Beyond News Headlines and TF-IDF: Enhancing Text-Based Forecasting Models With Validated Collocations and Improved Attention*

Gabriel Appau Abeyie

Department of Economics and Finance, University of North Carolina at Wilmington, 601 S. College Rd., Wilmington, NC, 28403, USA

Abstract

This paper proposes a method for improving text-based forecasting models, specifically focusing on forecasting crude oil prices. Utilizing advanced techniques, including pattern validation and attention mechanisms, the study demonstrates notable improvements in predictive power over traditional approaches. One key finding is that considering the full text of news articles, rather than limiting the analysis to news headlines, leads to significant gains in forecasting accuracy. Furthermore, the model featuring verb-noun and noun-verb collocation pattern validation consistently outperforms benchmarks and models based solely on news headlines across various forecasting horizons. The results suggest that the presence of such collocations as 'price fell,' 'prices tumbled,' and 'price dropped' in crude-oil-related news articles is associated with a decrease in oil price returns. Additionally, integrating macroeconomic data with text-based features enhances predictive performance, demonstrating that combining structured economic indicators with textual features improves forecasting accuracy.

Keywords: Attention mechanism, Crude oil prices, Macroeconomic data, Part-of-speech, Pattern validation

1. Introduction

The use of text as data for forecasting and decision-making has surged in popularity in recent years for many domains. This trend can be attributed to the growing availability of text data, ranging from social media and newspapers to extensive documents, and is further facilitated by the decreasing cost of computation. Businesses are capitalizing on this trend, utilizing text from online customer satisfaction surveys and product reviews to inform decision-making and drive business improvement. For example, Schneider and Gupta (2016) use consumer reviews from Amazon.com to forecast sales for both existing and newly-introduced products. In finance, text data sourced from financial news, social media, and company documents are employed to forecast movements in asset prices (An et al. (2023); Jiao et al. (2022); Bai et al. (2022); Lei et al. (2021); Zhang et al. (2021); Li et al. (2020); Li et al. (2019)). In economics, text from economic news articles and central bank-related documents are utilized for forecasting macroeconomic variables (Aprigliano et al. (2023); Ellingsen et al. (2022); Kalamara et al. (2022); Ochs (2021); Handlan (2020); Bailliu et al. (2019); Song and Shin (2019); Nowak and Smith (2017)).

Many of these influential studies tackled the challenge of high dimensionality by limiting their investigation to words found in a time-invariant pre-established dictionary (set of words). While this approach simplfies the analysis, it introduces some limitations. First, by relying on a static dictionary, researchers may overlook emerging terminologies or phrases that become relevant over time. For example, terms imbued with recessionary implications might be sidelined during boom times, only to become crucial during economic downturns. Secondly, the sentiment or significance attached to certain words can evolve, and a fixed dictionary does not account for these nuances. This might lead to misinterpretations or oversights in the analysis. For instance, the term 'fracking' was once primarily technical but has gained broad economic and environmental implications, altering its sentiment significantly in discussions related to energy sectors. Similarly, 'oil glut' might transition from a neutral to a negative connotation depending on market conditions.[1] Recent studies by Lima et al. (2021); Lima and Godeiro (2023) have demonstrated that substantial gains are achieved when employing a time-varying dictionary, underscoring the importance of adaptability in textual analysis. While these studies have made significant strides in advocating for the use of a time-varying dictionary, there is a need to ensure that the words, specifically collocations, considered are contextually relevant and interpretable. In Aruoba and Drechsel (2022); Calomiris et al. (2021), the researchers manually sifted through words and collocations, selecting those that carry an economic interpretation based on their own discretion. While this approach may capture the nuances of the data, it is inherently subjective, potentially introducing biases and hindering replicability. In this paper, I leverage Part-of-Speech (POS) tagging to address this problem.

*The numerical results gathered in this manuscript were reproduced by the Editor-in-Chief on 13 September 2025.

Email address: abeyieg@uncw.edu (Gabriel Appau Abeyie)

[1] Example terms such as 'fracking' and 'oil glut' illustrate how specific words related to the crude oil market can vary in sentiment and significance over time, reflecting changes in industry practices and market conditions.

