://sraf.nd.edu/textual-analysis/resources/). LM create the list by examining word usage in annual reports (i.e., Form 10-K). Since our oil news article context is slightly different, we drop three words (*late, closing,* and *closed*) from the list of LM negative words. The token *late* is the second most frequently occurring LM negative word in our oil news article corpus. Clearly, this word does not have negative meaning in its most frequent context within news articles (i.e., *late* today, *late* March, *late* this afternoon, *late* Tuesday, etc.). "*Closing* oil prices" and "the market *closed*" are also commonly appearing phrases that do not indicate pessimistic language in the oil news articles. After dropping these three words, the ten most frequently occurring Loughran and McDonald (2011) negative words in the oil news articles are *cut, decline, concerns, against, losses, weaker, weak, break, sharply,* and *concern*.

To control for the intensity of released information, the number of oil news articles appearing on day t-1 is included in all regressions. All other control variable data (i.e., a trade-weighted U.S dollar index, spot price of gold, 10-year Treasury constant maturity rate, and VIX Index) are obtained from the Federal Reserve Bank of St. Louis economic database (https://fred.stlouisfed.org). Like the crude oil futures price, the spot price of gold is converted into a daily nominal percentage return, while the other control variables are expressed as first differences of indexes to account for the unit root. We focus on percentage return and on first differences because the level of the dependent and control variables is not stationary based on the prior literature. Exceptions are the *% Tone Index* and the *% Negative,* which are stationary.

[4] See Agarwal, Chen, and Zhang (2016), Froot, Kang, Ozik, and Sadka (2017), Gissler, Oldfather, and Ruffino (2016), Law and Mills (2015), Manela and Moreira (2017), and Segal and Segal (2016).

The summary statistics (Panel A) and correlations (Panel B) for our key variables are reported in Table 3. The mean and median *% Tone Index* has respective values of 0.20% and 0.18%. Thus, the typical DJES news article has more oil price increasing words or phrases than decreasing oil price text. The average oil price per barrel during our time period is just over \$63. The 5th percentile for oil price (\$26.45) is a fraction of the 95th percentile value (\$105.73). The *% Negative* variable's mean (1.59%) and median (1.58%) values for oil news articles are similar to the pessimism contained in broader samples of news articles and slightly higher than the values reported in annual reports. For example, Gurun and Butler (2012) report a similar mean value (1.69%) for the fraction of negative words appearing in a composite news article in a given month. In contrast, Loughran and McDonald (2011) report mean and median *% Negative* values of 1.39% and 1.36% respectively for a sample of Form 10-Ks during the 1994-2008 time period. On average, about 10 oil news articles appear in the DJES database each day.

Panel B of Table 3 reports the correlations between the transformed key variables. *Oil returns* are positively correlated with the *% Tone Index* (0.07) and negatively correlated with *% Negative* (-0.06). The positive correlation between the *% Tone Index* and changes in oil prices is consistent with the *% Tone Index* capturing supply and demand content from oil news articles on the same day. It is important to note that the *% Tone Index* and *% Negative* are positively correlated with each other (0.16).

At first pass, it might seem surprising that the *% Tone Index* and *% Negative* are positively linked. However, a number of the most frequently occurring oil keywords are also on the Loughran and McDonald (2011) negative word list. For example, *problems, disruptions, shortage,* and *delayed* are some of the most commonly occurring keywords which imply an increase in oil prices, as reported in Panel A of Table 2. All of these mentioned words are also on the negative word list. As noted earlier, the token *cut* is the most frequently appearing negative word in our oil article corpus. The phrases *output cut* and *production cut* are the top two keyword and modifier combinations. Both of these phrases imply an increase in oil prices that will elevate the *% Tone Index* value. As the text of oil news articles discusses *disruptions, problems,* and *output cut,* both the *% Tone Index* and *% Negative* will rise in value.

## 3. Empirical Results

### 3.1 OLS Regressions

In an attempt to be complete, our analysis uses both OLS and GARCH regressions to examine the relation between oil prices and the content of DJES oil news articles. As a first pass, we present regression results of various lags of tone on changes in oil prices with no sentiment control variables. Table 4 presents the results from estimating the OLS regressions of *Oil Price Returns* on the lagged tone variables:

$$\begin{aligned} \textit{Oil Price Returns}_{t} &= \alpha + \beta_1 \textit{\% Tone Index}_{t} + \beta_2 \textit{\% Tone Index}_{t-1} + \beta_3 \textit{\% Tone Index}_{t-2} + \\ &\quad \beta_4 \textit{\% Tone Index}_{t-3} + \beta_5 \textit{\% Tone Index}_{t-4} + \varepsilon_t \, , \qquad (1) \end{aligned}$$

where *% Tone Index* is defined as (number of oil price increasing phrases - the number of oil price decreasing phrases) / (number of words in the article) for all DJES oil news articles in a given day. The dependent variable, *Oil Price Returns*$_t$, is defined as 100*ln (*Price of Oil*$_t$ / *Price of Oil*$_{t-1}$). As noted by the subscripts, our independent variables are known before the oil contract opening, with the exception of the contemporaneous *% Tone Index*$_t$. Thus, the lagged right hand side variables are predicting future oil prices. In all of our regressions, calendar year dummies, lagged values of the dependent variable up to four lags, and the number of daily oil news articles on day t-1 are usually included. The standard errors include a Newey-West correction for heteroscedasticity and autocorrelation up to five lags. Robust *t*-statistics, in parentheses, are reported below the coefficient estimates. All of the regressions have 3,568 trading day observations.

In column (1) of Table 4, the only independent variable, besides calendar year dummies, the lagged dependent variable, and the number of daily oil news articles, is the lagged *% Tone Index* value. The coefficient on *% Tone Index*$_{t-1}$ is negative (-0.16) and statistically significant at the 1% level (*t*-statistic of -2.39). Notably, the more language in the DJES oil news articles suggesting higher oil prices, the *lower* are subsequent oil futures prices. This evidence is consistent with overreaction on the part of oil traders and investors. In the second column of Table 4, we exclude the lagged oil returns from the regression with only a very minor effect on the *% Tone Index*$_{t-1}$ coefficient value (-0.16 versus -0.17). Thus, the significant levels of our main variable is not being driven by having lagged oil returns in the same regression.

In the third column of Table 4, we include lagged values of *% Tone Index* during day t-1 to day t-4. As shown by Ahern and Sosyura (2015) and Antweiler and Frank (2004), investors have been known to overreact to the content of news articles and stock message board postings. The only significant variable at the 1% level is the *% Tone Index* from day t-1. There is only a minor change in the *% Tone Index*$_{t-1}$ coefficient value when other lagged *% Tone Index* variables are included in the regression (-0.16 versus -0.17).

Column (4) of Table 4 provides a simple test of the validity of our *% Tone Index* measure. Although investors cannot trade on this information, column (4) of Table 4 includes the contemporaneous *% Tone Index* value in the regression. Unlike Kilian and Vega (2011), who find no statistically significant impact of macroeconomic news on oil prices, the results here indicate a strong instantaneous effect of oil related news. The contemporaneous *% Tone Index*$_t$ has a positive (0.33) and statistically significant coefficient value (*t*-statistic of 4.30). As news articles during a trading day have more words/phrases like *production cut, shortage, demand up*, and *recovery*, the higher are oil prices at the close of the same trading day. When all the lagged values of *% Tone Index* are included in the last column, both *% Tone Index*$_t$ and *% Tone Index*$_{t-1}$ retain their coefficient sign and significance levels. The contemporaneous *% Tone*

*Index* has a positive coefficient (0.37) while lagged *% Tone Index* has a negative coefficient value (-0.21).

It is important to point out that all the $R^2$ values in our Table 4 regressions are quite low. For example, the $R^2$ value is 0.34% in the first regression of the table where lagged *% Tone Index* is the main variable of interest. These low values reflect the difficulty in predicting changes in oil prices. In the context of using negative words contained in news stories to predict stock returns, Tetlock et al. (2008) report adjusted $R^2$ values of only 0.24%.

In the next set of OLS regressions, sentiment and additional control variables are added to the analysis. In Table 5, we estimate the OLS regressions of *Oil Price Returns*:

$$\begin{aligned} \textit{Oil Price Returns}_t = \alpha + \beta_1 \textit{\% Tone Index}_t + \beta_2 \textit{\% Tone Index}_{t-1} + \beta_3 \textit{\% Negative}_t + \\ \beta_4 \textit{\% Negative}_{t-1} + \beta_5 \Delta(Exchange\ rate_{t-1}) + \beta_6 \Delta(Gold_{t-1}) + \\ \beta_7 \Delta(10yr\ rate_{t-1}) + \beta_8 \Delta(VIX_{t-1}) + \varepsilon_t , \qquad (2) \end{aligned}$$

where *% Tone Index* is defined as before and *% Negative* variable is the fraction of words in the daily DJES news articles that are in the Loughran and McDonald (2011) negative word list. The control variables are the change in exchange rates, gold price per ounce, 10-year Treasury rates, and the VIX Index from the prior trading day.

