Mining for Oil Forecasts

Charles W. Calomiris, Nida Çakır Melek, and Harry Mamaysky

December 2020

Abstract

We study the usefulness of a large number of traditional variables and novel text-based measures for in-sample and out-of-sample forecasting of oil spot and futures returns, energy company stock returns, oil volatility, oil production, and oil inventories. After carefully controlling for small-sample biases, we find compelling evidence of in-sample predictability. Our text measures hold their own against traditional variables for oil forecasting. However, none of this translates to out-of-sample predictability until we data mine our set of predictive variables. Our study highlights that it is difficult to forecast oil market outcomes robustly.

Keywords: Asset Pricing, Commodity Markets, Energy Forecasting, Model Validation

JEL: C52, G10, G14, G17, Q47

1Office of the Comptroller of the Currency, charles.calomiris@occ.treas.gov, Federal Reserve Bank of Kansas City, nida.cakirmelek@kc.frb.org, and Columbia Business School, hm2646@columbia.edu, respectively. The authors are listed in alphabetical order. We thank Christiane Baumeister, Kateryna Holland (discussant), Huixin Bi, Deepa Datta, Jim Hamilton, Lee Smith, Jim Stock, and participants at the EIA 2019 Annual Workshop on Financial and Physical Energy Market Linkages, the FRB-IMF Workshop on New Techniques and Data in Macro Finance, the 2020 CEBRA Workshop for Commodities and Macroeconomics, and the Federal Reserve Bank of Kansas City for helpful comments and suggestions. Daliah Al-Shakhshir and Hongyu Wu provided excellent research assistance for this study. Roya Arab Loodaricheh helped with the textual analysis and Colton Tousey helped with construction of the energy words list. The views expressed herein are solely those of the authors and do not necessarily reflect the views of the Office of the Comptroller of the Currency, the U.S. Government, the Federal Reserve Bank of Kansas City, or the Federal Reserve System.

Electronic copy available at: https://ssrn.com/abstract=3755487
1. Introduction

The oil market receives more attention among macroeconomists and financial economists, and among general news outlets, than any other commodity market. This reflects its unparalleled importance as a major input to production and consumption goods, as well as its regional and geopolitical significance for U.S. states or foreign countries that produce and consume petroleum products in large quantity. Furthermore, the price of oil is a key variable in generating macroeconomic projections and in assessing macroeconomic risks.

In this study, we consider a broad range of predictors to forecast several important variables in the oil market, including oil futures returns, oil spot returns, the realized volatility of oil prices, the equity returns of oil companies, oil inventories and oil production for the period 1998-2020. Our goal is to construct a fully transparent empirical methodology for considering a comprehensive list of potential forecasting variables and investigating their usefulness both in sample and out of sample. We consider a wide range of potential explanatory variables – many of which have been included in prior studies – including macroeconomic and financial indicators and various additional measures that capture time-varying oil returns risk.2

In addition, we consider a set of new natural language processing (NLP) measures derived from the analysis of a corpus of oil news articles from Thomson Reuters (TR). Recent work has shown the usefulness of text measures for forecasting the returns and risks of individual stock and stock indexes, and we find that these techniques have value for oil forecasting. While some commonly used predictors of oil price changes, such as global oil production or global economic activity, are monthly and become available only with

2 A range of aggregate and commodity-market specific financial and macroeconomic variables used to predict commodity market outcomes are examined in Baumeister and Kilian 2017. Hamilton and Wu 2014 document significant changes in risk premia in crude oil futures contracts since 2005.
considerable delays, text measures can capture a wide range of energy market developments in real-time. Given the volume of news coverage of the energy sector, application of NLP tools in this space seems particularly promising. Indeed, we show that our textual measures can algorithmically identify important historical episodes in energy markets, in a way that traditional energy variables are unable to do. Our NLP measures include topic-specific frequency and sentiment derived from energy news, as well as a measure of the unusualness or “entropy” of oil news (i.e., the frequency of occurrence of unusual strings of words). Topics are obtained from a corpus of TR articles using an algorithm that identifies co-occurring lists of words. We employ a network modularity approach for identifying topics, as in Calomiris and Mamaysky (2019a).

Several features distinguish our approach from the literature: we begin with a comprehensive list of forecasting variables; our methodology for selecting variables is explicit; we adjust standard errors for selection bias with respect to determining which forecasting variables to include; we bootstrap our R-squareds to adjust for overlapping observations; and we consider a wide range of approaches to out-of-sample validation of our models. For these reasons, we believe our approach avoids reporting biases that are likely to arise when constructing such forecasting models.

For example, a study might show the significance of a particular variable in a forecasting model, but does that variable prove significant if it is forced to compete for inclusion with a full range of other candidate variables? How should that variable’s standard error be adjusted upward to reflect the fact that it was selected from a list of other variables because it was found to be a useful forecaster? And did the study in question report on all the other variables that were tried but that did not work?

---

3 Foster et al. (1997) propose techniques for assessing the R-squareds of OLS regressions in asset pricing when researchers implicitly select the best k of m regressors to use in the forecasting model.
Additional reporting biases may arise from selective reporting of out-of-sample tests. For example, one could do an exhaustive search across many possible specifications to identify a forecasting models that “works” both in-sample and out-of-sample, and only report in-sample results for models that pass this test. But such a search undermines the legitimacy of out-of-sample testing. Can one have confidence in any out-of-sample test that is reported simultaneously with the construction of an in-sample model? What out-of-sample validation technique can one use to provide a convincing validation?

Our approach takes explicit account of a broad set of possible modeling choices, both in our in-sample analysis and our out-of-sample validation. And we adjust standard errors to avoid selection biases that otherwise would occur, and also to correct for biases in R-squareds related to the use of overlapping time-series observations. We first employ a forward-selection model capable of selecting parsimonious time series forecasting specifications from the entire list of potential forecasting variables. The forward-selection approach accomplishes this via successively choosing each new variable as the one with the greatest contribution to the model R-squared. While there are other model selection techniques (see Hastie, Tibshirani, and Tibshirani 2017), we believe the iterative process of forward selection closely mirrors what researchers have done – though not done explicitly – in practice. Our approach avoids any bias associated with selecting in-sample models based on their out-of-sample properties, since we do not even consider out-of-sample performance at this stage of the process. Crucially, we adjust our standard errors to take into account the selection of the in-sample best-performing explanatory variables that is a feature of the forward selection algorithm. We show that the failure to adopt this methodological approach could (1) result in spuriously small in-sample
standard errors, and (2) invite the selection of in-sample models that pass out-of-sample tests by construction.

We consider various approaches to out-of-sample testing. In a time series context such as ours (as panel models are not generally appropriate for energy forecasting due to a lack of like time series), parsimonious models are attractive. Nevertheless, even when adopting a parsimonious modeling discipline (selecting only a small number of potential forecasting variables), we are unable to construct a methodology for systematically identifying, in real time, a set of forecasting variables that works well enough out-of-sample. We do identify interesting candidates for parsimonious oil forecasting models (that include our new NLP measures) by resorting to a brute force search over all possible forecasting models, but we remain agnostic about whether these models will prove useful in the future. A truly convincing out-of-sample test of these models can only be done via the passage of time.

Forecasting the price of crude oil is a central question in the oil-macro space. Leading contributions include Alquist, Kilian and Vigfusson 2013, Baumeister and Kilian 2015, Manescu and van Robays 2016, and Baumeister et al. 2020. Different approaches ranging from reduced-form monthly VARs containing the fundamental drivers of oil prices to DSGE models, product-spread models, or time-varying parameter models have been used. Although the models, the frequency of analysis and the variables considered in the oil-macro literature are generally different from ours, the general conclusion that it is hard to beat a random walk for out-of-sample oil price forecasting is in line with our findings.

A more related literature in finance investigates predictability in oil and other commodity markets. Examples include Bessembinder and Chan 1992, De Roon, Nijman and Veld 2000, Hong and Yogo 2012, Gorton et al. 2013, and Yang 2013. These studies provide evidence that
returns in commodity futures markets can be predicted using a range of aggregate and commodity-specific financial and macroeconomic variables. They are typically based on in-sample analysis with baseline models that contain six or seven predictors. The papers usually propose new predictors and examine whether these improve predictability. Our approach considers a wide range of financial and macro variables, including variables considered in this literature, as well as new text measures, and we test predictability in oil markets both in-sample and out-of-sample. In addition, we carefully control for small-sample biases. Our comprehensive approach to model selection and validation is informed by studies of equity markets that have shown the potential importance of reporting biases (e.g., Welch and Goyal 2008, and Harvey, Liu, and Zhu 2016).

Identifying relevant news and how it is associated with changes in market returns and risks is a central topic in asset pricing. Recently, economists have brought new tools to bear in examining this question, including the analysis of various aspects of the words that appear in newspaper articles or other textual sources, which have been applied to equity and exchange rate markets (for example, Tetlock 2007, Tetlock, Saar-Tsechansky, and Macskassy, 2008, Calomiris and Mamaysky 2019a, 2019b, Glasserman and Mamaysky 2019). Motivated by the ability of text-based measures to predict risk and returns in stock markets and their availability on a daily basis, we build on this literature and construct NLP measures for energy markets and examine their usefulness in predicting a set of market and fundamental variables. In this context, our work also relates to recent work using textual analysis in analyzing the oil market (Brandt and Gao 2019, Datta and Dias 2019, Loughran et al. 2019 and Plante 2019), but we consider a more comprehensive set of NLP measures.
The remainder of our paper proceeds as follows. Section 2 presents the list of forecasting variables we consider, and describes our data sources and our methods for constructing the NLP measures included in the models. Section 3 explains our choice of in-sample modeling structure, which involves a forward selection model using overlapping weekly observations to forecast eight-week ahead energy market outcomes. Section 3 also discusses our methodology for correcting standard errors for variable selection bias, and for correcting R-squareds for the use of overlapping observations, and presents our in-sample results. Section 4 presents our out-of-sample findings, using a variety of approaches and modeling choices. Section 5 concludes. All data series, including the energy topics, are available at https://sites.google.com/view/hmamaysky.

2. Data and Construction of Variables

We consider a variety of traditional variables that capture returns and risks in the macroeconomy and the oil market, as well as new predictors constructed using TR news articles about the energy sector. The raw data used to construct the variables used in our analysis come from Bloomberg, the Energy Information Administration (EIA), the Wall Street Journal, and the Federal Reserve Board. This section describes the data and the construction of the key variables of the empirical analysis. We forecast on an eight-week ahead basis, using weekly observations; four-week ahead results are qualitatively similar, and are reported in the Online Appendix.

2.1. Energy and Macro Series

Our dependent variables – oil spot and futures returns, company stock returns for oil majors, realized oil volatility, oil production, and oil inventories – are crucial for investors, policymakers, and analysts, as they seek to understand the dynamics of oil markets. Our
explanatory variables include lags of the dependent variables, as well as many other variables described below.

We would like to use observations at the highest possible frequency to take full advantage of the links between information arriving in the market and market reactions to that information. Although oil prices are available daily, production and inventory data are available at a weekly frequency in the U.S. We therefore perform our analysis using weekly observations. U.S. crude oil production and crude oil inventories (including the strategic petroleum reserve) data are released by the EIA on Wednesdays at 10:30am Eastern time. For some weeks, typically those involving holidays, releases are delayed by one or two days. As a result, when inventories or production levels are used as forecasting variables, for the dependent variables we use weekly returns (based on spot or futures prices, and oil company stocks) that use the closing price on Friday and measure changes to the Friday close of the following week. Our oil realized volatility series ends on Fridays, and looks back 30 trading days; we only use this measure in the eight-week ahead forecasting analysis. This timing convention ensures that the inventory and production numbers do not overlap with the variables they are intended to forecast. When we are forecasting inventories and production levels, we use forecasting variables that end on the Tuesday of a given week. For the other forecasting regressions, we use explanatory variables that end on Thursdays to avoid overlap with the 2:30 pm Friday close of oil futures markets.