An important aspect of textual analysis in forecasting is the role of Part-of-Speech (POS) tagging. POS tagging classifies each word in the text into a specific grammatical category, such as an adjective, adverb, noun, or verb among others. These grammatical categories have been shown to possess different strengths in sentiment analysis which can be harnessed for forecasting. For example, while verbs and adjectives are critical to identifying sentiments, adverbs modify them, providing deeper sentiment nuances. Nouns, on the other hand, are central to identifying the topic under discussion (Benamara et al. (2007); Nicholls and Song (2009); Khong et al. (2018); Yadav et al. (2020)).

The syntactic structure offered by POS tagging can be leveraged for more sophisticated text-based analyses (Manning (2011)). A notable application is in validating collocations or combinations of adjacent words to ensure their semantic or syntactic relevance (Justeson and Katz (1995)). By emphasizing specific POS combinations, such as adjective-noun or verb-noun pairs, researchers can discern nuanced meanings and directional sentiments. This, in turn, enhances the predictive capabilities of their models.

To address the challenges of high dimensionality, time-invariant dictionaries, and the need for contextually relevant and interpretable collocations, this research introduces a collocation pattern validation (CPV) stage into the text processing pipeline. This step not only streamlines the dimensionality of the data but also markedly enhances the replicability and interpretability of the results. Crucially, by systematically validating the relevance and context of words and collocations, this approach guarantees a more robust and improved forecasting performance.

In this paper, I conduct a text-based crude oil price forecasting exercise. I combine a CPV process with the attention mechanism[2] from Lima and Godeiro (2023), allowing the dictionary to vary over time in the forecasting exercise. The goal of integrating these two processes is to maximize the predictive power of the dictionary while ensuring that only contextually relevant and interpretable collocations are considered. The methodology can be summarized in four steps:

1. **Numerical Representation of Words**: The words (or terms) from the news articles are converted into numerical values to generate time series data. This transformation does not rely on a pre-specified, fixed dictionary. Given the high dimensionality and sparsity of this representation, a subsequent step focuses on dimensionality reduction.

2. **Pattern Validation of Collocations**: In this step, all extracted collocations are assessed and those that conform to specific part-of-speech patterns are selected. I utilize the Justeson and Katz (1995) pattern validation method, a well-regarded POS-based method in the linguistic literature. In addition, I introduce a novel Verb-Noun/Noun-Verb pattern validation approach. This filtering process not only reduces the data's dimensionality but also ensures that the collocations included in the analysis possess meaningful relationships and are contextually pertinent.

3. **Supervised Machine Learning for Predictor Construction**: Employing supervised machine learning (SML), I identify and select the time series (words) with the highest predictive power. From these selected words, I then derive new predictors. These predictors are obtained through feature extraction, specifically using principal component analysis (PCA).

4. **Out-of-Sample Forecasting**: With the newly formulated predictors, I proceed to generate out-of-sample (OOS) forecasts for the outcome variable.

Note that this four-stage process is recursively executed until the sample's conclusion. This recursive process results in a dynamic evolution in the content of the dictionary, highlighting its adaptability and responsiveness to changes over time.

The choice of PCA for feature extraction in this study is primarily due to its computational efficiency, which is especially advantageous given the recursive nature of our forecasting exercise. Alternative methods such as Latent Dirichlet Allocation (LDA) and Non-negative Matrix Factorization (NMF) are less suitable due to their high computational demands. Thorsrud (2020) points out that LDA struggles with maintaining topic consistency across updates, a significant issue for analyses that depend on continuous topics. Similarly, NMF, which is effective for shorter texts and smaller datasets, faces the same computational challenges.

This paper contributes to the literature by investigating the benefits of employing pattern validation and attention mechanism in text-based forecasting for crude oil prices. While the attention mechanism enables a time-varying dictionary and allows the model to focus on pertinent information within the text, pattern validation ensures that the extracted features are contextually relevant and meaningful. By comparing these advanced models to traditional benchmarks, the study assesses their relative predictive power.

Much of the research in text-based forecasting literature has considered news headlines rather than the full text of news articles for their forecasting exercises (Wu et al. (2021); Semiromi et al. (2020)). This paper contributes to the literature by investigating the benefits of considering the full text of news articles as opposed to just the news headlines. The aim is to determine whether there is a significant difference in forecast performance between using news headlines and the full text of news articles.

All codes and supplementary materials utilized in this research are publicly accessible on GitHub[3]. This underscores the commitment to transparent and reproducible research.