The OLS regression results with only the lagged control variables as the independent variables (besides calendar year dummies, the lagged dependent variable, and the number of oil news articles on day t-1) are presented in column (1) of Table 5. Among the control variables, only *% Negative* is statistically significant. The coefficient on *% Negative* has the expected sign (-0.28) and a *t*-statistic of -3.78. This implies that the more negative the sentiment in the trailing oil news stories, the lower are subsequent oil prices. Even in a different context then in which the negative word list was created, the fraction of negative LM words has predictive power. Similarly, Tetlock et al. (2008) find that more pessimistic newspaper tone concerning a particular firm, using the negative word category of the Harvard IV psychosocial dictionary, is associated with lower stock returns for the company on the following day.

In column (2) of Table 5, the lagged *% Tone Index* is added with the control variables. In the presence of the control variables, lagged *% Tone Index* has a negative and statistically significant coefficient value. The coefficient on *% Negative* continues to be significant with only a minor change in its coefficient value (-0.25 versus -0.28). This provides evidence that our *% Tone Index* variable is independent of the sentiment of the news oil articles.

When the contemporaneous values of the *% Tone Index* and *% Negative* are added to the column (3) OLS regression in the presence of the control variables, the two variables remain statistically significant. The *% Tone Index$_t$* coefficient value of 0.39 has a *t*-statistic of 5.00. The coefficient on the contemporaneous fraction of negative words has a negative value (-0.34) and is significant at the 1% level. As the content of oil-related news articles suggests that oil prices should increase, the value of price per barrel does increase even after controlling for the overall sentiment of the story.

The last column of Table 5 includes all the variables in an OLS regression. All the key variables retain their coefficient signs and their statistical significance levels. *% Tone Index$_t$* has a positive coefficient value (0.41) while the *% Tone Index* in day t-1 has a negative coefficient value (-0.17). Thus, controlling for contemporaneous DJES oil news content, the lagged value of *% Tone Index* implies an overreaction on the part of oil traders and investors. Both *% Negative$_t$* and *% Negative$_{t-1}$* have negative coefficient values, respectively -0.32 and -0.20. The significant value on lagged *% Negative* is consistent with a slow assimilation of oil news article sentiment into oil prices. Clearly, traders do not always properly and instantaneously incorporate all news story content into commodity prices.

### 3.2 GARCH Regressions

In Tables 6 and 7, we rerun the analysis in a GARCH regression setting. GARCH models potentially improve on the OLS framework in financial markets where volatility can change over time. The volatility of oil prices, like other financial assets, does change dramatically in periods of turbulence. In stable economic periods, oil prices are usually much less volatile. The GARCH regression framework attempts to minimize errors in forecasting by accounting for errors in prior forecasting.

In Table 6, we report GARCH (1,1) regressions between *% Tone Index* and changes in oil prices. As before, the dependent variable is *Oil Price Returns*$_t$ (defined as 100*ln (*Price of Oil*$_t$/*Price of Oil*$_{t-1}$)), calendar year dummies, four lags of the dependent variable, and the number of oil news articles on day t-1 are included in all regressions. Table 6 mirrors the OLS analysis of Table 4 under the GARCH framework. Contrasting the OLS results with the GARCH regressions, one can see more negative coefficient values for *% Tone Index*$_{t-1}$ under the GARCH framework. In column (1), the coefficient on *% Tone Index*$_{t-1}$ is -0.17 (*z*-statistic of -3.17). The coefficient on *% Tone Index*$_{t-1}$ remains significant at the 1% level when other lagged values of the variable are added to the regression reported in column (2).

As before, the coefficient on contemporaneous *% Tone Index* in column (3) is positive (0.22) and statistically significant (*z*-statistic of 3.65). More words like *closures, cutbacks, shutdowns,* and *strikes* in a DJES news article are associated with higher oil prices on the same trading day. When both contemporaneous *% Tone Index* and lagged *% Tone Index* values are included in the same regression (column (4) of Table 6), both the day t and the day t-1 coefficients of the variables are significant. *% Tone Index*$_t$ has a coefficient value of 0.24 while *% Tone Index*$_{t-1}$ has a negative coefficient value (-0.19). Once again, the evidence for the lagged *% Tone Index* variable is consistent with overreaction on the part of investors.

Once the sentiment and control variables are introduced in the Table 7 GARCH regressions, the *% Tone Index*$_t$ and *% Tone Index*$_{t-1}$ variables remain significantly associated with subsequent oil prices. In column (2), the coefficient on *% Tone Index*$_{t-1}$ is -0.14 and on *% Negative*$_{t-1}$ is -0.15, with respective *z*-statistics of -2.50 and -2.35. In the third column of Table 7, the contemporaneous values of the *% Tone Index* and *% Negative* are statistically significant at the 1% level and have the expected coefficient signs. When all the independent variables are included in the final column of Table 7, *% Tone Index*$_t$, *% Tone Index*$_{t-1}$, and *% Negative*$_t$ all have significant coefficient values. The lagged *% Tone Index* variable has a coefficient of -0.17 (*z*-statistic of -3.02). In this regression, the coefficient on *% Negative*$_{t-1}$ is only significant at the 10% level of significance (*z*-statistic of -1.88). Thus, controlling for contemporaneous sentiment and article content, the prior day *% Tone Index* suggests overreaction by investors in the oil futures market.

### 3.3 Profitability of a Simple Trading Strategy

Can an investor successfully profit by selling oil contracts when the *% Tone Index* is high and buying contracts when the index is low? That is, can investors profit from the documented short-term overreaction to DJES oil news stories? We adopt a simple strategy based on expected returns on the following day to assess if the forecasting result has any economic value besides its statistical significance.

An investor can calculate the daily *% Tone Index* on day t-1 and if the value lies in the bottom 20% of the prior period’s distribution, the investor purchases a contract on day t-1 and closes the position by selling the contract one day later. Conversely, if the *% Tone Index* lies in the top 20% of the prior period’s distribution, the investor sells an oil futures contract on day t-1 and buys it back in the next day. If the lagged *% Tone Index* lies between the top and bottom

20%, the investor does no trading on that particular day. We assume that this round-trip transaction incurs a one basis point trading cost.

The trading strategy distribution is based on the prior year's period of January to December. Each year, we determine the top and bottom 20% of the lagged *% Tone Index* and use these values to construct our trading strategy during the following year. Thus, we use the distribution of the lagged *% Tone Index* from calendar year 2000 to construct our strategy for year 2001 and then year's 2001 distribution for the investment strategy for calendar year 2002 and so forth. In Table 8, we report the yearly outcome of the investment strategy.

Using the top and bottom 20th percentiles of the lagged *% Tone Index* as the investment criterion, the strategy sells and buys contracts during the 2001-2016 period 729 and 705 times, respectively. The short position is much more profitable than the long position; the trading strategy has positive returns of 21 and 2 basis points on days in which the hypothetical strategy sells and buys the contract, respectively. Only the average short position return is statistically significant (t-statistic of 2.18). Over the 16 years, a positive return is generated from the short position a total of 12 times. Three of years with negative returns for the short position have values of 4 basis points or less. For the long position, 10 of the 16 years generate positive abnormal returns after controlling for transaction costs.

To place the Tone Index strategy returns in context, Tetlock (2007) finds the impact of a one standard deviation change in pessimism in the *WSJ's* "Abreast of the Market" column on the following day's Dow Jones returns is only 8.1 basis points. Adding both the short and long positions together, our strategy generates a total return of 23 basis points.

### 3.4 Linkage between *% Tone Index* and oil ETF returns

We have documented the relation between the tone of oil news articles and NYMEX oil futures contract prices. An obvious extension would be to see if there is a linkage between oil article tone and the prices of publicly-traded oil companies. However, because oil refineries (i.e., downstream companies) suffer when oil prices rise, while crude discoverers (i.e., upstream companies) benefit from higher oil levels, a market capitalization weighted index of oil buyers and sellers might not be expected to be impacted by oil article tone. Instead, we will focus our attention on a widely-traded oil exchange-traded fund (ETF), United States Oil Fund, ticker: USO. United States Oil Fund invests primarily in listed crude oil futures contracts and other oil-related futures contracts, and according to its website may invest in forwards and swap oil contracts. The time series of U.S. Oil Fund returns is obtained from Wharton WRDS starting at the ETF's inception in April of 2006 through the end of our time period in September of 2016. In this time period, there are 2,288 available trading days for the U.S. Oil Fund.