We consider the U.S. oil benchmark, West Texas Intermediate (WTI). Our measure of spot price returns \( \frac{P_{t+j}}{P_t} - 1 \) captures the percent spot price change over a \( j \) week period (here \( t \) is measured in weeks), where \( j = 4, 8 \). We use the front-month futures contract as the

\[ 4 \] There would be overlap between the four-week ahead realized volatility measures and our forecasting variables. For this reason, we focus our analysis on eight-week ahead forecasts, though the four-week ahead results are qualitatively similar for all other dependent variables.
measure of spot price (as in Kilian and Vega 2011 and Loughran et al. 2019). While modeling spot returns is useful for capturing the dynamics of oil price changes, spot price changes do not represent an investable return because they ignore storage and transportation costs.

To capture investable oil price returns, we measure realized returns from investing each week in the front-month oil futures contract. On weeks that the front month future expires, we measure returns using an investment in the second month oil future (which will become the front month at the end of the week). We construct \( j \)-week cumulative returns as the product of the past \( j \) weeks’ one-week returns. This measure captures the returns to a specific investment strategy, and reflects changes in spot prices, the realization of risk premia, and changes in risk premia over time.\(^5\) In a similar vein, energy company stock returns are also calculated as \( j \)-week percent changes in Friday to Friday stock prices. We consider three large multinational oil and gas companies’ stock returns (BP, Shell, ExxonMobil).\(^6\) Our measure of oil price volatility is the 30-trading-day realized volatility of WTI prices from Bloomberg.

To summarize, our eight dependent variables are oil spot and future returns, oil realized volatility, the stock returns of BP, Shell, and ExxonMobil, and oil inventories and oil production. Our forecasting variables include lags of these, as well as several measures commonly used in the literature to predict commodity returns. These predictors include the VIX, the yield on the ten-year Treasury note, the trade-weighted value of the dollar, S&P 500 returns, a measure of economic activity, and the oil future basis. Our measure of global economic activity is the month-over-month growth rate of world industrial production (WIPI) introduced by Baumeister.

\(^{5}\) Further details about the futures return calculation are in the Online Appendix.

\(^{6}\) For BP and ExxonMobil, we use NYSE stock prices. For Royal Dutch Shell, we use Euronext Amsterdam Royal Dutch Shell Class A prices; therefore, the Friday close may be before the release of the EIA inventory and production data. For this reason, we use Monday to Monday closing prices for our Shell return calculation.
and Hamilton 2019. Our basis measure is the annualized ratio of the 3-month to 1-month price for crude oil futures, namely \( \text{basis}_t = (F_{3,t}/F_{1,t})^6 - 1 \) (raising to the power of 6 converts this to an annualized measure). A positive basis indicates the curve in contango, and all other things being equal buying longer-dated futures will lose money as they roll down the curve. All right-hand side variables are released into the market prior to the Friday 2:30 pm oil futures market close. We refer to the variables defined in this section as our \textit{baseline} measures.

Table I presents definitions for all the variables used in the empirical analysis. Table II reports summary statistics for all the variables used as either dependent variables or forecasting variables in the 1998-2020 sample period. For example, the average eight-week return on oil futures has been 1.35% with a standard deviation of 13.78%. The average eight-week return on oil spot has been lower over the same period, at 0.64%, with a higher standard deviation of 14.8%. Energy company stocks, on the other hand, have lower average returns (ranging between -0.35% and +0.20%) and are less volatile (ranging between 7.61% and 9.94%).

2.2. Risk Premium Measures

In addition to the traditional energy market and macro predictors, we include several measures that are useful for gauging market risk premia. The first of these, \( \text{vix}_\text{spx} \), measures the difference between the VIX index of short-term implied volatility of S&P 500 options and the last 30-day realized volatility of the S&P 500 index. Many researchers, for example Bekaert and Hoerova 2014, argue that the difference between the VIX index and forecasts of future realized volatility reflect the variance risk premium. Here we assume lagged realized volatility is a reasonable proxy for expected future volatility. In a similar vein, we include \( \text{ovx}_\text{cl1} \) which is

---

7 Baumeister et al. 2020 evaluate global economic activity indicators based on their forecasting performance for the real price of oil and find that WIP is one of the most useful indicators that has been proposed in the literature. For example, it outperforms measures based on shipping rates.

8 Overall, four-week and eight-week summary statistics look similar.
the difference between the OVX index of implied volatility on an ETF which owns WTI futures and the last 30-day realized volatility of crude oil prices. We use ovx_cl1 as a proxy for the volatility risk premium in the oil markets.

To construct an alternative measure of the risk-premium in energy markets, we follow the method of Hansen and Jagannathan 1991. Letting \( R \) be an \( n \)-dimensional vector of daily gross returns from a candidate set of securities, the unconditional version of the basic no-arbitrage condition of asset pricing is \( 1 = E[mR] \) (note 1 is an \( n \)-dimensional vector), where \( m \) is the stochastic discount factor (SDF). Assuming \( m \) is in the linear span of the security returns implies

\[
m = 1^T E[RR^T]^{-1} R. \tag{1}
\]

Furthermore, it is well known that the expected excess return on a security is proportional to the negative of its covariance with the SDF (Cochrane 2005). The conditional version of this relationship can be written as

\[
E_t R^e_i = - \frac{\text{cov}_t(m, R^e_i)}{E_t m}, \tag{2}
\]

where \( R^e \) is the daily excess return on security \( i \) and the expectations are taken as of week \( t \). We estimate (1) in windows of our data using daily returns on the Credit-Suisse WTI futures total return index, the total return of the S&P 500 index, a U.S. Treasury total return index from Bloomberg (which roughly tracks 10-year bonds), the total return from investing in 6-month U.S. T-bills, and the total return of the MSCI World Energy Sector index. Then using the estimated SDF \( \hat{m}_t \), we approximate the week \( t \) conditional expectation in (2) by calculating the covariance between the excess return of the WTI futures index and \( \hat{m}_t \) over the prior 252 trading days. We use this as our estimate of the WTI risk premium.

We use three different estimation methods for \( \hat{m}_t \). In one, we use a rolling 756-day window (roughly three years), and use the \( \hat{E} R^e_{WTI} \) estimate from the window ending on day \( t \) as
the then prevailing estimate of the WTI risk premium. We refer to this series as \textit{SDF\_rolling}. In another variant, we use an expanding window that starts at a minimum of 756 days, and then expands for each successive day in the sample. We refer to the WTI risk premium estimate from this approach as \textit{SDF\_growing}. Both the rolling and growing SDF is used in our out-of-sample analysis. In our in-sample analysis, we use the SDF constructed with the entire sample, which we label \textit{SDF\_fullSample}. All calculations are done in windows that end on Tuesdays (for inventory and production forecasts) and Thursday (for all other forecasts).

2.3. Text Analytics

Our corpus for NLP analysis includes all 2.07 million articles in Thomson Reuters (TR) that are labeled as being energy-related from 1998 to 2020. We say an article is energy related if it is classified by TR as belonging to one of the 98 topics, the full list of which is available in the Online Appendix. To perform topical analysis, we compiled a list of energy-related words, bigrams and trigrams (two- and three-word phrases respectively) from several energy industry glossaries and other sources of energy words and phrases. This resulted in a list of 387 tokens. We then construct a $387 \times 387$ co-occurrence matrix which measures the cosine similarity between this initial list of tokens. The cosine similarity between tokens $i$ and $j$ is given by

$$\frac{w_i^T w_j}{\|w_i\| \|w_j\|}$$

where $w_i$ is the vector measuring the number of times token $i$ appears in all the documents in our TR corpus, and $\|w\|$ is the Euclidean norm of $w$. We then employ the Louvain clustering algorithm (see Blondel et al. 2008) to identify disjoint (i.e., non-overlapping) word groups that maximize the modularity (see Newman and Girvan 2004) of the network represented by the word co-occurrence matrix. In this step, we set the diagonal of the co-occurrence matrix to zero, which then yields eight topics from the Louvain algorithm. The eighth topic contained only
several tokens, so we reallocated these tokens from the eighth topic to the other seven topics to maximize the resultant seven-topic network’s modularity.

Once we identified the initial set of seven topics, we calculated the average co-occurrence of a large set of additional candidate energy related words, bigrams and trigrams with the 387 initial energy words, bigrams and trigrams from the energy industry glossaries. We then identified from the list of additional potential energy words those whose maximum topical co-occurrence was very high relative to its average topical co-occurrence. For example, the candidate token *shell*, which was not part of our original 387-token list, had an average cosine similarity with the existing tokens in topic 1 of 0.2076, whereas its average co-occurrence across all seven topics was 0.0374. The resultant difference of 0.1702 was the second highest of all our candidate tokens. We therefore included *shell* in our augmented token list. The intuition behind this metric is that we wanted to exclude words that had high co-occurrence with *all* our topical clusters because these tended to be generic words (such as *said* or *though*). However, words that had a high co-occurrence with a single topic tended to be energy-related words or bi- or trigrams. Applying this process to a large set of candidate tokens yielded an additional 54 tokens, which we then placed into one of the existing seven topical groups so as to maximize the network modularity of the new, 441-token network. We refer to these 441 tokens as the *energy words*.

Figure 1 displays the word clouds for each of our seven topics.9 We label the topical categories based on our interpretation of the common topical link defined by the words that appear in each of these word clouds. Interestingly, the topics defined by the word clouds have readily interpretable meaning and exhibit sufficient variation over time to be useful in our analysis. We discuss this further in the next subsection. We label the topics as follows: company

---

9 As a robustness check, we verified that latent Dirichlet allocation (LDA) produced similar topics to the Louvain-based ones. A summary of this analysis is in the Online Appendix.
(Co), global oil market (Gom), environment (Env), energy/power generation (Epg), crude oil physical (Bbl), refining and petrochemicals (Rpc), and exploration and production (Ep). We classify article $i$ into topical category $\tau$ by looking at the fraction of the energy words appearing in this article that belong to topic $\tau$, or

$$f_{i,\tau} = \frac{N_{i,\tau}}{\sum_{j=1}^{\tau} N_{j,\tau}}$$

where $N_{i,\tau}$ is the number of energy words in article $i$ that belong to topic $\tau$. Notice the article topic weight sum to one.

The sentiment of article $i$ is defined using the Loughran-McDonald sentiment dictionary as follows

$$s_i = \frac{\text{Pos}_i - \text{Neg}_i}{\text{Total}_i}.$$ 

Here $\text{Pos}_i$, $\text{Neg}_i$, and $\text{Total}_i$ are the number of positive, negative and total words in article $i$ after stop words have been removed. We define as article’s topic sentiment as the product of topic frequency and sentiment, or

$$s_{i,\tau} = f_{i,\tau} \times s_i.$$ 

Given that article frequencies sum to one, topical-sentiment sums to sentiment for each article. Unusualness is defined using the entropy concept introduced in Glasserman and Mamaysky (2019) and Calomiris and Mamaysky (2019a). Specifically, we define article $i$’s unusualness as the negative average log probability of all 4-grams appearing in that article, or

$$e_i \equiv -\sum_{j \in \text{4-grams in the article}} p_j \times \log \hat{m}_j,$$

where $p_j$ is the fraction of all 4-grams represented by the $j^{th}$ 4-gram in article $i$, and $\hat{m}_j$ is the empirical probability of the fourth word in the 4-gram conditional on the first three, estimated.
over a training corpus using all articles from months $t - 27$ to $t - 4$. Glasserman and Mamaysky 2019 showed that entropy can be used to measure the novelty of an article, and that higher entropy news flow is more informative for forecasting future market outcomes.