The rest of this paper is organized as follows: Section 2 describes the text preprocessing procedure used to convert crude

[2]Note that this attention mechanism is not related to the self-attention mechanism used in transformer models (e.g., Vaswani et al. (2017)).

[3]All codes and supplementary materials are available on Github. However, due to ProQuest's distribution restrictions, the text data cannot be shared directly. I have included detailed instructions within the repository on how to access and process the data from ProQuest for research purposes.

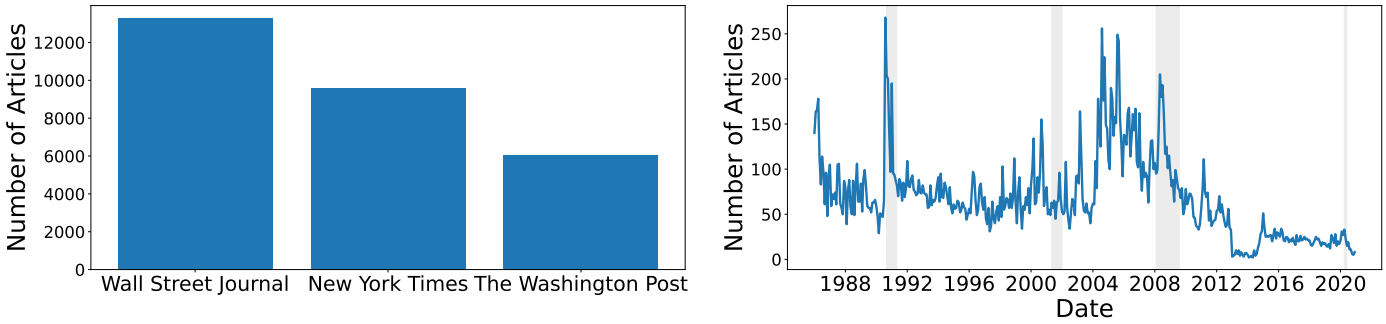

Figure 1: The bar chart (left) shows the number of articles published by the WSJ, NYT, and the WP within the corpora. The time series graph (right) shows the monthly number of articles related to crude oil prices from January 1986 to December 2020 within each corpus. The shaded regions represents NBER-defined recessions.

Table 1: Average Word Counts per Article and Headline Across Newspapers

Sources	Mean Article Length (Words)	Mean Headline Length (Words)
New York Times	798.23	7.27
Wall Street Journal	778.79	7.74
Washington Post	642.51	5.53

oil news articles into time-series features. Section 3 presents the methodological framework for extracting the most predictive textual features. Section 4 introduces the dataset and target variable. Section 5 outlines the forecast evaluation strategy and benchmark models. Section 6 reports the empirical results. Section 7 discusses robustness checks. Section 8 concludes.

2. Text Preprocessing

The corpus comprises a collection of monthly historical news articles and their corresponding headlines related to crude oil prices. The analysis is conducted at the monthly level to produce a coherent and structured representation of news content over time. Articles were sourced from the ProQuest repository using the search term *"crude oil price"*, chosen for its direct relevance to the study's focus. To ensure reputable coverage of economic and financial news, the corpus is restricted to three prominent publications: The New York Times (NYT), The Wall Street Journal (WSJ), and The Washington Post (WP), all of which have historically provided influential reporting on crude oil markets. This targeted approach yielded a substantial corpus of 28,904 newspaper articles and headlines spanning January 1986 to December 2020. Of these, 13,287 (45.97%) are from the WSJ, 9,573 (33.12%) from the NYT, and 6,044 (20.91%) from the WP. Figure 1 shows the distribution of crude-oil-price-related articles across the three newspapers and the monthly number of articles over the sample period. Table 1 reports the average word counts per article and per headline by newspaper source. Articles average between 640 and 800 words, while headlines average between five and eight words, highlighting the sharp contrast in length between full-text content and headlines.

2.1. Text Cleaning

The first step in text preprocessing is cleaning the corpora, which involves removing HTML tags, URLs, email addresses, numbers, short words, accented characters, hyphens, apostrophes, and stopwords (e.g., the, and, in). Unlike standard practice, I do not apply stemming or lemmatization, since retaining full word forms preserves semantic meaning and part-of-speech information. After conversion to lowercase, the text is tokenized into unigrams (single words).

Figure 2 shows a time series plot of the number of words in articles and headlines within each monthly corpus for the sample period.