Table 9 reports the regression results with U.S. Oil Fund ETF returns as the dependent variable. As before, calendar year dummies, lagged values of the dependent variable, and the number of daily oil news articles on day t-1 are included in each regression. The first two columns of the table report the OLS regressions while columns (3) and (4) are GARCH regressions. In all four regressions, the variable *% Tone Index*$_{t-1}$ is consistently associated with lower subsequent U.S. Oil Fund returns. The more phrases and words in a DJES oil news article implying higher oil prices, the lower are subsequent returns on the oil-based ETF. As in the prior setting, in both OLS and GARCH regressions, the contemporaneous *% Tone Index* and *% Negative* coefficients have the expected sign and are statistically significant. Thus, our main relation between oil news article tone and subsequent oil returns also exists in the oil EFT setting.

## 4. Conclusion

Do investors and traders rationally react to the content of oil news articles? As an initial contribution to the literature, we create a list of 130 oil keywords to assist researchers in gauging the content of oil-related news articles. The keywords are selected by reading hundreds of Dow Jones Energy Service (DJES) oil news articles. The 130 keywords are placed into three categories: (1) 59 words that should be associated with an increase in oil prices; (2) 19 words that should be linked to a decrease in oil prices; and (3) 52 words whose effect is dependent on their modifier. The three most frequently occurring oil keywords in the news articles are *recovery, problems,* and *attacks*.

For the 52 keywords that need to be signed by a modifier, we create a list of 291positive and 536 negative modifiers. Our keywords and modifiers allow us to measure the content of 41,432 unique DJES news articles during 2000 to 2016. For each day during the time period 12:01 AM EST to 2:15 PM EST, we create a *% Tone Index* to measure the content of DJES oil news articles. The *% Tone Index* is defined as (number of oil price increasing words or phrases - number of oil price decreasing words) / (number of words in the article). The *% Tone Index* is created prior to the 2:28 PM to 2:30 PM EST daily settlement price for the NYMEX oil futures contract.

Consistent with the prior research, we find that media news stories affect security prices. In both the OLS and GARCH regressions with oil returns as the dependent variable, the contemporaneous *% Tone Index* has a statistically significant positive coefficient. This shows that the 130 keywords are capturing the content of DJES oil news stories. Likewise, the contemporaneous fraction of Loughran and McDonald (2011) negative words in the news article is negatively associated with contemporaneous oil prices. Higher counts of negative words like *cut, decline,* and *concerns* are linked with lower oil prices on the same day.

For the one-day lagged *% Negative* variable in OLS regressions, its coefficient value in the presence of contemporaneous *% Tone Index* and *% Negative* variables is negative and statistically significant. Thus, at least in the OLS setting, investors are slow to incorporate the broader negative article sentiment into oil prices. Consistent with the evidence of Singleton

(2014), differing investor opinions concerning the content of new information can lead oil prices away from their fundamental values.

When a one-day lagged value of the *% Tone Index* is included in the OLS or GARCH regressions, its coefficient is negative and statistically significant. Thus, the higher (lower) the counts of words or phrases implying an increase in oil prices, the lower (higher) are subsequent oil prices. The negative coefficient on lagged *% Tone Index* remains significant when contemporaneous *% Tone Index* and *% Negative* values as well as control variables are included in the regressions. Investors are overreacting to the content of oil-related DJES articles. Our results add to the body of empirical evidence documenting the behavioral phenomenon of overreaction in a liquid and active market in a different setting from equity markets.

## References

Agarwal, S., Chen, V.Y. and Zhang, W., 2016. The information value of credit rating action reports: A textual analysis. *Management Science* 62: 2218-2240.

Ahern, K.R. and Sosyura, D., 2015. Rumor has it: Sensationalism in financial media. *Review of Financial Studies* 28: 2050-2093.

Alquist, R., Kilian, L. and Vigfusson, R.J., 2013. Forecasting the price of oil. *Handbook of Economic Forecasting* 2: 427-507.

Amano, R.A. and Van Norden, S., 1998. Oil prices and the rise and fall of the US real exchange rate. *Journal of International Money and Finance* 17: 299-316.

Antweiler, W. and Frank, M.Z., 2004. Is all that talk just noise? The information content of internet stock message boards. *Journal of Finance* 59: 1259-1294.

Backus, D.K. and Crucini, M.J., 2000. Oil prices and the terms of trade. *Journal of International Economics* 50: 185-213.

Barrero, J.M., Bloom, N. and Wright, I., 2017. Short and long run uncertainty, Stanford University working paper.

Bodnaruk, A., Loughran, T. and McDonald, B., 2015. Using 10-K text to gauge financial constraints. *Journal of Financial and Quantitative Analysis* 50: 623-646.

Chen, S.S. and Chen, H.C., 2007. Oil prices and real exchange rates. *Energy Economics* 29: 390-404.

Chen, H., De, P., Hu, Y. and Hwang, B.H., 2014. Wisdom of crowds: The value of stock opinions transmitted through social media. *Review of Financial Studies* 27: 1367-1403.

Cologni, A. and Manera, M., 2008. Oil prices, inflation and interest rates in a structural cointegrated VAR model for the G-7 countries. *Energy Economics* 30: 856-888.

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

Duffie, D., 2010. Presidential address: Asset price dynamics with slow-moving capital. *Journal of Finance* 65: 237-1267.

Elder, J., Miao, H. and Ramchander, S., 2013. Jumps in oil prices: the role of economic news. *The Energy Journal* 34: 217-237.

Engelberg, J.E. and Parsons, C.A., 2011. The causal impact of media in financial markets. *Journal of Finance* 66: 67-97.

Froot, K., Kang, N., Ozik, G. and Sadka, R., 2017. What do measures of real-time corporate sales say about earnings surprises and post-announcement returns? *Journal of Financial Economics* 125: 143-162.

Gissler, S., Oldfather, J. and Ruffino, D., 2016. Lending on hold: Regulatory uncertainty and bank lending standards. *Journal of Monetary Economics* 81: 89-101.

Gurun, U.G. and Butler, A.W., 2012. Don't believe the hype: Local media slant, local advertising, and firm value. *Journal of Finance* 67: 561-598.

Hamilton, J.D., 1983. Oil and the macroeconomy since World War II. *Journal of Political Economy* 91: 228-248.

Hamilton, J.D., 1996. This is what happened to the oil price-macroeconomy relationship. *Journal of Monetary Economics* 38: 215-220.

Han, L., Lv, Q. and Yin, L., 2017. Can Investor Attention Predict Oil Prices? *Energy Economics* 66: 547-558.

Kilian, L. and Vega, C., 2011. Do energy prices respond to US macroeconomic news? A test of the hypothesis of predetermined energy prices. *Review of Economics and Statistics* 93: 660-671.

Larcker, D.F. and Zakolyukina, A.A., 2012. Detecting deceptive discussions in conference calls. *Journal of Accounting Research* 50: pp.495-540.

Law, K.K. and Mills, L.F., 2015. Taxes and financial constraints: Evidence from linguistic cues. *Journal of Accounting Research* 53: 777-819.

Liu, B. and McConnell, J.J., 2013. The role of the media in corporate governance: Do the media influence managers' capital allocation decisions? *Journal of Financial Economics* 110: 1-17.

Loughran, T. and McDonald, B., 2011. When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *Journal of Finance* 66: 35-65.

Loughran, T. and McDonald, B., 2016. Textual analysis in accounting and finance: A survey. *Journal of Accounting Research* 54: 1187-1230.

Manela, A. and Moreira, A., 2017. News implied volatility and disaster concerns. *Journal of Financial Economics* 123: 137-162.

Nandha, M. and Faff, R., 2008. Does oil move equity prices? A global view. *Energy Economics* 30: 986-997.

Peress, J., 2014. The media and the diffusion of information in financial markets: Evidence from newspaper strikes. *Journal of Finance* 69: 2007-2043.

Segal, B. and Segal, D., 2016. Are managers strategic in reporting non-earnings news? Evidence on timing and news bundling. *Review of Accounting Studies* 21: 1203-1244.

Singleton, K.J., 2014. Investor flows and the 2008 boom/bust in oil prices. *Management Science* 60: 300-318.

Solomon, D.H., Soltes, E. and Sosyura, D., 2014. Winners in the spotlight: Media coverage of fund holdings as a driver of flows. *Journal of Financial Economics* 113: 53-72.

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

Tetlock, P.C., Saar-Tsechansky, M. and Macskassy, S., 2008. More than words: Quantifying language to measure firms' fundamentals. *Journal of Finance*, 63; 1437-1467.

## Appendix A

In this section, we present four examples of oil news articles. Words in bold indicate a decrease in oil prices while italicized words indicate an increase in oil prices. In the parsed news articles, only words in the updated Loughran-McDonald Master Dictionary are included in the analysis.

1. Neutral Example (April 17, 2011). Number of positive instances: 7 and number of negative instances: 6. *% Tone Index* is 0.34% = (7-6)/289 words.