We aggregate our article-level news measures to the daily level by taking a word-weighted average of all articles released between 2:30 pm of the prior business day and 2:30 pm of the present business day. For Mondays, we count articles from 2:30 pm to midnight on Friday, in addition to articles from 2:30 pm on Sunday to 2:30 pm on Monday. We then take an equal-weighted average of the daily news flow measures (article count, topical frequency, topical sentiment, and entropy) ending on Tuesday or Friday of a given week. We end on Tuesdays for news series meant to forecast oil production and oil inventories, and end on Fridays for news series used to forecast all other variables. Finally, we calculate the average number of daily articles that mention energy markets in the TR corpus in weeks ending on Tuesday or Fridays. This yields 16 distinct text-based series: article count, entropy, the seven topical frequency series (labeled $f[Topic]$), and the seven topical sentiment series (labeled $s[Topic]$). We standardize all our text-based series, except entropy, to have mean zero and unit variance. In the regression analysis, we use four-week rolling averages of all the weekly standardized text series. In addition to these, we also add three measures of aggregate news flow: the first principal components (PCAs) of the seven topical frequency series ($PCAfreq$), of the seven topical sentiment series ($PCAsent$), and of all fourteen series together ($PCAall$). The PCAs are calculated using the four-week averages of the weekly series, where the four-week averages have been normalized to be mean zero and unit variance.

---

$^{10}$ $\hat{m}_l$ for 4-gram $w_1w_2w_3w_4$ (the $w_k$’s refer to words) is the fraction of times $w_4$ follows the word sequence $w_1w_2w_3$ in the training corpus. Prior to doing this analysis, we tokenize and stem the documents, but do not remove the stop words. When a 4-gram has not been seen before in the training corpus, we assign to it a probability of 0.1. For more details of this methodology, see Glasserman and Mamaysky 2019 and Calomiris and Mamaysky 2019a.
2.4. Behavior of Energy News

We plot four-week averages of the nineteen series in Figure 2. As is clear from the figure, the text-based measures of news flow in energy markets display a large amount of time variation. To gain further insights into our measures of energy news flow, we explore whether unusual movements in our text measures correspond to important real world events in energy markets. To identify potentially interesting events, we look for the two most negative 4-week in the four-week average series of our seven topical sentiment series. For each episode, we then identify a set of candidate articles. Candidate articles are those that have entropy scores equal to or higher than 2, that contain 100 or more words after stop words are removed, and that have a topic allocation above 0.8, i.e., \( f_{i,t} > 0.8 \). These articles typically present stories about specific situations, and are not news alerts, daily summaries, or statistical tables. We then manually looked through the headlines and connected them to specific energy market episodes. Almost all extreme moves in topical sentiment were associated with important events in energy markets, but we chose to focus on six in particular, each of which belongs to a distinct topic. The end dates of the 4-week topical sentiment changes associated with these six episodes are marked with stars in Panels A and B of Figure 2.

While the events were identified based on changes in topical sentiment (Panel B), it is clear from Panel A that all of these events are also associated with large increases in the fraction of total news coverage devoted to that particular topic category. This points to a more general feature of the topical sentiment and frequency series, namely that for each topic the two aggregate series are very negatively correlated (the correlations range from -0.57 to -0.93). Spikes in topical frequency tend to occur at times of very negative topical sentiment.
Table III shows the six episodes of interest that we identified (in the same order as they appear in Figure 2). For each episode, we show the sentiment, entropy and headlines of the five most negative sentiment articles. The particular historical episodes associated with sharp drops in topic sentiment, with the associated topic category in parentheses, are: the UK fuel protests in September of 2000 (company), the attempted Venezuelan coup in 2002 (global oil markets), the Volkswagen emissions scandal in 2015 (environment), the Enron bankruptcy hearings of 2002 (energy/power generation), Hurricane Katrina in 2005 (crude oil physical), and the BP oil spill in 2010 (exploration & production).

It is notable that each event is classified into an appropriate topic. For example, many articles discussing the UK fuel protests focused on their impact on business. Others discuss the reduction in OPEC output caused by the civil unrest in Venezuela affecting global oil markets. Furthermore, these events were identified algorithmically, and not cherry-picked by us. It should be further noted that we assigned names to topics by looking only at the word clouds; the close match of headlines with their associated topics is a validation of the usefulness of our methodology.

These results indicate that our news-based measures of energy markets capture important aspects of energy news in a way that traditional series cannot do. In sections 3 and 4, we look to exploit the information content of these news series for both in- and out-of-sample forecasting of our eight dependent variables associated with the U.S. energy markets.

3. In-sample Predictability

Given the limits of degrees of freedom inherent in time series analysis, to avoid overfitting, we employ a forward-selection model to choose parsimonious time series forecasting specifications from a broad list of potential forecasting variables. The forward-selection
approach accomplishes this via successively choosing each new variable as the one with the greatest contribution to the model R-squared.\textsuperscript{11} We believe this is the method implicitly followed by many papers on predictability in energy markets: they select a smaller set of variables that work “best” from a larger list. Our goal is to apply the same methodology formally – not implicitly – to all our dependent variables, to let the rich data speak, and come up with reliable inference that accounts for the selection criterion of our variables.

Our 35 forecasting variables include lagged measures of spot and futures returns, energy company stock returns, oil price realized volatility, the change in oil price realized volatility, the change in oil production, the change in oil inventories, the VIX, the difference between the VIX and the realized volatility of the S&P 500 index, the market return on the S&P 500, the yield on the ten-year Treasury note, the change in the trade-weighted value of the dollar, the basis, the year-on-year growth rate of WIPI, and 19 NLP measures including article count, entropy, the seven topical frequency series, the seven topical sentiment series, and the three PCAs.\textsuperscript{12} To maintain consistency of our forecasting variables, they are calculated over the last four weeks for both the four- and eight-week ahead forecasting regressions; similarly, lagged realized volatility is always measured over the prior 30 trading days. Prior to running the in-sample forward selection procedure, we first detrend all variables, to ensure that trend does not contribute to forecastability, and then residualize the data by regressing out the four-week version of the lagged dependent variable from both the left- and right-hand sides of the in-sample specification. We residualize because, otherwise, the lagged dependent variable would frequently be chosen in the forward selection procedure. Our forecast horizon is either four- or eight-weeks ahead.\textsuperscript{13} We

\textsuperscript{11} We use the \texttt{fs()} method from the \texttt{selectiveInference} package to perform this analysis.
\textsuperscript{12} To be conservative, we use only one lag because we already have numerous forecasting variables.
\textsuperscript{13} The four-week horizon results, reported in the Appendix in Table A.IV, are consistent with the eight-week results.
use forward selection to select seven variables out of our set of 35, after all data has been detrended and residualized.\footnote{The number of variables considered in studies examining return predictability in commodity markets ranges between one and seven (see Table 1 in Baumeister and Kilian 2017). In our out-of-sample analysis, we also consider a two-variable in-sample selection model.}

The model is estimated using weekly observations with either four- or eight-week ahead overlapping observations, which substantially increases the possibility of finding spurious forecasting relationships. It is well-known that the use of overlapping observations will downwardly bias standard errors and upwardly bias R-squareds (see, for example, Hodrick 1992, Kirby 1997, Ang and Bekaert 2007, and Boudoukh et al. 2008). Furthermore, we employ forward selection for choosing a parsimonious set of in-sample regressors, which tends to introduce upward bias in the R-squareds, and downward bias in the standard errors as well. To control for both of these sources of finite sample bias, we construct bootstrapped distributions for our t-statistics and R-squareds, a methodological contribution of our paper. We now describe our methodology.

3.1. Controlling for Overlapping Observations and Sample Selection

We assess the in-sample forecasting power of our model by simulating the data and checking whether the empirical R-squareds are anomalous relative to the simulated R-squareds. We first estimate an AR(8) process for the dependent variable. We then simulate a new dependent series based on the AR(8) model. Next, we rerun our in-sample regressions, using all of the actual 35 forecasting variables, except replacing the lagged dependent variable with the simulated series. By construction this simulated dependent series is independent of all our forecasting variables, except for the lagged dependent series itself which controls for the mechanical autoregressive properties of the series being forecast. In one round of this
simulation, we calculate the standard OLS t-statistics for the selected variables, keeping track of
the order of selection, i.e. the t-statistic for the first selected variables, for the second selected
variable, and so on. We also record the R-squared of this one simulation round. We then repeat
this process 1,000 times to build a bootstrapped distribution for the ordered t-statistics, as well as
for the model R-squared. This process controls for both the selection and overlapping
observation properties of our in-sample procedure. More details are in the Appendix.

To give a sense for the impact of small-sample biases, Figure 3 shows the bootstrapped
R-squared distributions for forecasting eight-week ahead oil futures returns and oil volatility.
Under the null hypothesis of no predictability, other than the mechanical autoregressive
properties of both series, there is a wide range of R-squareds in our simulated runs. In fact, the
dual small-sample problems of overlapping observations and variable selection lead to very high
in-sample R-squared. When reporting our actual R-squareds in Table IV, we show the
percentage of simulated R-squareds that are lower than the actual ones (in the table row labeled
“CDF”). Rather than interpreting the outright value of the R-squared, a very high CDF value
indicates that there is evidence of in-sample predictive ability even in the face of these biases.

To understand the impact of small-sample biases on p-values, Figure 4 shows the
distribution of the ordered t-statistics for the seven forward selected variables, under the null
hypothesis of no predictability for forecasting oil futures returns and oil volatility. The butterfly
shaped distributions show the extreme bias that forward selection introduces to standard OLS t-
statistics. The first chosen t-statistic (the widest bimodal distribution) shows that the modes for
the t-statistic of the first selected variables are close to -6 and +6 respectively. The modes for the
seventh selected variable are expectedly smaller in magnitude, at approximately -3 and +3. To
the extent that other energy forecasting papers implicitly follow a variable selection methodology
reminiscent of forward selection (i.e., choose a subset of the best forecasting variables from a larger set), their standard error distribution under the null has the butterfly pattern shown in Figure 4. Not adjusting for this introduces obvious biases. We adjust for this issue by calculating p-values in our in-sample regressions by comparing the OLS t-statistics in our actual regressions to these distributions. Let \( \hat{p} \) be the fraction of simulated t-statistics for a given ordered selected variable (e.g. the second selected variable in a given specification) that are less than the t-statistic for the actual ordered selected predictor. Our bootstrapped p-value is reported as \( \min(\hat{p}, 1 - \hat{p}) \). A p-value less than or equal to 0.025 (0.05) indicates significance at the 5% (10%) level. We don’t present bootstrapped distributions of R-squareds and t-statistics for all dependent variables (they are available from the authors), but Table IV summarizes this information.