2.2. Parts-of-Speech Tagging

The next stage in our text analysis pipeline involves categorizing tokens according to their parts of speech (POS). POS tagging clarifies the syntactic structure of sentences within each monthly corpus and highlights the linguistic style of the news articles and headlines. I use the `spaCy` package in Python,[4] which assigns parts of speech such as nouns, verbs, adjectives, and adverbs to each token.

Each unigram extracted in the text cleaning phase is subjected to POS tagging. Standard categories include nouns (denoting entities), verbs (actions or states), adjectives (qualifying nouns), and adverbs (modifying verbs, adjectives, or other adverbs).

Table A.10 in Appendix A reports the distribution and frequency of these categories across the corpus, revealing patterns in language use. For instance, a higher frequency of verbs may indicate more dynamic reporting, while nouns suggest a

[4] spaCy's pretrained models provide robust and efficient POS tagging, balancing accuracy with computational speed for large corpora.

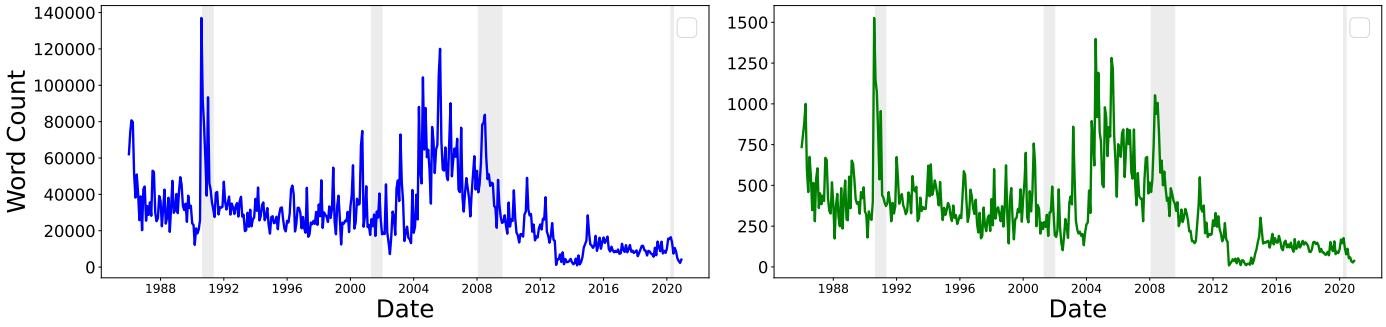

Figure 2: The graph shows a time series plot of the number of words in articles (left) and headlines (right) from January 1986 to December 2020 within each monthly corpus. The shaded regions represents NBER-defined recessions.

focus on entities and facts. Adjectives and adverbs often reflect the writer's stance or tone, providing information relevant for assessing subjectivity or sentiment. Figure A.11 in Appendix A shows how the use of these grammatical categories fluctuates over time, in relation to economic conditions identified by NBER-defined recessions.

2.3. Collocation Analysis

Following the categorization of tokens by their parts of speech, the next stage in our text analysis pipeline involves the extraction of collocations within each monthly corpus. Collocations are significant pairs of words that co-occur more frequently than would be expected by chance. In this study, I initially consider all pairs of consecutive tokens, and in Section 2.4 incorporate frequency-based thresholds.

I focus on bigrams,[5] pairs of consecutive words, identified within the cleaned corpus.

Table A.11 in Appendix A reports the distribution and frequency of collocations segmented by combinations such as noun-noun, adjective-noun, verb-noun, noun-verb, noun-adjective, and verb-adjective pairings. Collocations involving adverbs were excluded due to their rarity in the corpus. This categorization aids in understanding how different types of word pairs contribute to the overall discourse and how they affect crude oil prices. Furthermore, Figure 3 employs word clouds to visualize the top 100 most frequent collocations for both articles and headlines, offering a graphical representation of the common linguistic patterns in the corpus.

2.4. TF-IDF Weighting

Term Frequency-Inverse Document Frequency (TF-IDF) is a statistical measure used in text mining and information retrieval to evaluate the importance of a word or phrase within a corpus. This metric helps in identifying words and collocations that are not only common within each monthly training corpus, but also distinctive across the full set of training corpora.