KUWAIT (Dow Jones) -- Saudi Arabia's oil minister said Sunday that the global crude market is **oversupplied** and that the kingdom's **output** could **rise** this month compared with March. Saudi's crude oil output stood at 8.292 million barrels a day last month, down from 9.125 million barrels a day in February, Ali al-Naimi said in remarks carried by state-run Kuwait News Agency, or KUNA. The kingdom's **output** could **rise** this month compared with March, Naimi said, adding that Saudi Arabia has an output capacity of 12.5 million barrels a day. Saudi Arabia has *cut* oil *production* by 500,000 barrels a day, reversing a previous boost decided in response to the Libyan crisis, people familiar with the matter said last week, after meeting tepid demand for the extra output. In late February, Saudi Arabia--the largest producer in the Organization of Petroleum Exporting Countries--**increased production** to soothe market concerns as *turmoil* brought Libyan exports to a standstill. The reduction, confirmed by Saudi officials, brings the country's production back to about 8.5 million barrels a day. The Gulf nation had encountered only limited interest from buyers, due in part to high prices, maintenance at **refineries**, and **reduced** Japanese **demand** after the earthquake and tsunami. Naimi earlier this month said Saudi Arabi has enough spare crude oil capacity to meet any *increased* global *demand* or potential *supply shortage* in the market. Current high prices are caused primarily by speculation, misinformation and unjustified fear about the future of supply and demand, he added, in an effort to reassure the market in the wake of rising anxiety over the Libyan *outages*. Oil prices have risen sharply in recent weeks as political *upheaval* in North Africa has spread from Egypt and Tunisia into Libya, a major oil exporter. Saudi Arabia and other OPEC members have said they will make up for any *shortfall* in Libyan *output*, but markets remain edgy.

2. Neutral example (April 6, 2016). Number of positive instances: 6 and number of negative instances: 5. *% Tone Index* is 0.28% = (6-5)/357 words.

By Georgi Kantchev and Jenny W. Hsu NEW YORK -- Oil prices rallied Wednesday after an industry group reported an unexpected *decline* in U.S. crude *stockpiles*, and maintained their gains when the official U.S. data was released. Light, sweet crude for May delivery recently rose $1.17, or 3.1%, to $37.01 a barrel on the New York Mercantile Exchange. Brent, the global benchmark, rose 85 cents, or 2.2%, to $38.72 a barrel on ICE Futures Europe. U.S. crude-oil **supplies** stand at the **highest** level in more than 80 years, and market watchers think they are still growing due to continued **robust production** and **weaker demand** as **refiners** perform seasonal maintenance. The Energy Information Administration reported Wednesday that U.S. crude *supplies dropped* by 4.9 million barrels last week, far from analysts' estimates of a 3.3-million-barrels rise. The American Petroleum Institute, an industry group, said late Tuesday that its own data for the same week showed a 4.3-million-barrel *decline* in U.S. crude *inventories*. The agency will also publish its latest U.S. oil output estimate. Production has held above 9 million barrels a day in recent months, down from a peak of 9.7 million barrels a day last April. Investors are closely watching the 9-million-barrel mark, and a drop below that level could boost prices further, said Commerzbank in a note. In China, a private gauge of service activity showed a faster pace of expansion last month following moves by Beijing to prop up growth after a shaky start of the year. China is the world's second largest oil consumer. The Caixin China services purchasing managers index rose to 52.2 in March from 51.2 in February, Caixin Media Co. and research firm Markit said overnight. A reading above 50 indicates a month-to-month expansion, while a level below that points to a contraction. Oil prices have gained in recent months on speculation of a possible *production freeze* among major producing nations. However, the rally stalled last week after Saudi Arabia said Friday it would back out unless Iran was on board. Tehran plans to **increase output** until it reaches pre-sanction levels of around 4 million barrels a day. Kuwait, a heavyweight in the Organization of the Petroleum Exporting Countries, expressed confidence Tuesday that players will agree to *limit* their *output* when OPEC and non-OPEC producers, including Russia, meet in Doha, Qatar, on April 17. "Market watchers will be keeping their ears sharp until the suspense of a possible *production freeze* is over," said Barnabas Gan, an OCBC commodities analyst. Gasoline futures recently rose 0.2% to $1.3798 a gallon. Diesel futures rose 1.7% to $1.0927 a gallon.

3. Positive tone example (August 5, 2012). Number of positive instances: 7 and number of negative instances: 0. *% Tone Index* is 4.14% = (7-0)/169 words.

KUWAIT CITY (AFP)--A *drop* in Iranian *production* coupled with regional *tensions* were pushing oil prices higher, Kuwaiti Oil Minister Hani Hussein said in remarks published Sunday. “Iranian *production* has *dropped* which has contributed to raising prices,” Hussein was quoted as saying by Al-Watan newspaper. "Fears from regional *tensions*" and economic issues have also pushed prices higher, he added. Global oil prices rebounded Friday after better-than-expected jobs data in the United States and ongoing *tensions* over key producer Iran. New York's main contract, West Texas Intermediate light sweet crude for September, jumped $4.27 to $91.40 a barrel. Brent North Sea crude for delivery in September soared $3.04 to $108.94 a barrel in London deals. Hussein said that despite geopolitical *tensions*, “oil supplies are going well and there is enough production to meet market demand which is a positive signal to the market. Iranian oil *production* has *dropped* sharply following European and U.S. sanctions on the Islamic republic over its nuclear program”, according to the Organization of Petroleum Exporting Countries. U.S. President Barack Obama on Tuesday imposed new economic sanctions on Iran's oil export sector and on a pair of Chinese and Iraqi banks accused of doing business with Tehran.

4. Negative tone example (May 23, 2004). Number of positive instances: 0 and number of negative instances: 7. *% Tone Index* is -2.94% = (0-7)/238 words.

Dow Jones--President of the Organization of Petroleum Exporting Countries Purnomo Yusgiantoro said Sunday that the group is not opposed to Saudi Arabia's plan to **raise output** unilaterally. "Right now, we encourage the member countries to do as much as they can," he told reporters on the sidelines of the International Energy Forum in Amsterdam. "The Saudi proposal is there, and the Saudi position is fine by us," he added. Saudi Oil Minister Ali Naimi said in an interview with pan-Arab daily al-Hayat published Sunday OPEC should **raise** its **production** ceiling by 2.3 million-2.5 million barrels a day, this is more than his call Friday for a 2 million b/d ceiling hike. However, Libya's Oil Minister Fathi bin Shatwan said Sunday it was "wrong" for Saudi Arabia to unilaterally **boost** its crude **production** ahead of OPEC sanctioning a **rise** in **output quotas**. One senior OPEC source said, though there would undoubtedly be a ceiling rise agreed in Beirut, he was very surprised by the high amount Naimi was touting. The source also mirrored concerns expressed by the oil ministers of Libya and Iran that there remained the potential for a sharp price fall late in the year if global - and particularly U.S. - oil **inventories** continued to **climb**. Naimi said it was Saudi Arabia's unique spare capacity that will add real meaning to any OPEC ceiling increase at the group's June 3 meeting in Beirut, given that the producer group is already **overproducing** by more than 2 million barrels a day. "There's still spare capacity left," Yusgiantoro said, adding that OPEC is producing at around 88% of capacity.

## Appendix B

List of positive and negative modifiers.

The table presents in Panel A the positive modifiers that imply an increase in oil prices while Panel B reports the negative modifiers that imply a decrease in oil prices.

*Panel A: Positive modifiers*

Abound
Abounded
Abounding
Abounds
Abundance
Abundances
Abundant
Abundantly
Accelerate
Accelerated
Accelerates
Accelerating
Acceleration
Accumulate
Accumulated
Accumulates
Accumulating
Accumulation
Accumulative
Add
Added
Additional
Adds
Ample
Amply
Arise
Arisen
Arises
Arising
Awash
Benefit
Benefited
Benefiting
Benefits
Better
Bigger
Binge
Bolster
Bolstered
Bolstering
Boom
Booming
Boost
Boosted
Boosting
Bounce
Bounced
Brim
Brimful
Brimmed

Brimming
Brims
Brisk
Brisked
Brisking
Briskly
Build
Building
Bull
Bullish
Bulls
Buoy
Buoyant
Buoyed
Buoying
Cheating
Climb
Climbed
Climbing
Deluge
Deluged
Deluges
Deluging
Elevate
Elevated
Elevating
Enhance
Enhanced
Enhancing
Escalate
Escalated
Escalates
Escalating
Exceed
Exceeded
Exceeding
Exceeds
Excellent
Excess
Excessive
Excessively
Expand
Expanded
Expanding
Expands
Expansion
Expansionary
Expansions
Expansive
Extend

Extended
Extending
Extends
Extra
Extravagantly
Forward
Foster
Fostered
Fostering
Fosters
Fuelled
Fuelling
Fuels
Gain
Gained
Gaining
Gains
Grew
Grow
Growing
Grown
Growths
Headway
Healthier
Hefty
Heightens
High
Higher
Highest
Highly
Highs
Hike
Hiked
Hikes
Huge
Improve
Improved
Improvement
Improvements
Improves
Improving
Increase
Increased
Increases
Increasing
Increasingly
Inflate
Inflated
Inflates
Inflating