3.2. Results

Table IV presents the regression results for our 8 dependent variables using stepwise forward selection that chooses seven variables for each model. Only the predictors that were chosen by at least one model are presented in the table. For each dependent variable, we present coefficient estimates of the selected predictors, which are standardized, along with corresponding p-values as described in the last section. Our standardized coefficients report the standard deviation change in the dependent variable due to a one standard deviation change in the forecasting variable.\(^{15}\)

The standardized coefficients for the selected predictors range between 0.07 and 0.71 in absolute value. For example, a one standard deviation increase in economic activity growth

\[^{15}\text{These are } b \times sd(RHS)/sd(LHS) \text{ where } b \text{ is the estimated coefficient, } sd(RHS) \text{ is the standard deviation of the forecasting variable, and } sd(LHS) \text{ is the standard deviation of the dependent variable, calculated for the set of dates available for each individual forecasting regression.}\]
(WIPImom) over the previous month increases eight-week ahead oil futures returns by 3.7% 
(0.27 × 13.78%). Or, a one standard deviation increase in average $sGom$ over the past month – 
positive sentiment about global oil markets – increases BP returns by 1.3% (0.13 × 9.94%) over 
the next eight weeks. Moreover, around 63% of the selected variables are statistically significant, 
even after adjusting for overlapping observations and variable selection. The variables chosen 
by the forward selection method are generally both economically and statistically significant.

The actual adjusted R-squareds of the forecasting regressions also look good, ranging 
from 13% to 37%. At the bottom of Table IV, we present mean of the bootstrapped adjusted R-
squareds for each regression and their corresponding CDFs (i.e., the percentage of bootstrapped 
R-squareds that are lower than the empirical one). They suggest that the impressive empirical R-
squareds observed in our models are highly unlikely to have been generated by chance. That is, 
under the null hypothesis of no relationship between the dependent and the independent variables 
(except for the presence of the lagged independent variable), the probability of adjusted R-
squareds being greater than or equal to the empirical R-squareds reported is less than 0.5% for all 
models.

Turning to the composition of the selected variables, out of the 56 predictors selected 
across all the models, 26 of them are text measures (about 46 percent), and of these about 73% 
are statistically significant. (The candidate explanatory variables are roughly equally split 
between text-based and traditional measures.) And, four of the text measures are chosen 
statistically significantly at least two times, namely entropy, $sEnv$, $fRp$,$c$, and $sEp$. For example, 
BP returns, Shell returns, change in oil production and change in inventories are all forecastable 
by $sEp$. While 54 percent of the predictors selected are non-text measures, only 53 percent of 
them are statistically significant. And, six of the selected non-text variables are chosen
statistically significant at least two times, which are change in inventories, \(DFX\), \(basis\), \(tnote_{10y}\), \(sp500Ret\), and \(WIPImom\). For instance, not surprisingly, \(DFX\), \(basis\), and \(WIPImom\) are all economically and statistically significant in forecasting oil futures returns.\(^{16}\) These results suggest not only that our new text measures are selected frequently (about as much as the traditional non-text measures), but also that they are statistically significant more often than the traditional measures. We conclude, therefore, that text measures are important in-sample forecasting variables for the oil market.

However, some variables are more forecastable by text measures than others. For example, overall, non-text measures seem more useful in forecasting futures returns, while oil spot price changes and changes in realized oil volatility are forecastable by text measures. This is a surprising result, as one may expect text measures to be more useful in predicting risk premia, and hence futures returns. This takes us to the next question. Could text measures be selected because they are proxies for risk?

To address this question, we take the forward selection models considered above and presented in Table IV, and add each of our risk measures presented in Section 2.2 – \(vix\), \(vix_spx\), \(ovx_{cl1}\), and \(sdf\_fullSample\) – one by one after the seven variables were selected by stepwise forward selection. As \(vix\) and \(vix_spx\) were included in the list of candidate variables for our forward selection procedure, they are included in this test only if they were not selected in the first place. Risk measures are natural predictors of returns because time variation in expected returns may reflect forward-looking compensation for risk. Looking at how coefficients on text measures change, we find that adding the risk measures does not reduce the coefficients on the

\(^{16}\) The \(basis\) variable comes in with a positive coefficient in Table IV for both futures returns and spot changes, indicating when the curve is in contango, expected oil returns are higher. This is surprising. But the result is due to the interaction of \(basis\) with the other regressors, because in univariate regressions, \(basis\) forecasts with the expected negative sign.
text measures towards zero (Appendix Table A.III). In other words, one should not interpret the selected text measures as proxies for some omitted risk factor. Interestingly, sdf_fullSample enters significantly only once in these augmented regressions; though we will see the SDF variable often plays a larger role in the out-of-sample analysis in the next section.

To sum up, we find compelling evidence of in-sample predictability after carefully controlling for small-sample biases. Next, we explore whether it translates into out-of-sample performance.

4. Out-of-sample Predictability

We consider three out-of-sample approaches. The first takes the variables chosen in the in-sample forward selection model as given. Because this approach suffers from lookahead bias, it gives our out-of-sample tests the best chance for success. If a constant set of variables that are chosen using future information do not work well for out-of-sample forecasting, then perhaps the set of forecasting variables changes over time. In recognition of this, our second approach identifies a time-varying set of candidate forecasting variables, and checks their out-of-sample performance. The third approach engages in a brute force search over all possible two-variable forecasting models, and checks their out-of-sample performance. For our set of forecasting variables, we augment the series that are available for the in-sample forward selection model in Section 3 with the two SDF-based expected return forecasts (SDF_rolling, SDF_growing), OVX_CL1, and sent, which is the sum of the four-week topical sentiments in a given week.

4.1. Ability of in-sample regressors to forecast out-of-sample

To evaluate the effectiveness of our in-sample model selection technique for out-of-sample forecasting, we first lower the bar by considering the seven forecasting variables selected for each dependent variable using the forward selection methodology over the entire sample.
Our dependent variables are eight-week ahead returns or changes, as they were for the in-sample analysis; for each dependent variable, the seven forecasting variables are those that were the most successful in-sample forecasters according to the forward selection model. In our out-of-sample analysis, to avoid imposing a forward-looking trend, we do not detrend any series, nor do we regress out the lagged dependent variable as we had done in the in-sample analysis. Instead, we add the four-week version of the lagged dependent series as an additional forecasting variable. This approach clearly suffers from lookahead bias because the best full-sample set of forecasting variables would not have been known for most of our data sample; in this sense, this methodology is extremely favorable for the out-of-sample test.

We then run rolling five-year lasso regressions, with automatic penalty parameter selection using ten-fold cross validation, to estimate rolling coefficients for out-of-sample forecasting of eight-week ahead changes or returns. We ensure each five-year training window uses data only from inside the window. Using the lasso coefficient estimates in each training window, we then use the dependent variables available at the end of the window to make an eight-week ahead forecast. We then march the training window forward by one week, re-estimate the model, and make another eight-week ahead forecast.

To control for the possibility of overfitting a seven independent variable model in five-year training windows, we redo the out-of-sample procedure just described using a two-variable version of the forward selection model run over the entire sample. As before, we do not detrend,

---

17 We consider rolling regressions rather than expanding window to account for possible regime shifts in the data. An expanding window would ultimately settle on a single regime, and not allow for structural breaks in the forecasting relationships.
18 The first right-hand side observation in a five-year training window occurs eight weeks after the window’s start to allow for the lagged dependent variable as a regressor. The last right-hand side observation in the training window occurs eight weeks prior to the end of the window to ensure that the left-hand side variable does not extend beyond the five-year training window. For specifications where PCAsent (the other PCAs are never chosen in the full-sample forward selection model) was chosen in the in-sample model selection, we re-estimate PCAsent using normalized four-week averages of the topical sentiment series in each five-year training window.
do not regress out the lagged dependent variable, but we do add the four-week version of the lagged dependent variable as the third forecasting variable. This test also suffers from look-ahead bias, but is perhaps less susceptible to overfitting in the five-year rolling training windows.

To measure the out-of-sample efficacy of the two versions (the two- and seven-variable models) of the lasso model, we use as a baseline model the five-year rolling averages of the left-hand side variable. We are again careful to make sure that the averages of eight-week changes used for what we call the constant model do not extend outside of the five-year training window. Our main test is to look at the ratio of the mean-squared-error (MSE) of our two- or seven-variable lasso models (which suffers from lookahead bias) relative to the mean-squared-error of the constant model, i.e.

\[
MSE \text{ ratio} = \frac{MSE_{lasso}}{MSE_{constant}}
\]

Note that the out-of-sample R-squared is simply one minus this ratio. Therefore, when this ratio is above one, i.e. when the lasso model generates a higher error variance than the constant model, the out-of-sample R-squared will be negative.

Table V shows that the out-of-sample forecasting power, for eight-week ahead returns and changes, of the in-sample forward selected model is generally poor, as measured by the MSE ratio. Of the 16 cases considered (eight forecasting variables for the two- and seven-variable model versions), the constant model outperforms the lasso model 12 times. In the cases of \textit{xomRet} and \textit{bpRet}, the two-factor lasso model barely ekes out an outperformance with an MSE ratio of just below one. Only for changes in oil realized volatility is there strong evidence that the forward selected variables outperform the constant model in an out-of-sample setting. This outperformance is not surprising because volatility is known to be stationary, and therefore changes in volatility are mean-reverting. This is exactly what the analysis picks out. In all other
cases, the out-of-sample forecasting power of our model, which includes traditional and text-based forecasting variables, and which suffers from lookahead bias, is extremely poor.

4.2. Ability of a time-varying variable set to forecast out-of-sample

One problem with the approach in the prior section is that the two or seven variables chosen by the in-sample model selection are fixed over time for all out-of-sample forecasting windows. However, it is likely that the usefulness of forecasting variables ebbs and flows over time. To allow for this possibility, we run univariate regressions of each dependent variable (eight-week ahead changes or returns) on each forecasting variable in rolling five-year training windows. Within each training window, we then rank each forecasting variable (for each dependent variable) based on its standalone R-squared. We classify our forecasting variables into two groups: the text group contains our text-based measures and the baseline group contains all the other forecasting variables, including the SDF variables (both the rolling and the growing version), as well as the \textit{OIVX\_CL1} variable.

With the R-squared rank of each of the text and baseline variables in hand, we then form three candidate forecasting models: the first contains only the baseline variable with the highest R-squared in the training window (we refer to this as the 1-0 model); the second contains only the text variable with the highest training window R-squared (the 0-1 model); and the third contains both the top baseline and text variables (the 1-1 model). For each of the three models, we estimate a lasso regression in the training window, as described in the previous section. We then use the coefficient estimates from the lasso and the last observation of the forecasting variables in the training window, to make one eight-week ahead out-of-sample forecast. We then march forward by one week, re-estimate the training window, and make a new eight-week ahead forecast.
We also explore a variation of the above in which we choose 2 baseline variables (the 2-0 model), or 2 text variables (the 0-2 model), or the two top baseline and text variables (the 2-2 model), where all variables are ranked by their training window R-squareds within their peer variable set (i.e., text relative to text, and baseline relative to baseline). This variation of the approach allows for the possibility that more than one forecasting variable from each set may be useful.

Unlike our prior methodology, which uses the variable set from the in-sample forward selection model, these approaches are truly out-of-sample. They do not use any forward information to select the forecasting variables. Furthermore, they allow for the set of forecasting variable to vary over time. For each of our eight-week ahead dependent variables, Figure A.2 in the Appendix shows which forecasting variable is chosen at each point in our data sample in the 1-1 model. Two features are notable. First, there is some persistence in which variable is chosen over time, though this is not surprising given our use of overlapping five-year training windows. Second, despite this persistence there is as a large amount of time series variation in the selected baseline and text variables. This suggests that allowing for time variation in the forecast variable set may be useful.