The TF-IDF value is computed as follows:

$$\text{TF-IDF(w, d)} = tf(w,d) \times \left[\log\left(\frac{n}{df(w)}\right) + 1\right] \quad (1)$$

where $tf(w,d)$ denotes the frequency of word (or term) w in document d, n is the total number of documents, and $df(w)$ is the document frequency of w, indicating the number of documents in which the term appears.[6]

In this study, I apply TF-IDF to the bigrams within our corpus, which is divided into a training set of 262 documents and a testing set of 158 documents. TF-IDF values are computed once, using only the training set. To avoid data leakage and look-ahead bias, I adopt a training/test split combined with 5-fold time-series cross-validation. The selection of TF-IDF thresholds, ranging from 0.01 to 0.2, was achieved by fitting a penalized regression model on the training set. The optimal threshold was 0.01, meaning that terms appearing in less than 1% of documents were ignored. This penalization process, detailed in Section 3.1.1, was essential for optimizing feature selection and enhancing model performance.

Table 2 illustrates the effect of applying the TF-IDF thresholds. From a total of 4,439,908 and 94,709 unique bigrams in news articles and headlines, respectively, the implementation of TF-IDF thresholds reduces the counts to 2,157 and 162. This substantial reduction underscores the efficacy of TF-IDF in distilling the corpus to its most relevant terms, thereby reducing noise and streamlining the feature set for subsequent analysis.

2.5. Validating Collocations

The next stage after TF-IDF weighting in our text analysis pipeline is the validation of collocations through their part-of-speech (POS) patterns. This validation process serves a dual purpose: it reduces the dimensionality of our text data and ensures that the collocations integrated into our forecasting models are both interpretable and relevant.

In this paper, I rely on the Justeson and Katz (1995) grammatical pattern, widely used in the linguistics literature for colloca-

[5] While collocations can extend to trigrams (three-word sequences), quadgrams (four-word sequences), or higher n-grams, our study is confined to bigrams for tractability. Future work could extend the analysis to higher n-grams.

[6] Here, "documents" refer to the monthly corpora of news articles (or headlines) used in the training set. Each corpus is constructed by combining all articles or headlines from a given newspaper within a particular month.

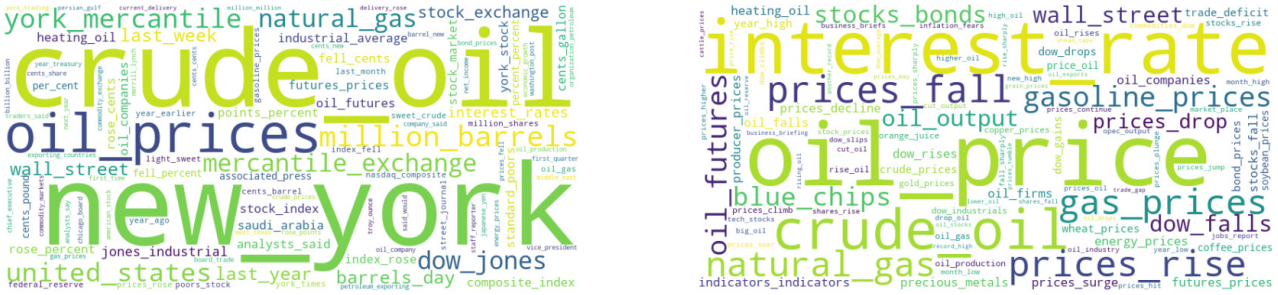

Figure 3: Word cloud of the top 100 most frequent collocations (bigrams) in news articles (left) and headlines (right) across the corpus.

Table 2: Bigram Counts Pre- and Post-TF-IDF

	Count	TF-IDF
News Articles	4,439,908	2,157
News Headlines	94,709	162

This table displays the number of unique bigram collocations identified in news articles and headlines, before and after applying TF-IDF filtering. The "TF-IDF" column shows the number of bigrams retained after excluding low-relevance terms based on their TF-IDF scores.

tion extraction. According to their framework, probable collocations are those multi-word noun phrases specified by the regular expression:[7]

$$((A|N)^*)N \tag{2}$$

$$((A|N)^+|(A|N)^*(NP)^?(A|N)^*)N \tag{3}$$

where A (adjective), N (noun), and P (preposition) are the parts-of-speech. Since our collocations are only bigrams (two-word), our pattern is limited to A-N and N-N. This pattern posits that a significant collocation is likely to be either an adjective followed by a noun, or a sequence of two nouns. Examples include "petroleum industry," "fossil fuel," "global demand," "exploration production," "crude oil," and "energy consumption." This step ensures that the features (bigrams) included in the model have semantic or syntactic relevance, improving interpretability and efficiency.