Jump
Jumped
Jumping
Jumps
Large
Larger
Lavishly
Lift
Lifted
Lifted
Lifts
Massive
Maximize
Maximized
Maximizes
Maximizing
More
Mount
Optimistic
Outpace
Outpaced
Outpaces
Outpacing
Outperform
Outperformed
Outperforming
Outperforms
Outrun
Outstrip
Overly
Overshoot
Overwhelmingly
Peak
Peaked
Peaking
Persist
Persisted
Persistent
Plentiful
Positive
Positively
Progress
Progressed
Progressing
Progression
Raise
Raised
Raising
Rallying
Ramp

Ramped
Ramping
Ramps
Rebound
Record
Recoup
Recover
Recovered
Recovering
Recovery
Replenish
Replenished
Replenishes
Replenishing
Resilient
Resilient
Resiliently
Revive
Revived
Revives
Rise
Rised
Risen
Rising
Robust
Rocketed
Rose
Rush
Soar
Soared
Soaring
Soaring
Solid
Speculative
Spur
Spurred
Spurring
Spurted
Spurting
Spurts
Steam
Stimulate
Stimulated
Stimulates
Stimulating
Strength
Strengthen
Strengthened
Strengthening
Stretch

Stretched
Stretches
Stretching
Strong
Strongly
Success
Successful
Successfully
Surge
Surged
Surges
Surging
Surpass
Surpassed
Surpasses
Surpassing
Sustain
Top
Tops
Unrestrained
Up
Upgrade
Upgraded
Upgrades
Upgrading
Upped
Upper
Ups
Upside
Uptick
Upturn
Upturned
Upturning
Upturns
Upward
Upward
Upwardly
Warm
Warmer
Warming
Widening

*Panel B: Negative modifiers*

Abandon
Abandoned
Abandoning
Abandons
Abate
Abated
Abatement
Abatements
Abates
Abating
Abbreviate
Abbreviated
Abbreviates
Abbreviating
Absorb
Absorbed
Absorbing
Absorbs
Adverse
Adversely
Ail
Ailed
Ailing
Ails
Alleviate
Alleviates
Alleviating
Alleviation
Alleviations
Anemic
Bad
Badly
Battered
Bear
Bearish
Bottom
Bottomed
Breach
Breached
Breaches
Breaching
Burden
Burdened
Burdening
Burdens
Cap
Capped
Cease
Ceased
Ceases
Ceasing
Cold
Colder
Coldly
Collapse
Collapsed
Collapses
Collapsing
Condemn

Contracting
Contraction
Cool
Cooled
Cooler
Coolest
Cools
Crash
Crashed
Crashes
Crashing
Crimp
Crimped
Crimping
Crimps
Cripple
Crippled
Cripples
Crippling
Crop
Crumble
Crumbled
Crumbles
Crumbling
Crunch
Crunched
Crunches
Crush
Crushed
Crushes
Crushing
Curb
Curbed
Curbing
Curbs
Curtail
Curtail
Curtailed
Curtailing
Curtailment
Curtailments
Curtails
Cut
Cuts
Cutting
Damage
Damaged
Damages
Damaging
Damp
Damped
Dampen
Dampening
Darken
Darkened
Decay
Decayed
Decaying
Decays

Depressing
Depressingly
Destabilization
Destabilize
Destabilized
Destabilizes
Destabilizing
Destroy
Destroyed
Destroying
Destroys
Destruction
Destructions
Deteriorate
Deteriorated
Deteriorates
Deteriorating
Deterioration
Deteriorations
Devastate
Devastated
Devastates
Devastating
Devastatingly
Devastation
Devastations
Alleviated
Cancelled
Coldest
Diminish
Diminished
Diminishing
Dimmed
Dipped
Dipped
Dipping
Dips
Disappoint
Disappointed
Disappointing
Discouraging
Disrupt
Disrupted
Disrupting
Disruption
Disruptions
Disrupts
Distress
Distressed
Dogged
Down
Downbeat
Downbeats
Downgrade
Downgraded
Downgrades
Downgrading
Downs
Downscale

Drop
Dropped
Dropping
Drops
Dwindle
Dwindled
Dwindles
Dwindling
Dwindling
Ended
Ending
Erode
Eroded
Erodes
Eroding
Evaporate
Evaporated
Evaporates
Evaporating
Evaporation
Evaporations
Exacerbate
Exacerbated
Exacerbates
Exacerbating
Fade
Faded
Fades
Fading
Fall
Fallen
Falling
Falls
Falter
Faltered
Faltering
Falters
Fell
Fewer
Fewest
Fizzle
Fizzled
Fizzles
Fizzling
Flip
Freeze
Freezes
Freezing
Frigid
Futile
Gap
Gloomy
Halt
Halted
Halting
Halts
Hammer
Hammered
Hammering

Hollow
Hollowed
Hollowing
Icily
Icy
Impede
Impeded
Impedes
Impeding
Inadequate
Insufficient
Intermittent
Interrupt
Interrupted
Interrupting
Interrupts
Intimidate
Intimidated
Intimidates
Intimidating
Intimidatingly
Lack
Lackluster
Lag
Lagged
Lagging
Less
Lessened
Limit
Limited
Limiting
Limits
Lingering
Lost
Low
Lower
Lowered
Lowest
Lows
Melt
Minimize
Minimized
Minimizing
Muted
Negative
Negatively
Pause
Paused
Pausing
Perilously
Pessimistic
Pinch
Plummet
Plummeted
Plummeting
Plummets
Plunge
Plunged
Plunges

Reduces
Reducing
Reduction
Rein
Reining
Restrain
Restrained
Restraining
Restraint
Restraints
Restrict
Restricted
Restricting
Restricts
Retreat
Retreated
Retreating
Retreats
Rupture
Ruptured
Ruptures
Rupturing
Sagging
Sap
Saturated
Scant
Severe
Severely
Severing
Shabby
Shortfall
Shrank
Shrink
Shrinked
Shrinking
Shrinks
Shut
Shutdown
Shutdowns
Slack
Slap
Slash
Slashed
Slashes
Slashing
Slid
Slide
Slip
Slipped
Slipping
Slips
Slow
Slowdown
Slowed
Slower
Slowing
Sluggish
Slump
Slumped

Stopped
Struck
Stuck
Stunt
Stunted
Stunting
Stunts
Stymie
Stymied
Stymieing
Stymies
Subdue
Subdued
Subdues
Subside
Subsiding
Suffer
Suffered
Suffering
Suffers
Sunk
Suppress
Suppressed
Suppresses
Suppressing
Suspend
Suspended
Suspending
Suspends
Suspension
Taper
Tapered
Tapering
Thwart
Thwarted
Thwarting
Thwarts
Tight
Tighten
Tightening
Tighter
Trim
Trimmed
Trimming
Trims
Truncate
Truncated
Truncates
Truncating
Tumble
Tumbled
Undercut
Undermine
Undermined
Undermines
Underperform
Underperformed
Underperforming
Underperforms

Condemned
Condemning
Condemns
Conservative
Constrain
Constrained
Constraining
Constrains
Constraint
Constraints
Constrict
Constricted
Constricting
Constriction
Constrictions
Constrictive
Contract
Contracted
Decline
Declined
Declines
Declining
Decrease
Decreased
Decreases
Decreasing
Decreasingly
Deflate
Deflated
Deflates
Deflating
Deflect
Dented
Depress
Depressed
Depresses
Downscaled
Downscales
Downscaling
Downsize
Downturn
Downturns
Downward
Downwards
Drag
Dragged
Dragging
Drain
Drained
Draining
Drastically
Drawdown
Drawdowns
Drawn
Hammers
Hamper
Hampered
Hampering
Hampers
Hardship
Harm
Harmed
Harming
Harsh
Hinder
Hindered
Hindering
Hinders
Hobble
Hobbled
Hobbles
Hobbling
Plunging
Poor
Poorly
Pressed
Pressing
Pressure
Pressured
Pressures
Quash
Quashed
Quashes
Quashing
Recede
Receded
Recedes
Receding
Reduce
Reduced
Slumping
Slumps
Small
Smaller
Smolder
Snag
Soften
Softening
Spook
Spooked
Stagger
Staggered
Staggering
Staggers
Stall
Stifled
Stifles
Stifling
Stilted
Stop Wane
Waned
Wanes
Weak
Weaken
Weakened
Weakening
Whack
Withdraw
Worse
Worsen
Worsening
Worst