Unfortunately, as Table VI shows, it is not. The top panel of the table shows the performance of 0-1 and 1-1 models as measured by the MSE ratio relative to the constant model (evaluated as described in the prior section). As the left columns in the top panel show, the MSE ratio is above one in 14 of the 16 cases. Only for the 0-1 model for eight-week ahead production changes and for the 1-1 model for changes in realized oil volatility are the MSE ratios less than one. As the bottom panel of the table shows, when we move the 0-2 and 2-2 models the performance is even worse, as 15 out of 16 cases have above one MSE ratios.
We also test the incremental value of our text measures in out-of-sample forecasting relative to using the traditional forecasting variables only. The two rightmost columns of Table VI show the ratio of the MSE of models 0-1 (text only) and 1-1 (baseline and text) respectively, relative to model 1-0 (baseline only). The results are now more positive. In many cases, we find that the MSE ratio of the models that include the text measures are lower than the MSE ratios of the baseline-only models. This is true also for the two variable variation as shown in the bottom panel of the table. Apparently, adding text-based information to traditional forecasters of energy market outcomes is useful, though not in general sufficiently useful to outperform the forecast coming from the constant model.

4.3. Data mining for out-of-sample forecasting

Thus far, we have attempted to devise a transparent and systematic methodology that identifies a subset of successful out-of-sample forecasters from a large pool of potential forecasters. Our lack of success suggests that it is very difficult to select, in real-time, a model that beats the random walk for out-of-sample forecasting of oil market outcomes. But are there any combinations of forecasting variables that appear to have successful out-of-sample performance? To address this question we undertake a brute-force search over all possible forecasting combinations. Of course, whether this set of variables – data-mined to work for a particular sample – will continue to work in the future is a question that will only be answered by the passage of time.

To conduct a brute-force data mining exercise to find out-of-sample predictability in our data, for every dependent variable we check all combinations of any two forecasting variables. As can be seen from the rows and columns of Figure 5, there are 39 forecasting variables in our study, and therefore there are \( \binom{39}{2} = 741 \) potential models. For each candidate model, we run a
lasso in five-year training windows using the methodology already described. We then make an out-of-sample forecast, move the window forward for one week, and repeat. Table VII reports the number of successful two variables combinations, which we define to be those that generate an MSE ratio relative to the constant model of less than one. For futures returns, there are 16 such pairs out of 741 possibilities. The other dependent variables are associated with a larger number of successful forecasting models, ranging from 23 for changes in WTI prices to a high of 213 for changes in realized oil volatility. The mean MSE ratio of successful models ranges from 0.9925 for ExxonMobil and Shell returns to 0.9123 for changes in realized oil volatility. Not surprisingly, changes in realized oil volatility are associated with the greatest number of successful forecasting combinations and low MSE ratios, whereas future returns are associated with fewer successful models that also have higher MSE ratios.

Figure 5 summarizes the successful forecasting models. In total, there are 554 successful models (shown in Table VII) which are associated with 1,108 selected variables. Figure 5 shows that 408 of these variables are text variables, and 700 are baseline variables. The most frequently appearing forecasting variables in the successful models are $OVX\_CL1$, $SDF\_growing$, and $SDF\_rolling$, all of which represent different sets of risk premia estimates for energy markets. Of the text variables, the most frequently selected ones are $sEp$ (the sentiment of stories related to the topic of exploration and production), $sCo$ (company sentiment), and $fBbl$ (the topical frequency of physical crude oil related articles). In Table A.V of the Online Appendix, we list all 554 successful variable pairs, with an MSE ratio of less than one, as well as the associated MSE ratio.

The takeaway from this analysis is twofold. First, it is straightforward to find evidence of out-of-sample predictability in the energy space, if one simply searches all the possible models.
Out of a total of $741 \times 8 = 5,928$ potential models, we find 554 with sub-one MSE ratios. Selectively reporting on just several of these would give the impression of readily identifiable out-of-sample forecasting performance. The second takeaway is the more sobering finding that we could not identify any systematic algorithm for finding such successful forecasting pairs on an ex-ante basis. Two open questions therefore remain. First, does such a systematic algorithm exist? (Brute force methods don’t count.) And second, will any of the 554 identified successful out-of-sample models continue to outperform in the future?

5. Conclusions

Energy forecasting is a challenging task. There are only a few time series that can be forecasted, and panel approaches are not applicable. Furthermore, the time period for which data are available is relatively short, and even if it weren’t, regime shifts would undermine the usefulness of much of the earlier data. Traditional in-sample approaches suffer from an implicit variable selection bias, although researchers typically do not formally adjust for this. In our work, we add text-based measures to the list of standard forecasting variables to consider, and we formally model the variable selection problem, control for selection bias in our R-squared evaluation and in our standard errors. With these adjustments, we still find systematic patterns of in-sample predictability. Many of the successful in-sample forecasting variables derive from measures based on the text of TR news articles about the energy space. This is one of the novel findings of our paper.

Despite successful in-sample forecasting results, we find it difficult to identify a transparent and systematic strategy for finding forecasting variables that lead to out-of-sample performance that is better than a rolling mean of the dependent variable. When we data mine, we identify many successful out-of-sample forecasting models. What is unclear is whether a
systematic variable selection method exists that will identify successful out-of-sample forecasting models without engaging in data mining. Furthermore, we do not know how many of our 554 successful out-of-sample models will remain so in the future. Answering these two questions should be a focus of future research in this area, which can take the results reported here as a point of departure.
References


Manescu, C., and I. van Robays, 2016, “Forecasting the Brent Oil Price: Addressing Time-Variation in Forecast Performance,” mimeo, ECB.


### Table I

**Data Definitions Summary**

Topic is one of company (Co), global oil market (Gom), environment, (Env), energy/power generation (Epg), crude oil physical (Bbl), refining and petrochemicals (Rpg), or exploration and production (Ep).

<table>
<thead>
<tr>
<th>Variable</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>FutRet$^8$</td>
<td>WTI front-month futures cumulative weekly returns (in %) starting in week $t$ through week $t+8$</td>
</tr>
<tr>
<td>DSpot$^8$</td>
<td>Percent change in the WTI spot price from week $t$ to $t+8$</td>
</tr>
<tr>
<td>DOilVol$^8$</td>
<td>Level difference in the rolling 30-day realized volatility of WTI physical futures 1-month nearby contract between weeks $t+8$ and $t$</td>
</tr>
<tr>
<td>xomRet$^8$</td>
<td>Exxon Mobil stock returns (in %) from week $t$ to week $t+8$</td>
</tr>
<tr>
<td>bpRet$^8$</td>
<td>British Petrol stock returns from week $t$ to week $t+8$</td>
</tr>
<tr>
<td>rdsaRet$^8$</td>
<td>Royal Dutch Shell class A stock returns from week $t$ to week $t+8$</td>
</tr>
<tr>
<td>DInv$^8$</td>
<td>Percent change in U.S. crude inventories including SPR (EOP, mil. bbl) from week $t$ to week $t+8$</td>
</tr>
<tr>
<td>DProd$^8$</td>
<td>Average weekly percent change in U.S. crude oil field production (mil. bbl/day) from week $t$ to week $t+8$</td>
</tr>
<tr>
<td>OilVol</td>
<td>Rolling 30-day realized volatility of WTI physical futures 1-month nearby contract</td>
</tr>
<tr>
<td>VIX</td>
<td>CBOE market volatility index</td>
</tr>
<tr>
<td>DFX</td>
<td>Percent change in the nominal broad dollar index - goods only (Jan 1997 = 100) relative to 4 weeks ago</td>
</tr>
<tr>
<td>tnote_10y</td>
<td>10-year treasury note yield at constant maturity (EOP, % p.a.)</td>
</tr>
<tr>
<td>sp500Ret</td>
<td>Standard and Poor’s 500 stock returns relative to 4 weeks ago</td>
</tr>
<tr>
<td>basis</td>
<td>WTI physical annualized 3-month to 1-month basis (when positive curve is upward sloping, capturing contango)</td>
</tr>
<tr>
<td>WIPImom</td>
<td>Month-over-month growth rate of Baumeister and Hamilton’s (2019) monthly World Industrial Production Index</td>
</tr>
<tr>
<td>trend</td>
<td>Weekly linear time trend</td>
</tr>
<tr>
<td>vix_spx</td>
<td>The difference between CBOE market volatility index and the 30-day volatility of Standard and Poor’s 500 index</td>
</tr>
<tr>
<td>ovx_cl1</td>
<td>The difference between CBOE crude oil volatility index and the 30-day volatility of WTI crude oil prices</td>
</tr>
<tr>
<td>sdf_fullSample</td>
<td>Risk premium calculated from annual covariance with full-sample stochastic discount factor</td>
</tr>
<tr>
<td>f[Topic]</td>
<td>Average frequency of articles over the previous 4 weeks in Topic</td>
</tr>
<tr>
<td>s[Topic]</td>
<td>Average sentiment over the previous 4 weeks due to Topic</td>
</tr>
<tr>
<td>artcount</td>
<td>Average number of articles in the energy corpus over the past 4 weeks</td>
</tr>
<tr>
<td>entropy</td>
<td>Average measure of article unusualness over the past 4 weeks</td>
</tr>
</tbody>
</table>

Table II
Descriptive Statistics

Data summary using weekly observations from April 1998 to March 2020. The variables labeled t8 show eight-week changes (the t8 is suppressed in other tables). The other non-text series are observed weekly, some as changes and some as levels, and the text variables are four-week averages of weekly observations. The data are observed on Tuesday for non-price series, and on Thursday for price-based series. For each variable, the table shows the mean, standard deviation, median, and the 5th and 95th percentiles. N is the number of observations in the sample. Variable definitions are presented in Table I. The text measures, which except entropy are standardized to mean zero and unit variance in the regressions, are not standardized here.