I extend this framework by considering verb-noun (V-N) (e.g., "increase rate"), noun-verb (N-V) (e.g., "price fall"), noun-adjective (N-A), and verb-adjective (V-A) collocations.

The rationale behind focusing on POS patterns stems from existing research indicating that collocations of different POS compositions exhibit distinct strengths. For example, noun-verb pairs may convey action-related information crucial for understanding dynamics, whereas adjective-noun pairs might offer insights into qualitative aspects of the subjects discussed.

Table 3 summarizes the distribution of collocation patterns in news articles and headlines, based on their adherence to the TF-IDF threshold. For news articles, out of the 2,157 collocations identified, 951 exhibited a noun-noun or adjective-noun pattern (N-N/A-N), 464 a verb-noun or noun-verb pattern (V-N/N-V), and 89 a noun-adjective or verb-adjective pattern (N-A/V-A). Similarly, in news headlines, of the 162 collocations meeting the threshold, 91 were N-N/A-N, 41 were V-N/N-V, and 7 were N-A/V-A. At this point, the corpus is ready to be transformed into a structured form.

Table 3: Distribution of Collocation Patterns in News Articles and Headlines

	TF-IDF	N-N/A-N	V-N/N-V	N-A/V-A
News Articles	2,157	951	464	89
News Headlines	162	91	41	7

This table shows the number of extracted bigram collocations that pass a TF-IDF threshold, grouped by their part-of-speech (POS) patterns. "N-N/A-N" refers to noun-noun and adjective-noun patterns, "V-N/N-V" to verb-noun and noun-verb patterns, and "N-A/V-A" to noun-adjective and verb-adjective patterns.

2.6. *Converting Text into Time Series*

The vector space model (Salton and Yang (1973)) allows us to map each document into a vector in an m-dimensional space, where m represents the number of unique words (or terms) in the document. Repeating this process for all n documents in the corpus, I can construct an $n \times m$ document-term frequency matrix, denoted as $\mathrm{D_t}(n, m)$. In this matrix, each row i corresponds to the document at time t, and each column j corresponds to a unique word (or term). The value in each cell $\mathrm{D_t}(i, j)$ represents the frequency of the j^{th} word in the i^{th} document at time t.

Building upon the foundation laid by the distribution of collocation patterns in news articles and headlines, as summarized in Table 3, our analysis advances by transforming these identified collocations in the documents into a document-term frequency matrix. Specifically, I generate four distinct matrices, each corresponding to a different collocation pattern category identified in the table: noun-noun/adjective-noun (N-N/A-N), verb-noun/noun-verb (V-N/N-V), and noun-adjective/verb-

[7] In the context of regular expressions, * means the preceding token can occur zero or more times; + means it must occur one or more times; and ? means it can occur zero or one time.

adjective (N-A/V-A), in addition to a comprehensive matrix derived from the TF-IDF weighted collocations.

Table 4 details the structure of our transformed text data into four distinct time series matrices based on identified collocation patterns. Each matrix, labeled from $D_{1,t}$ to $D_{4,t}$, corresponds to a unique aspect of our corpus analysis, ranging from broad TF-IDF weighted collocations to more focused grammatical patterns. Each matrix dimension is represented by the number of documents (420) versus the number of unique collocations. At this point, our matrices are ready to be used for our forecasting exercise.

3. Methodology

This paper compares the predictive performance of a standard text-based crude oil price forecasting model with an alternative model that incorporates text-based features alongside a CPV process and an attention mechanism. It also investigates whether using the full text, rather than only news headlines, improves forecasting accuracy. The forecasting equation is given by:

$$r_{t+h} = \alpha + \eta r_t + \Gamma' D_t + \epsilon_{t+h}, \quad h = 1, 3, 6, 9 \tag{4}$$

where r_{t+h} denotes the return at time $t + h$, h represents the forecast horizon (1, 3, 6, or 9 months ahead), α is the intercept, η captures the effect of the lagged target variable r_t, D_t is the document-term frequency matrix derived from the text data, Γ is the associated coefficient matrix, and ϵ_{t+h} is the forecast error.