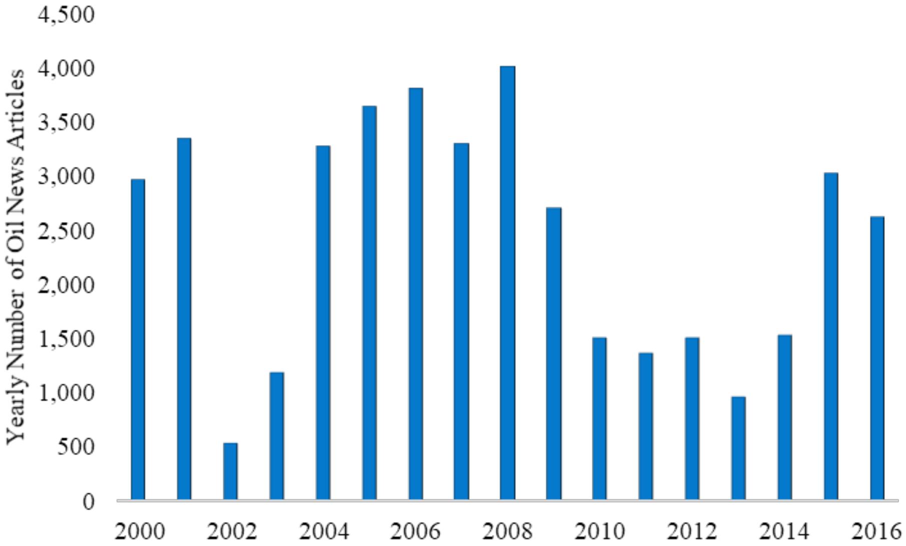

Figure 1. The figure depicts the number of annual Dow Jones Energy Service (DJES) news articles published during January, 2000 to September, 2016. The sample includes oil news articles published during 12:01 AM to 2:15 PM each day. Overall, 41,432 DJES news articles were retrieved and analysed.

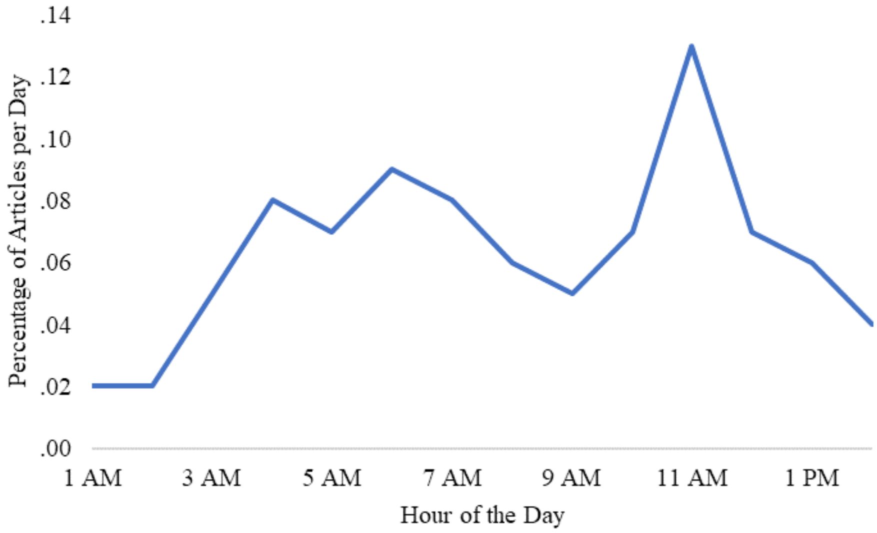

Figure 2. Number of DJES oil articles by the hour-of-the-day.

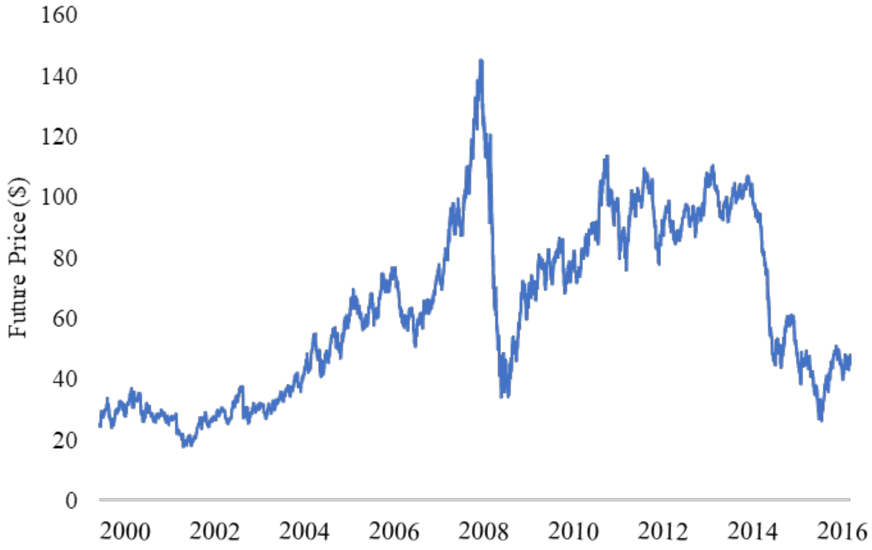

Figure 3. The figure reports the NYMEX Cushing, OK Crude Oil Future Contract 1 prices during the January 2000 to September 2016 time-period.

**Table 1**
**List of 130 keywords created from the Dow Jones Energy Service news database that are expected to affect oil prices**

| 59 Keywords that should be expected to increase oil Prices (1) | | 19 Keywords that should be expected to decrease oil prices (2) | 52 Keywords that need to be signed by a modifier (3) | |
|---|---|---|---|---|
| Attack | Outage | Discovered | Allotment | Producing |
| Attacker | Outages | Discoveries | Allotments | Production |
| Attackers | Problem | Discovery | Buying | Pumping |
| Attacks | Problematic | Glut | Capacities | Pumpings |
| Bomb | Problematicly | Gluts | Cargo | Quota |
| Bombed | Problems | Mistrust | Cargoe | Quotas |
| Bomber | Recoveries | Overproducer | Cargoes | Refineries |
| Bombers | Recovery | Overproducers | Constructing | Refinery |
| Bombing | Sabotage | Overproducing | Construction | Reserves |
| Bombings | Sabotages | Overproduction | Consumption | Rig |
| Bombs | Shortage | Oversupplied | Demand | Rigs |
| Closures | Shortages | Oversupply | Demands | Stock |
| Concerned | Storm | Oversupplying | Difficulties | Stockpile |
| Conflict | Storms | Recession | Drilling | Stockpiles |
| Conflicts | Strike | Recessions | Drillings | Stocks |
| Cutback | Strikers | Repaired | Economic | Supplies |
| Cutbacks | Strikes | Repairing | Economies | Supply |
| Delay | Tension | Surplus | Economy | Temperature |
| Delayed | Tensions | Surpluses | Embargo | Temperatures |
| Delaying | Turmoil | | Exploration | Weather |
| Delays | Turmoils | | Export | |
| Dispute | Unrest | | Exports | |
| Disputed | Upheaval | | Import | |
| Disputes | Upheavals | | Imports | |
| Disruptions | Withdraw | | Inventories | |
| Explosion | Withdrawing | | Inventory | |
| Explosions | Withdrawn | | Output | |
| Fire | | | Outputs | |
| Fires | | | Pipeline | |
| Hurricane | | | Pipelines | |
| Hurricanes | | | Platform | |
| Instability | | | Platforms | |

**Table 2**
**Most frequent keywords and word combinations in Dow Jones Energy Service oil news articles**

*Panel A: List of Most Frequent Keywords*

| | % of total key word count | Cumulative % | | % of total key word count | Cumulative % |
|---|---|---|---|---|---|
| Recovery | 9.31% | 9.31% | Problem | 2.75% | 65.98% |
| Problems | 5.45% | 14.76% | Oversupplied | 2.58% | 68.56% |
| Attacks | 5.09% | 19.85% | Delayed | 2.33% | 70.89% |
| Oversupply | 4.92% | 24.77% | Fire | 2.27% | 73.16% |
| Hurricane | 4.89% | 29.66% | Outages | 2.08% | 75.24% |
| Glut | 4.07% | 33.72% | Delays | 2.02% | 77.26% |
| Concerned | 3.74% | 37.46% | Delay | 1.96% | 79.23% |
| Disruptions | 3.60% | 41.07% | Conflict | 1.48% | 80.71% |
| Shortage | 3.48% | 44.54% | Shortages | 1.45% | 82.17% |
| Recession | 3.38% | 47.92% | Overproduction | 1.29% | 83.45% |
| Strike | 3.25% | 51.17% | Unrest | 1.17% | 84.63% |
| Tensions | 3.22% | 54.39% | Strikes | 1.14% | 85.76% |
| Storm | 3.04% | 57.43% | Dispute | 1.04% | 86.80% |
| Surplus | 2.92% | 60.35% | Discovery | 1.01% | 87.82% |
| Attack | 2.89% | 63.24% | Explosion | 0.95% | 88.77% |