<table>
<thead>
<tr>
<th>VARIABLES</th>
<th>mean</th>
<th>sd</th>
<th>p5</th>
<th>p50</th>
<th>p95</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>FutRet_t8</td>
<td>1.349</td>
<td>13.78</td>
<td>-22.71</td>
<td>2.512</td>
<td>21.49</td>
<td>1,139</td>
</tr>
<tr>
<td>DSpot_t8</td>
<td>0.637</td>
<td>14.80</td>
<td>-25.91</td>
<td>2.876</td>
<td>20.31</td>
<td>1,077</td>
</tr>
<tr>
<td>DOilVol_t8</td>
<td>0.203</td>
<td>14.40</td>
<td>-22.72</td>
<td>-0.550</td>
<td>23.36</td>
<td>1,077</td>
</tr>
<tr>
<td>xomRet_t8</td>
<td>0.203</td>
<td>7.612</td>
<td>-11.87</td>
<td>0.545</td>
<td>11.52</td>
<td>1,099</td>
</tr>
<tr>
<td>bpRet_t8</td>
<td>-0.295</td>
<td>9.944</td>
<td>-15.24</td>
<td>0.359</td>
<td>13.00</td>
<td>1,099</td>
</tr>
<tr>
<td>rdsaret_t8</td>
<td>-0.351</td>
<td>9.420</td>
<td>-14.83</td>
<td>0.548</td>
<td>12.47</td>
<td>1,074</td>
</tr>
<tr>
<td>Dlnv_t8</td>
<td>0.137</td>
<td>1.742</td>
<td>-2.596</td>
<td>0.145</td>
<td>2.966</td>
<td>1,139</td>
</tr>
<tr>
<td>DProd_t8</td>
<td>0.271</td>
<td>2.618</td>
<td>-2.388</td>
<td>0.190</td>
<td>3.438</td>
<td>1,139</td>
</tr>
<tr>
<td>OilVol</td>
<td>35.82</td>
<td>15.95</td>
<td>17.50</td>
<td>32.52</td>
<td>66.49</td>
<td>1,136</td>
</tr>
<tr>
<td>VIX</td>
<td>20.05</td>
<td>8.826</td>
<td>11.21</td>
<td>17.97</td>
<td>35.79</td>
<td>1,146</td>
</tr>
<tr>
<td>DVIX</td>
<td>0.134</td>
<td>5.739</td>
<td>-6.200</td>
<td>-0.440</td>
<td>8.070</td>
<td>1,141</td>
</tr>
<tr>
<td>DFX</td>
<td>0.0544</td>
<td>1.507</td>
<td>-2.269</td>
<td>-0.0220</td>
<td>2.424</td>
<td>1,141</td>
</tr>
<tr>
<td>tnote_10y</td>
<td>3.567</td>
<td>1.327</td>
<td>1.700</td>
<td>3.580</td>
<td>5.850</td>
<td>1,147</td>
</tr>
<tr>
<td>sp500Ret</td>
<td>0.308</td>
<td>4.707</td>
<td>-7.545</td>
<td>1.007</td>
<td>6.077</td>
<td>1,141</td>
</tr>
<tr>
<td>WIPImom</td>
<td>0.2089</td>
<td>0.6022</td>
<td>-0.674</td>
<td>0.2604</td>
<td>1.0027</td>
<td>1,143</td>
</tr>
<tr>
<td>basis</td>
<td>0.0751</td>
<td>0.319</td>
<td>-0.256</td>
<td>0.0465</td>
<td>0.473</td>
<td>1,136</td>
</tr>
<tr>
<td>trend</td>
<td>574</td>
<td>331.3</td>
<td>58</td>
<td>574</td>
<td>1,090</td>
<td>1,147</td>
</tr>
<tr>
<td>vix_spx</td>
<td>3.235</td>
<td>4.566</td>
<td>-4.415</td>
<td>3.620</td>
<td>9.478</td>
<td>1,146</td>
</tr>
<tr>
<td>oxv_cl1</td>
<td>1.82</td>
<td>8.400</td>
<td>-14.778</td>
<td>3.070</td>
<td>12.642</td>
<td>673</td>
</tr>
<tr>
<td>sdf_fullSample</td>
<td>0.041</td>
<td>0.030</td>
<td>0.007</td>
<td>0.033</td>
<td>0.093</td>
<td>1,039</td>
</tr>
<tr>
<td>PCAsent</td>
<td>0</td>
<td>1.503</td>
<td>-2.179</td>
<td>0.278</td>
<td>2.183</td>
<td>1,143</td>
</tr>
<tr>
<td>PCAfreq</td>
<td>0</td>
<td>1.776</td>
<td>-2.159</td>
<td>-0.766</td>
<td>2.940</td>
<td>1,143</td>
</tr>
<tr>
<td>PCAAll</td>
<td>0</td>
<td>2.443</td>
<td>-2.950</td>
<td>-1.065</td>
<td>3.744</td>
<td>1,143</td>
</tr>
<tr>
<td>artcount</td>
<td>332.4</td>
<td>113.5</td>
<td>173.3</td>
<td>358.0</td>
<td>517.8</td>
<td>1,143</td>
</tr>
<tr>
<td>entropy</td>
<td>2.150</td>
<td>0.116</td>
<td>1.949</td>
<td>2.170</td>
<td>2.305</td>
<td>1,143</td>
</tr>
<tr>
<td>sCo</td>
<td>-0.00119</td>
<td>0.000341</td>
<td>-0.00181</td>
<td>-0.00110</td>
<td>-0.000772</td>
<td>1,143</td>
</tr>
<tr>
<td>jCo</td>
<td>0.127</td>
<td>0.0468</td>
<td>0.0757</td>
<td>0.120</td>
<td>0.219</td>
<td>1,143</td>
</tr>
<tr>
<td>sGom</td>
<td>-0.00471</td>
<td>0.00178</td>
<td>-0.00796</td>
<td>-0.00436</td>
<td>-0.00239</td>
<td>1,143</td>
</tr>
<tr>
<td>fGom</td>
<td>0.346</td>
<td>0.103</td>
<td>0.213</td>
<td>0.333</td>
<td>0.507</td>
<td>1,143</td>
</tr>
<tr>
<td>sEnv</td>
<td>-0.000561</td>
<td>0.000324</td>
<td>-0.00114</td>
<td>-0.000561</td>
<td>-0.000151</td>
<td>1,143</td>
</tr>
<tr>
<td>fEnv</td>
<td>0.0318</td>
<td>0.0172</td>
<td>0.00841</td>
<td>0.0333</td>
<td>0.0581</td>
<td>1,143</td>
</tr>
<tr>
<td>sEpg</td>
<td>-0.00564</td>
<td>0.00135</td>
<td>-0.00777</td>
<td>-0.00550</td>
<td>-0.00354</td>
<td>1,143</td>
</tr>
<tr>
<td>fEpg</td>
<td>0.355</td>
<td>0.0539</td>
<td>0.260</td>
<td>0.369</td>
<td>0.429</td>
<td>1,143</td>
</tr>
<tr>
<td>sBbl</td>
<td>-0.000429</td>
<td>0.000225</td>
<td>-0.000932</td>
<td>-0.000357</td>
<td>-0.000198</td>
<td>1,143</td>
</tr>
<tr>
<td>fBbl</td>
<td>0.0387</td>
<td>0.0158</td>
<td>0.0197</td>
<td>0.0345</td>
<td>0.0670</td>
<td>1,143</td>
</tr>
<tr>
<td>sRpg</td>
<td>-0.000341</td>
<td>0.000104</td>
<td>-0.000560</td>
<td>-0.000328</td>
<td>-0.000195</td>
<td>1,143</td>
</tr>
<tr>
<td>fRpg</td>
<td>0.0203</td>
<td>0.00433</td>
<td>0.0148</td>
<td>0.0194</td>
<td>0.0292</td>
<td>1,143</td>
</tr>
<tr>
<td>sEp</td>
<td>-0.000472</td>
<td>0.000192</td>
<td>-0.000751</td>
<td>-0.000447</td>
<td>-0.000234</td>
<td>1,143</td>
</tr>
<tr>
<td>fEp</td>
<td>0.0358</td>
<td>0.0115</td>
<td>0.0212</td>
<td>0.0340</td>
<td>0.0550</td>
<td>1,143</td>
</tr>
</tbody>
</table>
## Sample Sentences

This table shows headlines associated with the episodes marked with stars in Panels A and B of Figure 2. Each episode is labeled with its respective time frame. Articles for each episode must belong predominantly \((f_{i,r} > 0.8)\) to the episode’s topical category. For each event, the headlines of the five most negative sentiment articles are chosen from the candidate set, which consists of articles with an entropy higher than 2 and with a total number of words higher than 100. The **Sentiment** and **Entropy** columns correspond to article sentiment and entropy respectively.

<table>
<thead>
<tr>
<th>Sentiment</th>
<th>Entropy</th>
<th>Date</th>
<th>Headline</th>
</tr>
</thead>
<tbody>
<tr>
<td>-0.115</td>
<td>2.298</td>
<td>9/12/2000</td>
<td>UK's Blair to hold urgent talks over fuel crisis</td>
</tr>
<tr>
<td>-0.092</td>
<td>2.361</td>
<td>9/12/2000</td>
<td>EU asks Belgium for information on trucks protest</td>
</tr>
<tr>
<td>-0.072</td>
<td>2.395</td>
<td>9/13/2000</td>
<td>UPDATE 1-UK business says fuel crisis hurting</td>
</tr>
<tr>
<td>-0.069</td>
<td>2.347</td>
<td>9/13/2000</td>
<td>Fuel crisis costs UK firms 250 mln stg a day -LCC</td>
</tr>
<tr>
<td>-0.068</td>
<td>2.447</td>
<td>9/19/2000</td>
<td>EU govs to hold crisis talks far from Brussels</td>
</tr>
</tbody>
</table>

**Co:** Failed Venezuelan coup from 2002-03-27 to 2002-04-24
-0.132 2.483 4/12/2002 Venezuela PDVSA staff say oil exports being restored
-0.128 2.476 4/12/2002 Venezuela PDVSA staff say oil exports being restored
-0.111 2.44 4/11/2002 U.S. concerned about Venezuela, urges moderation
-0.102 2.403 4/5/2002 UPDATE 1-Oil protest grips Venezuela, disruptions reported
-0.097 2.504 4/12/2002 IPE Brent lower as Venezuela supply concerns ease

**Env:** Volkswagen emissions scandal from 2015-09-16 to 2015-10-14
-0.107 2.347 9/24/2015 Nidera says suffers significant loss from biofuels fraud
-0.09 2.364 9/23/2015 BRIEF-Fitch places Volkswagen AG on Rating Watch Negative
-0.09 2.312 10/2/2015 UPDATE 1-VW faces French inquiry for 'aggravated deception' in emissions scandal
-0.071 2.391 9/20/2015 UPDATE 1-Volkswagen orders investigation into breach of US environment rules
-0.063 2.431 9/21/2015 UPDATE 1-Volkswagen shares plunge on U.S. emissions scandal

**Ep:** Post-bankruptcy Enron hearings from 2002-01-16 to 2002-02-13
-0.131 2.372 2/12/2002 Calif senate panel seeks contempt citation vs. Enron
-0.123 2.33 2/6/2002 Enron skips Calif. hearing, may face contempt charges
-0.114 2.333 2/4/2002 UPDATE 1-Global Crossing says panel to probe accounting
-0.108 2.312 2/8/2002 Court seen for Enron bigwigs as Congress probes
-0.095 2.34 1/23/2002 Calif. court orders Enron to save documents

**Bbl:** Hurricane Katrina from 2005-08-24 to 2005-09-21
-0.075 2.367 9/12/2005 UPDATE 1-FEMA chief Brown resigns in wake of Katrina
-0.059 2.342 9/12/2005 FEMA revises Brown's bio after exaggeration charges
-0.057 2.268 9/2/2005 Bush signs $10.5 bln spending bill for Katrina
-0.055 2.331 9/13/2005 U.S. lawmaker won't reopen bankruptcy for Katrina
-0.055 2.349 8/31/2005 UPDATE 1-Bush says will take years to recover from Katrina

**Ep:** BP oil spill aftermath from 2010-05-05 to 2010-06-02
-0.078 2.12 5/6/2010 UPDATE 1-Pioneer Drilling Q1 loss wider than expected
-0.072 2.488 6/1/2010 UPDTAE 1-Goldman removes Halliburton from conviction buy list
-0.061 2.367 5/27/2010 UPDATE 1-Carrefour, unions reach Belgian restructuring deal
-0.058 2.583 6/1/2010 Transocean, Halliburton credit default swaps surge
-0.057 2.32 5/13/2010 UPDATE 1-Transocean seeks to limit spill liability
Table IV

Stepwise Forward Selection at the Eight-Week Horizon

The table shows the regression results for all 8 dependent variables at the eight-week horizon using stepwise forward selection to choose 7 of all the variables described in Table I, except ovx_cl1 and sdf_fullSample. Only predictors that were chosen by at least one model are included in this table. Coefficients are standardized. Superscripts before coefficients indicate order in forward selection (1=chosen first). The p-values in parentheses are obtained using Monte Carlo simulations that use an AR8 process to simulate the LHS variable, as well as forward selection to produce both adjusted R^2 and t-statistic simulations. The p-values refer to the minimum of the fraction of simulated t-statistics less than the empirical t-statistic, and 1 minus the fraction of simulated t-statistics less than the empirical t-statistic, where the comparison is relative to the order in which the variables were chosen. The bootstrap was repeated 1,000 times. The table also reports the mean of simulated adjusted R^2 resulting from the same bootstrap, as well as the corresponding CDF percentage, computed as the percent of adjusted R^2 simulations less than the empirical adjusted R^2. Statistically significance shown in bold.