The analysis spans January 1986 to December 2020, covering 420 monthly observations ($T = 420$). Models are estimated using a growing-window approach, which expands the training sample as new observations become available. Each model is first estimated with the initial T_1 observations, then re-estimated as the sample expands, generating dynamic out-of-sample forecasts. This recursive estimation technique, as opposed to a fixed rolling window, provides a more adaptive simulation of real-world forecasting scenarios.[8]

The training set spans January 1986 to October 2007, with 262 observations ($T_1 = 262$). The remaining $P = T - T_1 = 158$ observations form the test set. For $h = 1$, the evaluation period runs from November 2007 to December 2020 (158 forecasts). For $h = 9$, it runs from December 2007 to June 2020 (150 forecasts).

To avoid data leakage, the document-term frequency matrices are standardized prior to estimation.[9] This ensures that information from future periods does not contaminate past forecasts.

Because the text data are high-dimensional, the document-term frequency matrix poses risks of overfitting. To address this, I adopt the attention mechanism proposed by Lima and Godeiro (2023).

The attention mechanism reweights terms in the document-term frequency matrix, emphasizing those most predictive of crude oil prices and down-weighting less informative features. This reduces dimensionality, mitigates overfitting, and enhances the model's generalization to unseen data. The following section describes the attention mechanism in detail.

3.1. Incorporating Attention into the Model

The approach of Lima and Godeiro (2023) incorporates attention into the forecasting framework through two main steps: feature selection and feature extraction. In the feature selection step, a regularization method is applied to identify the most predictive features from the full set ($D_{1,t}$, $D_{2,t}$, $D_{3,t}$, $D_{4,t}$). In the feature extraction step, principal component analysis (PCA) is then used to transform the selected features into a new set of uncorrelated variables. The following sections describe each step in detail.

3.1.1. Feature Selection – Time-Varying Dictionary

The feature selection process aims to create a time-adaptive dictionary that retains only the most predictive terms. To achieve this, an elastic net (Zou and Hastie (2005)) is estimated recursively for the following linear prediction equation:

$$r_{s+h} = \eta_h r_s + D_s' \beta_h + \epsilon_{s+h} \tag{5}$$

where $h = 1, 3, 6, 9$ denotes the forecast horizon, r_{s+h} is the crude oil return at month $s + h$, η_h captures the effect of past returns r_s, and β_h is estimated by minimizing the loss function:

$$\hat{\beta}_h = \arg\min_{\beta_h} \frac{1}{2(t-h)} \sum_{s=1}^{t-h} \left(r_{s+h} - \eta_h r_s - D_s' \beta_h\right)^2 + \underbrace{\alpha\rho\|\beta_h\|_1 + \frac{\alpha(1-\rho)}{2}\|\beta_h\|_2^2}_{\text{penalty}} \tag{6}$$

Here, $\alpha \geqslant 0$ controls the magnitude of the penalty, while $\rho \in [0, 1]$ governs the balance between ridge regression (Hoerl and Kennard (1970)) and LASSO regression (Tibshirani (1996)). The elastic net reduces to LASSO when $\rho = 1$ and approaches ridge regression as $\rho \to 0$. For other values of ρ, the penalty term balances between the l_1-norm and the l_2-norm.[10]

The observations r_{s+h} are regressed on the predictors D_s for $s = 1, \ldots, r-h$ and $r = R, \ldots, t-h$. For each forecast origin t, the loss function identifies the most predictive words (or terms) from the original time-series matrix $D_{i,t}$ for $i = 1, \ldots, 4$. This approach mirrors the strategy developed by Bai and Ng (2008).

An integral part of the approach presented in Lima and Godeiro (2023) is the implementation of an innovative updating scheme. This scheme employs the Sahm rule recession indicator[11] as a key tool for eliminating look-ahead bias. For example,

[8] For a detailed discussion of recursive versus rolling-window approaches, see Morales-Arias and Moura (2013).

[9] Standardization uses the `StandardScaler` module in scikit-learn, fitted only on the training set and then applied to both training and test sets.

[10] The tuning of α and ρ is performed using the `ElasticNetCV` function from scikit-learn.

[11] The Sahm Rule Recession Indicator, developed by Claudia Sahm, identifies the start of a recession when the three-month moving average of the national unemployment rate rises by at least 0.5 percentage points relative to its 12-month minimum.