*Panel B: Most Frequent Keyword-modifier Combinations*

| | | | | | |
|---|---|---|---|---|---|
| Output Cut | 2.68% | 2.68% | Output Boost | 0.47% | 14.15% |
| Production Cut | 1.51% | 4.19% | Output Cuts | 0.47% | 14.61% |
| Demand Strong | 0.95% | 5.14% | Demand Low | 0.45% | 15.06% |
| Production Increase | 0.92% | 6.06% | Output Rise | 0.44% | 15.50% |
| Output Increase | 0.91% | 6.97% | Supply Tight | 0.42% | 15.92% |
| Stocks Build | 0.85% | 7.83% | Demand Higher | 0.41% | 16.33% |
| Demand Weak | 0.84% | 8.66% | Inventories Build | 0.38% | 16.72% |
| Output Raise | 0.78% | 9.44% | Output Up | 0.38% | 17.09% |
| Demand Up | 0.70% | 10.14% | Demand More | 0.36% | 17.45% |
| Demand High | 0.67% | 10.81% | Stocks Low | 0.36% | 17.80% |
| Output Hike | 0.63% | 11.44% | Economic Recovery | 0.34% | 18.15% |
| Weather Cold | 0.60% | 12.05% | Demand Rising | 0.34% | 18.49% |
| Demand Lower | 0.60% | 12.65% | Demand Increase | 0.33% | 18.82% |
| Production Cuts | 0.57% | 13.21% | Refinery Struck | 0.33% | 19.15% |
| Supply Disruptions | 0.47% | 13.68% | Demand Poor | 0.33% | 19.47% |

This table presents the fractional and cumulative percentages of the 30 most frequent keywords and the 30 most frequent keyword-modifiers.

**Table 3**
**Summary statistics and correlations**

| *Panel A: Summary Statistics* | Mean | Std.dev. | 5th percentile | Median | 95th percentile |
|---|---|---|---|---|---|
| % Tone Index | 0.20% | 0.60% | -0.72% | 0.18% | 1.21% |
| Oil Price | $63.62 | $27.90 | $26.45 | $61.38 | $105.73 |
| % Negative | 1.59% | 0.52% | 0.75% | 1.58% | 2.47% |
| Number of articles | 9.81 | 6.83 | 2.00 | 9.00 | 22.00 |
| Dollar exchange rate | 84.55 | 11.33 | 71.01 | 81.94 | 108.09 |
| Gold price | $876.08 | $469.49 | $272.80 | $884.00 | $1,663.95 |
| 10-year Treasury rate | 3.59% | 1.25% | 1.72% | 3.74% | 5.68% |
| VIX index | 20.19 | 8.57 | 11.59 | 18.10 | 35.80 |

*Panel B: Correlation table of the transformed variables*

| | Oil returns | % Tone Index | % Negative | # of articles | $ exchange rate | Gold price | 10-Year rate |
|---|---|---|---|---|---|---|---|
| % Tone Index | 0.07 | | | | | | |
| % Negative | -0.06 | 0.16 | | | | | |
| # of articles | -0.05 | -0.08 | -0.01 | | | | |
| $ exchange rate | -0.22 | -0.02 | 0.02 | 0.02 | | | |
| Gold price | 0.05 | 0.02 | -0.03 | -0.01 | -0.21 | | |
| 10-Year rate | 0.19 | 0.06 | -0.02 | -0.02 | 0.06 | -0.03 | |
| VIX index | -0.21 | -0.02 | 0.00 | 0.01 | 0.10 | 0.02 | -0.31 |

This table presents in Panel A summary statistics for the variables of interest and for the control variables used in regressions. All variables in Panel A are presented in levels. *% Tone Index* is defined as (number of oil price increasing phrases - number of oil price decreasing phrases) / (number of words in the article). The *% Tone Index* is the daily average of all DJES articles in a given day. Oil price is the NYMEX Cushing, OK Crude Oil Future Contract 1 price per barrel. Control variables include *% Negative* (from the Loughran and McDonald (2011) word list), number of daily oil articles, dollar exchange rate, gold price, 10-year Treasury rate, and VIX index. Panel B presents correlations between the transformed key variables.

**Table 4**
**OLS regressions between oil price returns and the *% Tone Index* of oil news articles**

| | Dependent variable: oil price returns | | | | |
|---|---|---|---|---|---|
| | [1] | [2] | [3] | [4] | [5] |
| $\% \, Tone\ Index_t$ | | | | 0.33<br>(4.30) | 0.37<br>(4.57) |
| $\% \, Tone\ Index_{t-1}$ | -0.16<br>(-2.39) | -0.17<br>(-2.65) | -0.17<br>(-2.44) | | -0.21<br>(-2.95) |
| $\% \, Tone\ Index_{t-2}$ | | | 0.00<br>(0.05) | | -0.01<br>(-0.22) |
| $\% \, Tone\ Index_{t-3}$ | | | -0.06<br>(-0.89) | | -0.09<br>(-1.28) |
| $\% \, Tone\ Index_{\,t-4}$ | | | 0.08<br>(1.15) | | 0.05<br>(0.66) |
| Intercept | 0.02<br>(0.88) | 0.02<br>(0.87) | 0.02<br>(0.88) | 0.02<br>(0.96) | 0.02<br>(0.91) |
| Calendar Dummies | Yes | Yes | Yes | Yes | Yes |
| Number of Oil Articles | Yes | Yes | Yes | Yes | Yes |
| Lagged Oil Returns | Yes | No | Yes | Yes | Yes |
| Adj. $R^2$ | 0.34% | 0.21% | 0.31% | 0.76% | 0.93% |
| Observations | 3,568 | 3,568 | 3,568 | 3,568 | 3,568 |

This table presents regression results with the dependent variable defined as oil price returns. The key independent variable is the level of the *% Tone Index*. The *% Tone Index* is defined as (number of oil price increasing phrases - number of oil price decreasing phrases)/(number of words per article) for all DJES oil news articles in a given day published during the period 12:01 AM to 2:15 PM. The sample period is January 2000 – September 2016. Calendar year dummies, lagged values of the dependent variable up to 4 lags, and the number of daily oil news articles on day t-1 are included in all regressions. The standard errors include a Newey-West correction for heteroscedasticity and autocorrelation up to five lags. Robust *t*-statistics are reported in parentheses.

**Table 5**
**OLS regressions between oil price returns and the *% Tone Index* of oil news articles**

| | Dependent variable: oil price returns | | | |
|---|---|---|---|---|
| | [1] | [2] | [3] | [4] |
| $\% \, Tone \; Index_t$ | | | 0.39<br>(5.00) | 0.41<br>(5.21) |
| $\% \, Tone \; Index_{t-1}$ | | -0.12<br>(-1.80) | | -0.17<br>(-2.47) |
| $\% \, Negative_t$ | | | -0.34<br>(-4.67) | -0.32<br>(-4.42) |
| $\% \, Negative_{t-1}$ | -0.28<br>(-3.78) | -0.25<br>(-3.34) | | -0.20<br>(-2.91) |
| $\Delta(Exchange \; rate_{t-1})$ | 0.00<br>(0.24) | 0.00<br>(0.24) | 0.00<br>(0.06) | 0.00<br>(0.07) |
| $\Delta(Gold_{t-1})$ | 0.00<br>(1.82) | 0.00<br>(1.83) | 0.00<br>(1.86) | 0.00<br>(1.81) |
| $\Delta((10yr \; rate)_{t-1})$ | -0.08<br>(-0.99) | -0.00<br>(-0.93) | -0.05<br>(-0.97) | -0.04<br>(-0.92) |
| $\Delta(VIX_{t-1})$ | -0.00<br>(-1.88) | -0.00<br>(-1.79) | -0.00<br>(-1.61) | -0.00<br>(-1.65) |
| Intercept | 0.06<br>(2.50) | 0.05<br>(2.31) | 0.07<br>(2.95) | 0.01<br>(3.90) |
| Calendar Dummies | Yes | Yes | Yes | Yes |
| Number of Oil Articles | Yes | Yes | Yes | Yes |
| Lagged Oil Returns | Yes | Yes | Yes | Yes |
| Adj. $R^2$ | 0.61% | 0.66% | 1.29% | 1.62% |
| Observations | 3,568 | 3,568 | 3,568 | 3,568 |

This table presents regression results with the dependent variable defined as oil price returns and the independent variables *% Tone Index*, *% Negative*, a trade-weighted U.S dollar index, spot price of gold, 10-year Treasury constant maturity rate, and VIX index. The *% Tone Index* is defined as (number of oil price increasing phrases - number of oil price decreasing phrases)/(number of words in the article) for all DJES oil news articles in a given day. The sample period is January 2000 – September 2016. Calendar year dummies, lagged values of the dependent variable up to 4 lags, and the number of daily oil news articles on day t-1 are included in all regressions. The standard errors also include a Newey-West correction for heteroscedasticity and autocorrelation up to five lags. Robust *t*-statistics are reported in parentheses.