<table>
<thead>
<tr>
<th>Predictors</th>
<th>FutRet</th>
<th>Dspot</th>
<th>DOilVol</th>
<th>xomRet</th>
<th>bpRet</th>
<th>rdsaRet</th>
<th>DInv</th>
<th>DProd</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>coef</td>
<td>pval</td>
<td>coef</td>
<td>pval</td>
<td>coef</td>
<td>pval</td>
<td>coef</td>
<td>pval</td>
</tr>
<tr>
<td>DSpot</td>
<td>-0.33</td>
<td>0.00</td>
<td></td>
<td></td>
<td>0.11</td>
<td>0.16</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DOilVol</td>
<td>-0.16</td>
<td>0.06</td>
<td>0.13</td>
<td>0.10</td>
<td>0.13</td>
<td>0.10</td>
<td>0.12</td>
<td>0.05</td>
</tr>
<tr>
<td>OilVol</td>
<td>-0.10</td>
<td>0.39</td>
<td>0.71</td>
<td>0.00</td>
<td>0.10</td>
<td>0.01</td>
<td>0.12</td>
<td>0.01</td>
</tr>
<tr>
<td>DInv</td>
<td>0.10</td>
<td>0.39</td>
<td>0.13</td>
<td>0.10</td>
<td>0.13</td>
<td>0.10</td>
<td>0.12</td>
<td>0.01</td>
</tr>
<tr>
<td>DProd</td>
<td>0.16</td>
<td>0.06</td>
<td>0.10</td>
<td>0.01</td>
<td>0.12</td>
<td>0.05</td>
<td>0.15</td>
<td>0.04</td>
</tr>
<tr>
<td>tote_10y</td>
<td>0.16</td>
<td>0.03</td>
<td>0.13</td>
<td>0.00</td>
<td>0.13</td>
<td>0.05</td>
<td>0.07</td>
<td>0.13</td>
</tr>
<tr>
<td>DFX</td>
<td>0.14</td>
<td>0.07</td>
<td>0.10</td>
<td>0.09</td>
<td>0.12</td>
<td>0.01</td>
<td>0.19</td>
<td>0.10</td>
</tr>
<tr>
<td>sp500Ret</td>
<td>0.14</td>
<td>0.07</td>
<td>0.10</td>
<td>0.09</td>
<td>0.12</td>
<td>0.01</td>
<td>0.19</td>
<td>0.10</td>
</tr>
<tr>
<td>basis</td>
<td>0.10</td>
<td>0.32</td>
<td>0.13</td>
<td>0.02</td>
<td>0.13</td>
<td>0.02</td>
<td>0.19</td>
<td>0.01</td>
</tr>
<tr>
<td>WIPImom</td>
<td>0.27</td>
<td>0.00</td>
<td>0.18</td>
<td>0.05</td>
<td>0.12</td>
<td>0.01</td>
<td>0.18</td>
<td>0.01</td>
</tr>
<tr>
<td>PCAsent</td>
<td>0.12</td>
<td>0.01</td>
<td>0.12</td>
<td>0.01</td>
<td>0.12</td>
<td>0.01</td>
<td>0.18</td>
<td>0.01</td>
</tr>
<tr>
<td>artcount</td>
<td>0.10</td>
<td>0.07</td>
<td>0.34</td>
<td>0.00</td>
<td>0.17</td>
<td>0.00</td>
<td>0.16</td>
<td>0.03</td>
</tr>
<tr>
<td>entropy</td>
<td>0.10</td>
<td>0.07</td>
<td>0.34</td>
<td>0.00</td>
<td>0.17</td>
<td>0.00</td>
<td>0.16</td>
<td>0.03</td>
</tr>
<tr>
<td>fCo</td>
<td>0.16</td>
<td>0.03</td>
<td>0.10</td>
<td>0.32</td>
<td>0.13</td>
<td>0.02</td>
<td>0.19</td>
<td>0.01</td>
</tr>
<tr>
<td>sGom</td>
<td>0.10</td>
<td>0.02</td>
<td>0.13</td>
<td>0.08</td>
<td>0.13</td>
<td>0.02</td>
<td>0.25</td>
<td>0.01</td>
</tr>
<tr>
<td>fGom</td>
<td>0.16</td>
<td>0.03</td>
<td>0.10</td>
<td>0.32</td>
<td>0.13</td>
<td>0.02</td>
<td>0.19</td>
<td>0.01</td>
</tr>
<tr>
<td>sEnv</td>
<td>0.10</td>
<td>0.02</td>
<td>0.13</td>
<td>0.08</td>
<td>0.13</td>
<td>0.02</td>
<td>0.25</td>
<td>0.01</td>
</tr>
<tr>
<td>fBbl</td>
<td>0.10</td>
<td>0.17</td>
<td>0.10</td>
<td>0.32</td>
<td>0.13</td>
<td>0.02</td>
<td>0.19</td>
<td>0.01</td>
</tr>
<tr>
<td>sRpc</td>
<td>0.11</td>
<td>0.05</td>
<td>0.11</td>
<td>0.05</td>
<td>0.11</td>
<td>0.05</td>
<td>0.14</td>
<td>0.02</td>
</tr>
<tr>
<td>fRpc</td>
<td>0.11</td>
<td>0.05</td>
<td>0.11</td>
<td>0.05</td>
<td>0.11</td>
<td>0.05</td>
<td>0.14</td>
<td>0.02</td>
</tr>
<tr>
<td>sEp</td>
<td>0.33</td>
<td>0.00</td>
<td>0.33</td>
<td>0.00</td>
<td>0.15</td>
<td>0.19</td>
<td>0.15</td>
<td>0.07</td>
</tr>
<tr>
<td>fEp</td>
<td>0.11</td>
<td>0.09</td>
<td>0.11</td>
<td>0.09</td>
<td>0.11</td>
<td>0.09</td>
<td>0.11</td>
<td>0.09</td>
</tr>
</tbody>
</table>

Observations: 1110 1063 1063 1075 1075 1050 1119 1119
R^2 / R^2 adjusted: 0.155 / 0.149 0.176 / 0.171 0.369 / 0.365 0.254 / 0.249 0.236 / 0.231 0.132 / 0.126 0.154 / 0.149 0.169 / 0.164
Mean of sim. Adj. R2: 0.071 0.086 0.062 0.058 0.053 0.062 0.066 0.049
CDF (%): 99.6 99.7 100 100 99.6 100 100
The table displays MSE ratios of the in-sample based Forward Selection model against the benchmark Constant model. The in-sample Forward Selection provides 7 or 2 top predictors sequentially for each predicted variable. During this process, the lagged variables are regressed out to control for autocorrelation and are added back in the prediction phase. The RHS independent variables of the two models are thus the lagged predicted variable along with forward selected 7 or 2 top candidates, which are labeled as FW7 and FW2 respectively in the 1st column. Each of the remaining columns corresponds to a predicted variable and entries show the MSE ratios relative to the constant models. The out-of-sample prediction is handled weekly, by first re-evaluating the factor loadings based on a Lasso regression on a rolling 5-year lookback window, and forecasting eight-week ahead thereafter. The penalty parameter in the Lasso regressions are determined by a grid search algorithm with 10 cross validation using mean squared errors as the criterion. Since the regression lagged the explanatory variables by eight weeks, those predictor series in the last eight-week window in the lookback window are excluded from the coefficient updating practice by definition. Furthermore, the forecast is only founded on the last observation in the window instead of that eight-week window to ensure genuine out-of-sample prediction. The constant models take the average in the lookback window as the prediction. MSE of each tested model and the constant model is calculated after all the predictions are accomplished. The MSE ratios are produced by dividing the MSE of the forecasting model and that of the constant model. Boldface indicates better performance than the benchmark model.

<table>
<thead>
<tr>
<th>Model</th>
<th>FutRet</th>
<th>xomRet</th>
<th>bpRet</th>
<th>rdsaRet</th>
<th>DSpot</th>
<th>DOilVol</th>
<th>DInv</th>
<th>DProd</th>
</tr>
</thead>
<tbody>
<tr>
<td>FW2</td>
<td>1.139</td>
<td><strong>0.999</strong></td>
<td><strong>0.995</strong></td>
<td>1.055</td>
<td>1.077</td>
<td><strong>0.856</strong></td>
<td>1.008</td>
<td>1.021</td>
</tr>
<tr>
<td>FW7</td>
<td>1.299</td>
<td>1.025</td>
<td>1.047</td>
<td>1.097</td>
<td>1.190</td>
<td><strong>0.867</strong></td>
<td>1.114</td>
<td>1.089</td>
</tr>
</tbody>
</table>
Table VI
Forecast Accuracy of Out-of-Sample 1-1 and 2-2 R² Selected Lasso Updating model

The table displays the MSE ratio of the out-of-sample 1-1 and 2-2 Lasso updating model against benchmark models. The model selects n-n predictor pairs weekly (n=1 or 2 according to specification) from the base and text variable pool separately with the top n R² in the univariate OLS regression based on a 5-year lookback window. The univariate OLS regression takes the predicted variable as LHS and a lagged base or text candidate as RHS. There are two benchmark models: the Constant model and the Baseline model, with specification annotated in the table as \( 0 \text{ base} + 0 \text{ text} \) or \( n \text{ base} + n \text{ text} \) respectively. Alternative models are the Text model and the Full model, with specifications as \( 0 \text{ base} + n \text{ text} \) or \( n \text{ base} + n \text{ text} \). The model then update the coefficients of the predictors with Lasso regression on a rolling 5-year lookback window and predicts eight-week ahead using the last observation in the window. The MSE for a model are produced once all the weekly forecasts are accomplished. After the MSE calculations, the MSE ratio is determined by dividing the MSE of an alternative model on that of a benchmark model. Boldface indicates better performance of the alternative than the benchmark.