**Table 6**
**GARCH (1,1) regressions between *%Tone Index* and oil price returns**

| | Dependent variable: oil price returns | | | |
|---|---|---|---|---|
| | [1] | [2] | [3] | [4] |
| $\% \, Tone \; Index_t$ | | | 0.22<br>(3.65) | 0.24<br>(4.04) |
| $\% \, Tone \; Index_{t-1}$ | -0.17<br>(-3.17) | -0.17<br>(-3.16) | | -0.19<br>(-3.54) |
| $\% \, Tone \; Index_{t-2}$ | | 0.00<br>(0.05) | | 0.00<br>(0.12) |
| $\% \, Tone \; Index_{t-3}$ | | -0.03<br>(-0.58) | | -0.05<br>(-0.83) |
| $\% \, Tone \; Index_{t-4}$ | | 0.04<br>(0.67) | | 0.01<br>(0.21) |
| Intercept | 0.01<br>(0.33) | 0.02<br>(0.93) | 0.01<br>(1.06) | 0.01<br>(0.39) |
| Calendar Dummies | Yes | Yes | Yes | Yes |
| Number of Oil Articles | Yes | Yes | Yes | Yes |
| Lagged Oil Returns | Yes | Yes | Yes | Yes |
| Adj. $R^2$ | 0.10% | 0.05% | 0.44% | 0.59% |
| Observations | 3,568 | 3,568 | 3,568 | 3,568 |

This table presents GARCH (1,1) regression results with the dependent variable defined as oil price returns. The independent variable is the *% Tone Index*, defined as (number of oil price increasing phrases - number of oil price decreasing phrases)/(number of words per article). The *% Tone Index* is the daily average of all DJES articles in a given day. The sample period is January 2000 – September 2016. Calendar year dummies, lagged values of the dependent variable up to 4 lags, and the number of daily oil news articles on day t-1 are included in all regressions. The *z*-statistics are reported in parentheses.

**Table 7**
**GARCH (1,1) regressions between *% Tone Index* and oil price returns**

| | Dependent variable: oil price returns | | | |
|---|---|---|---|---|
| | [1] | [2] | [3] | [4] |
| $Tone\ Index_t$ | | | 0.25<br>(4.21) | 0.26<br>(4.42) |
| $Tone\ Index_{t-1}$ | | -0.14<br>(-2.50) | | -0.17<br>(-3.02) |
| $\%\ Negative_t$ | | | -0.21<br>(-3.38) | -0.19<br>(-3.14) |
| $\%\ Negative_{t-1}$ | -0.18<br>(-2.93) | -0.15<br>(-2.35) | | -0.11<br>(-1.88) |
| $\Delta(Exchange\ rate_{t-1})$ | 0.00<br>(0.13) | 0.00<br>(0.08) | 0.00<br>(0.38) | 0.00<br>(0.43) |
| $\Delta(Gold_{t-1})$ | 0.00<br>(1.55) | 0.00<br>(1.54) | 0.00<br>(1.34) | 0.00<br>(1.48) |
| $\Delta((10yr\ rate)_{t-1})$ | -0.01<br>(-1.60) | -0.01<br>(-1.54) | -0.01<br>(-1.47) | -0.01<br>(-1.46) |
| $\Delta(VIX_{t-1})$ | -0.00<br>(-1.67) | -0.00<br>(-1.67) | -0.01<br>(-1.53) | -0.01<br>(-1.53) |
| Intercept | 0.03<br>(1.19) | 0.02<br>(1.05) | 0.04<br>(3.59) | 0.05<br>(2.03) |
| Calendar Dummies | Yes | Yes | Yes | Yes |
| Number of Oil Articles | Yes | Yes | Yes | Yes |
| Lagged Oil Returns | Yes | Yes | Yes | Yes |
| Adj. $R^2$ | 0.29% | 0.34% | 0.86% | 1.16% |
| Observations | 3,568 | 3,568 | 3,568 | 3,568 |

This table presents GARCH (1,1) regression results with the dependent variable defined as oil price returns. The independent variables are *% Tone Index*, % Negative words of Loughran-McDonald (LM) word list, a trade-weighted U.S dollar index, spot price of gold, 10-year Treasury constant maturity rate, and VIX index. The *% Tone Index* is defined as (number of oil price increasing phrases - number of oil price decreasing phrases)/(number of words in the article). The *% Tone Index* is the daily average of all DJES articles in a given day. The sample period is January 2000 – September 2016. Calendar year dummies, lagged values of the dependent variable up to 4 lags, and the number of daily oil news articles on day t-1 are included in all regressions. The *z*-statistics are reported in parentheses.

**Table 8**
**Daily returns generated by an implementable trading strategy using lagged *% Tone Index***

| | Long position: Low *% Tone Index* | | Short position: High *% Tone Index* | |
|---|---|---|---|---|
| Calendar Year | Number of transactions | Average return per transaction | Number of transactions | Average return per transaction |
| 2001 | 3 | -0.80% | 83 | 0.08% |
| 2002 | 40 | -0.04% | 57 | 0.08% |
| 2003 | 26 | -0.18% | 31 | 0.06% |
| 2004 | 48 | 0.00% | 12 | -0.29% |
| 2005 | 58 | -0.14% | 58 | -0.04% |
| 2006 | 32 | 0.56% | 39 | 0.80% |
| 2007 | 54 | 0.32% | 31 | 0.44% |
| 2008 | 83 | -0.42% | 15 | 1.20% |
| 2009 | 49 | 0.38% | 69 | 0.49% |
| 2010 | 20 | 0.35% | 67 | 0.23% |
| 2011 | 71 | 0.03% | 32 | 0.02% |
| 2012 | 44 | -0.20% | 36 | -0.01% |
| 2013 | 44 | 0.04% | 74 | -0.03% |
| 2014 | 36 | 0.12% | 37 | 0.59% |
| 2015 | 77 | 0.17% | 15 | 1.47% |
| 2016 | 20 | 0.41% | 73 | 0.08% |
| Total | 705 | 0.02% | 729 | 0.21% |
| | | (0.44) | | (2.18) |

This table presents the number of transactions and the returns of an investment strategy based on the top and bottom 20th percentiles of the lagged *% Tone Index*. The investment strategy includes both long and short positions depending on whether the lagged *% Tone Index* lies below or above the 20th percentile. The prior year's distribution of the *% Tone Index* determines the daily buy/sell/no trade decision for the investor during following year. If the lagged *% Tone Index* lies between the top and bottom 20%, the investor does no trading on that particular day. The trading strategy spans January, 2001 to September, 2016. We assume that each round-trip transaction incurs a one basis point trading cost. *T*-statistics are in parentheses.

**Table 9**
**OLS and GARCH regressions between U.S. Oil Fund returns and the *% Tone Index***

| | Dependent variable: U.S. Oil Fund returns | | | |
|---|---|---|---|---|
| | [1] | [2] | [3] | [4] |
| $\% \, Tone \, Index_t$ | | 0.24<br>(3.05) | | 0.15<br>(2.93) |
| $\% \, Tone \, Index_{t-1}$ | -0.21<br>(-3.27) | -0.21<br>(-3.09) | -0.17<br>(-2.92) | -0.18<br>(-3.04) |
| $\% \, Negative_t$ | | -0.18<br>(-2.57) | | -0.15<br>(-2.29) |
| $\% \, Negative_{t-1}$ | | -0.17<br>(-2.23) | | -0.05<br>(-0.90) |
| Intercept | 0.01<br>(0.09) | 0.04<br>(1.85) | -0.01<br>(-1.50) | 0.01<br>(0.07) |
| Calendar Dummies | Yes | Yes | Yes | Yes |
| Number of Oil Articles | Yes | Yes | Yes | Yes |
| Lagged ETF Returns | Yes | Yes | Yes | Yes |
| Adj. $R^2$ | 0.52% | 1.14% | 0.00% | 0.44% |
| Observations | 2,288 | 2,288 | 2,288 | 2,288 |

This table presents regression results with the dependent variable defined as U.S. Oil Fund ETF returns. The independent variable is the level of the *% Tone Index*. The *% Tone Index* is defined as (number of oil price increasing phrases - number of oil price decreasing phrases)/(number of words per article) for all DJES oil news articles in a given day published during the period 12:01 AM to 2:15 PM. The sample period is April 2006 – September 2016. Calendar year dummies, lagged values of the dependent variable up to 4 lags, and the number of daily oil news articles on day t-1 are included in all regressions. The standard errors include a Newey-West correction for heteroscedasticity and autocorrelation up to five lags. Robust *t*-statistics or *z*-statistics are reported in parentheses. OLS regressions are reported in columns 1 and 2, while GARCH regressions are reported in columns 3 and 4.