Panel A: Full Model specification: 1 base + 1 text

<table>
<thead>
<tr>
<th>Benchmark Model</th>
<th>Constant (0 \text{ base} + 0 \text{ text})</th>
<th>Baseline (1 \text{ base} + 0 \text{ text})</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alternative Model</td>
<td>Text (0 \text{ base} + 1 \text{ text})</td>
<td>Full (1 \text{ base} + 1 \text{ text})</td>
</tr>
<tr>
<td>FutRet</td>
<td>1.115</td>
<td>1.179</td>
</tr>
<tr>
<td>xomRet</td>
<td>1.047</td>
<td>1.109</td>
</tr>
<tr>
<td>bpRet</td>
<td>1.136</td>
<td>1.238</td>
</tr>
<tr>
<td>rdsaRet</td>
<td>1.092</td>
<td>1.084</td>
</tr>
<tr>
<td>DSpot</td>
<td>1.079</td>
<td>1.108</td>
</tr>
<tr>
<td>DOilVol</td>
<td>1.061</td>
<td>\textbf{0.970}</td>
</tr>
<tr>
<td>DInv</td>
<td>1.182</td>
<td>1.226</td>
</tr>
<tr>
<td>DProd</td>
<td>\textbf{0.986}</td>
<td>1.039</td>
</tr>
</tbody>
</table>

Panel B: Full Model specification: 2 base + 2 text

<table>
<thead>
<tr>
<th>Benchmark Model</th>
<th>Constant (0 \text{ base} + 0 \text{ text})</th>
<th>Baseline (2 \text{ base} + 0 \text{ text})</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alternative Model</td>
<td>Text (0 \text{ base} + 2 \text{ text})</td>
<td>Full (2 \text{ base} + 2 \text{ text})</td>
</tr>
<tr>
<td>FutRet</td>
<td>1.082</td>
<td>1.210</td>
</tr>
<tr>
<td>xomRet</td>
<td>1.054</td>
<td>1.100</td>
</tr>
<tr>
<td>bpRet</td>
<td>1.069</td>
<td>1.270</td>
</tr>
<tr>
<td>Variable</td>
<td>dsaRet</td>
<td>DSot</td>
</tr>
<tr>
<td>---------------</td>
<td>--------</td>
<td>------</td>
</tr>
<tr>
<td>dsaRet</td>
<td>1.112</td>
<td>1.109</td>
</tr>
<tr>
<td>DSot</td>
<td>1.109</td>
<td>1.191</td>
</tr>
<tr>
<td>DOilVol</td>
<td>1.118</td>
<td>0.988</td>
</tr>
<tr>
<td>DInv</td>
<td>1.154</td>
<td>1.225</td>
</tr>
<tr>
<td>DProd</td>
<td>1.024</td>
<td>1.108</td>
</tr>
</tbody>
</table>
Table VII
Summary Statistics of the MSE ratios of the Successful Candidates in the Out-of-Sample Fixed Model

This table shows the summary statistics of the distribution of the MSE ratios of the winning models against the constant model in the out-of-sample fixed model. There are 8 rows with each corresponding to a predicted variable. The columns indicates the 5, 25, 50 (median), 75, 95 percentiles, mean, and standard deviations of the victorious MSE ratios in the out-of-sample fixed model forecast accuracy test; the column labeled N reports the total number of the desired (i.e. winning) models in each practice. The fixed model selects a pair of variable and sticks to them through the whole sample. The predictions are made weekly by updating the coefficients using a rolling 5-year lookback Lasso regression, and the MSE is calculated once all forecasts are done. The MSE ratios divides the MSE of the fixed model by that of the constant model, which predicts weekly as the average of predicted value in a rolling 5-year lookback window. A model is considered advantageous if the MSE ratio is less than 1. For each predicted variable, there are 741 fixed models at disposal (all the non-overlapping combinations of the 39 variables with 19 base and 20 text ones).

<table>
<thead>
<tr>
<th>Predicted Variables</th>
<th>Summary Statistics of MSE ratios for Successful Models</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>p5</td>
</tr>
<tr>
<td>FutRet</td>
<td>0.9510</td>
</tr>
<tr>
<td>xomRet</td>
<td>0.9798</td>
</tr>
<tr>
<td>bpRet</td>
<td>0.9776</td>
</tr>
<tr>
<td>rdsaRet</td>
<td>0.9803</td>
</tr>
<tr>
<td>DSpot</td>
<td>0.9645</td>
</tr>
<tr>
<td>DOilVol</td>
<td>0.6726</td>
</tr>
<tr>
<td>DInv</td>
<td>0.9589</td>
</tr>
<tr>
<td>DProd</td>
<td>0.9406</td>
</tr>
</tbody>
</table>
Figure 1. Word cloud plots for energy topics. This figure shows the word clouds of the energy topics extracted from the energy corpus using the Louvain clustering algorithm. Larger font indicates words that occur more frequently in a given cluster.
Panel A: Topical Frequency

- Company (Co)
- Global Oil Markets (Gom)
- Environment (Env)
- Energy/Power Generation (Epg)
- Crude Oil Physical (Bbl)
- Refining & Petrochemicals (Rpc)
- Exploration & Production (Ep)
Figure 2: NLP measures over time. This figure shows the time series plots of all the textual series in this paper. All series start from April 1998 and end in March 2020. We display the 4-week averages of topical frequencies in Panel A, and 4-week averages of topical sentiments in Panel B. The stars in Panels A and B mark the events detailed in Table III. The stars are positioned on the ending date of the time-period associated with the Table III episodes. In addition, Panel C shows 4-week averages of the article counts, the unusualness (entropy) and the first principal components of normalized 4-week average textual measures.
variables listed in Table I as candidate variables, in addition to the [vix_spx] risk premium. After this
Figure 15. Oil Futures Returns (8-week horizon) Oil Volatility (8–week horizon)

Note: the adjusted R-squared value in the baseline model is 14.94%. Note: the adjusted R-squared value in the baseline model is 36.47%.

Figure 3. Monte Carlo simulations of adjusted $R^2$ for the eight-week oil futures returns and eight-week difference in oil volatility models. The bootstrap uses forward selection to choose 7 variables after all variables have been de-trended, the lagged dependent variable has been regressed out of all variables, and the lagged dependent variable dropped from consideration. The forward selection process includes all variables listed in Table I as candidate variables, in addition to the [vix_spx] risk premium. After this adjustment, RHS variables are used as they are in the simulation, while the LHS is simulated using an AR8 process. The figure shows the p-value, which is the percent of simulated adjusted $R^2$ less than the empirical adjusted $R^2$. The difference shown in the legend refers to the difference between the empirical adjusted $R^2$ and the mean of adjusted $R^2$ simulations. The appendix presents a detailed overview of the bootstrap process.
Figure 4. Monte Carlo simulations of t-statistics for the 7 variables chosen via forward selection for the eight-week oil futures returns and eight-week difference in oil volatility models. Following the same bootstrap process outlined in the Appendix and used to produce Figure 3, we keep track of the t-statistics for the simulated regression results. The p-value is computed as the minimum of the percent of simulated t-statistics less than the empirical t-statistic, and 1 minus the percent of simulated t-statistics less than the empirical t-statistic. In computing the p-values, we preserve the order of variables chosen in the empirical and bootstrap processes, and compare the t-statistics in that order. The empirical t-statistics of the variables chosen via forward selection, in the order in which they were chosen, as listed in the notes below the figures. The p-values presented in Table IV are derived from this process for all variables.
Figure 5. Number of times that all possible 2-variable models outperform the constant model for out-of-sample forecasting. The figure shows the winning counts of the 2-variable fixed models against the constant model under MSE criterion in the prediction of the dependent variables. Each cell represents the number of left-hand side variables for which the corresponding fixed model, consisting of an explanatory variable from the indicated row and column of the matrix, beats the constant model. Darker color indicates greater number. The 2-variable fixed model fixes 2 predictors ex-ante and forecasts eight-week ahead weekly based on the update coefficients by applying Lasso regression on a 5-yr lookback window. All the possible combinations of 19 baseline and 20 textual variables are considered, rendering a model pool with 741 candidates for each predicted variable. The left and top edges display variable names starting by the 20 textual variables and then the 19 baseline variables. The matrix is symmetric and has zero diagonal by construction. The bottom and right edges show the sum of the labeling column or row respectively, and each number denotes the total time a variable enters the desired fixed models. Two boxes between the matrix and the color bar reveal the aggregated entry counts of textual and baseline variables accordingly.
Appendix

5.1. Bootstrap analysis

For our in-sample analysis, we need to control for two deviations from standard econometric assumptions. First, we are likely to have serial correlation in the residuals of our time-series regressions because of the use of overlapping observations. Furthermore, we employ forward selection for choosing a parsimonious set of in-sample regressors. Both of these considerations may introduce upward bias in the R-squareds, and downward bias in the standard errors. To control for both of these sources of finite sample bias, we bootstrap the data and construct bootstrapped distributions for our t-statistics and R-squareds.

Before we begin the bootstrap, we detrend all dependent and forecasting variables. For each dependent variable, we then residualize all series by regressing out the four-week version of the lagged dependent variable. Our in-sample analysis assumes the following specification for the detrended and residualized series:

\[ y_{t:t+h} = X_t^{(M)} \beta + \epsilon_{t:t+h}, \]  

(3)

where \( X_t^{(M)} \), \( \beta \in R^M \), \( M \) is the number of chosen explanatory variables, and the time index \( t \) is in weeks. We assume the \( X_t^{(M)} \)'s are chosen from the larger set \( X_t \) of \( N > M \) variables using the forward selection algorithm. Under the null we assume that \( \beta = 0 \) for all subsets of \( M \) variables. To match the empirical properties of the data, we estimate an \( AR(K) \) model for the dependent variable as follows:

\[ y_{t:t+h} = b_1 y_{t-1:t-1+h} + \cdots + b_K y_{t-K:t-K+h} + \epsilon_{t:t+h}. \]  

(4)

We run the analysis with \( K = h \), i.e. eight lags for eight-week ahead forecasts, and four lags for four-week ahead forecasts. We estimate the above model to get the empirical \( \hat{b}_1, \ldots, \hat{b}_K \) and the
innovation variance $\bar{\nu} r(e)$; these are the parameters that describe the behavior of the actual data. We then use this to calculate a single run of the simulation as follows:

1. Set $y_{1:h}, \ldots, y_{K:K+h} = 0$

2. Draw $e_{K+1:K+1+h}$ from a normal distribution with mean zero and variance $\bar{\nu} r(e)$.

3. Use the above relationship to generate the next element $y_{K+1:K+1+h}$ in (4).

4. Run the model for 100 steps as a burn-in period.

5. On the 101st step of the model, collect the $y$ variables until we match the number of empirical observations.

6. Run the forward selection algorithm using the simulated $y$’s and the detrended and residualized $X$’s. This selects a subset $X(M)$ of explanatory variables.

7. Keep track of the adjusted R-squared of this simulation run.

8. Keep track of the standard (no adjustments) OLS t-statistic for each of the variables that are selected by the forward-selection algorithm. In every simulated path this will result in $M$ t-statistics, $\{\hat{t}_1, \ldots, \hat{t}_M\}$. Here $\hat{t}_1$ correspond to the t-statistic of the first variable chosen by the forward selection algorithm, $\hat{t}_2$ is the t-statistics of the second chosen variable, and so on. We refer to these as the ordered t-statistics.

We repeat this procedure 1,000 times to generate a distribution for the observed R-squareds and the observed t-statistics. The simulated R-squareds and ordered t-statistics adjust for overlapping observations and variable selection under the null hypothesis of no relationship between the dependent and the independent variables.

We evaluate the adjusted R-squared for a given dependent variable via the percentage of simulated adjusted R-squareds that are lower. Since this is a one-sided test, a value of above 95% indicates significance at the 5% level. For p-values of the coefficient estimates in the actual
regression, we compare the t-statistic of the $n^{th}$ chosen variable in the forward selection method to the $n^{th}$ ordered t-statistic distribution under the null hypothesis. We report the outcome of the two-sided test $\min(\hat{p}, 1 - \hat{p})$ where $\hat{p}$ is the number of simulated t-statistics for the $n^{th}$ chosen variable that are less than the t-statistic for the actual $n^{th}$ chosen variable. For purposes of this test, all t-statistics are calculated using standard OLS assumptions of independence and homoscedasticity. The simulated t-statistic distributions will reflect all of the OLS biases.