Table 4: Summary of Time Series Matrices

Matrix	Description	News Articles Dimensions	News Headlines Dimensions
$D_{1,t}$	Contains all collocations that meet the TF-IDF threshold, representing a comprehensive overview of significant term pairs.	420×2,157	420×162
$D_{2,t}$	Captures collocations that adhere to the grammatical pattern outlined by Justeson and Katz (1995), emphasizing syntactically relevant pairings.	420×951	420×91
$D_{3,t}$	Captures collocations that adhere to the verb-noun and noun-verb patterns, highlighting dynamic actions and interactions within the text.	420×464	420×41
$D_{4,t}$	Captures collocations conforming to noun-adjective and verb-adjective patterns, representing the descriptive and qualitative aspects of the text.	420×89	420×7

Each matrix $D_{i,t}$ represents a different way of transforming text data into a document-term frequency matrix. $D_{1,t}$ includes all collocations selected using a TF-IDF threshold. $D_{2,t}$ includes syntactic collocations identified using the Justeson and Katz (1995) pattern (e.g., adjective-noun and noun-noun pairs). $D_{3,t}$ captures verb-noun and noun-verb pairings, while $D_{4,t}$ includes noun-adjective and verb-adjective combinations. Dimensions correspond to the number of time periods (rows) and the number of unique collocations (columns) for each data source.

if the indicator crosses the 0.5 threshold in December 2008, the dictionary is updated at the forecast origin of January 2009 and remains constant until the indicator falls below the threshold.

Figures 4 and 5 together illustrate how the number of predictive features evolves across horizons under the Sahm rule and continuous updating schemes. At horizons 1, 3, and 6, the Sahm rule updates the dictionary three times, while at horizon 9 it updates twice. By contrast, the continuous scheme updates the dictionary throughout the entire analysis period, ensuring responsiveness to new information but at a higher computational cost. The overarching goal of both approaches is to replicate real-time forecasting conditions where the decision to refresh the dictionary is data-driven.

3.1.2. Feature Extraction

After identifying the most predictive terms, the next stage proceeds to feature extraction. This step entails the estimation of common factors, which are linear combinations of selected predictive features of $D_{i,t}$, denoted as $D^*_{i,t}$ for each $i = 1, \ldots, 4$. This estimation is performed using principal components, with data $D^*_{k,s}$ considered up to the forecast origin t, where $k = 1, \ldots, K$ are the most predictive time series. To put it simply, the principal components estimators are defined as:

$$(\Lambda_t, F_t) = \underset{\{\lambda_k, f_s\}}{\operatorname{argmin}} \frac{1}{Kt} \sum_{k=1}^{K} \sum_{s=1}^{t} \left(D^*_{k,s} - \lambda'_{k,t} f_{s,t}\right)^2, \qquad (7)$$

where Λ_t denotes a vector of factor loadings and F_t is a vector of r common factors. Here, $f_{s,t}$ represents the *sth* observation on the vector of common factors, estimated using data accumulated until time point t.[12] This interpretation emphasizes that the factors in F_t are informed exclusively by the most predictive collocations D^*_t.

Following the approach in Lima and Godeiro (2023), the maximum number of factors to consider, denoted as r_{max}, is set to eight. The optimal number of factors is determined using the method outlined by Ahn and Horenstein (2013). Next, Equation 5 is estimated using the selected factors, and only those with a p-value less than 0.05 in the prediction equation are retained, as suggested by Bai and Ng (2008).[13] If none of the factors meet this threshold, the principal components are retained to preserve model continuity. Finally, Equation 5 is re-estimated using only the statistically significant (or retained) factors, as detailed in Algorithm 1.

Table 5 presents the average number of common factors for each forecast horizon, for both news articles and headlines. The table distinguishes between two approaches: feature selection guided by the Sahm rule and continuous updating of features

[12] The dual index (s,t) signifies the use of data up to each forecast origin for the estimation of the r common factors (as detailed in Gonçalves et al. (2017)). This conventional method in forecasting effectively circumvents the look-ahead bias.

[13] Bai and Ng (2005) demonstrate that the least squares estimates derived from factor-augmented forecasting regressions exhibit $\sqrt{T}$ consistency and asymptotic normality. Moreover, they establish that the pre-estimation of factors does not compromise the consistency of the second-stage parameter estimates or their associated standard errors.

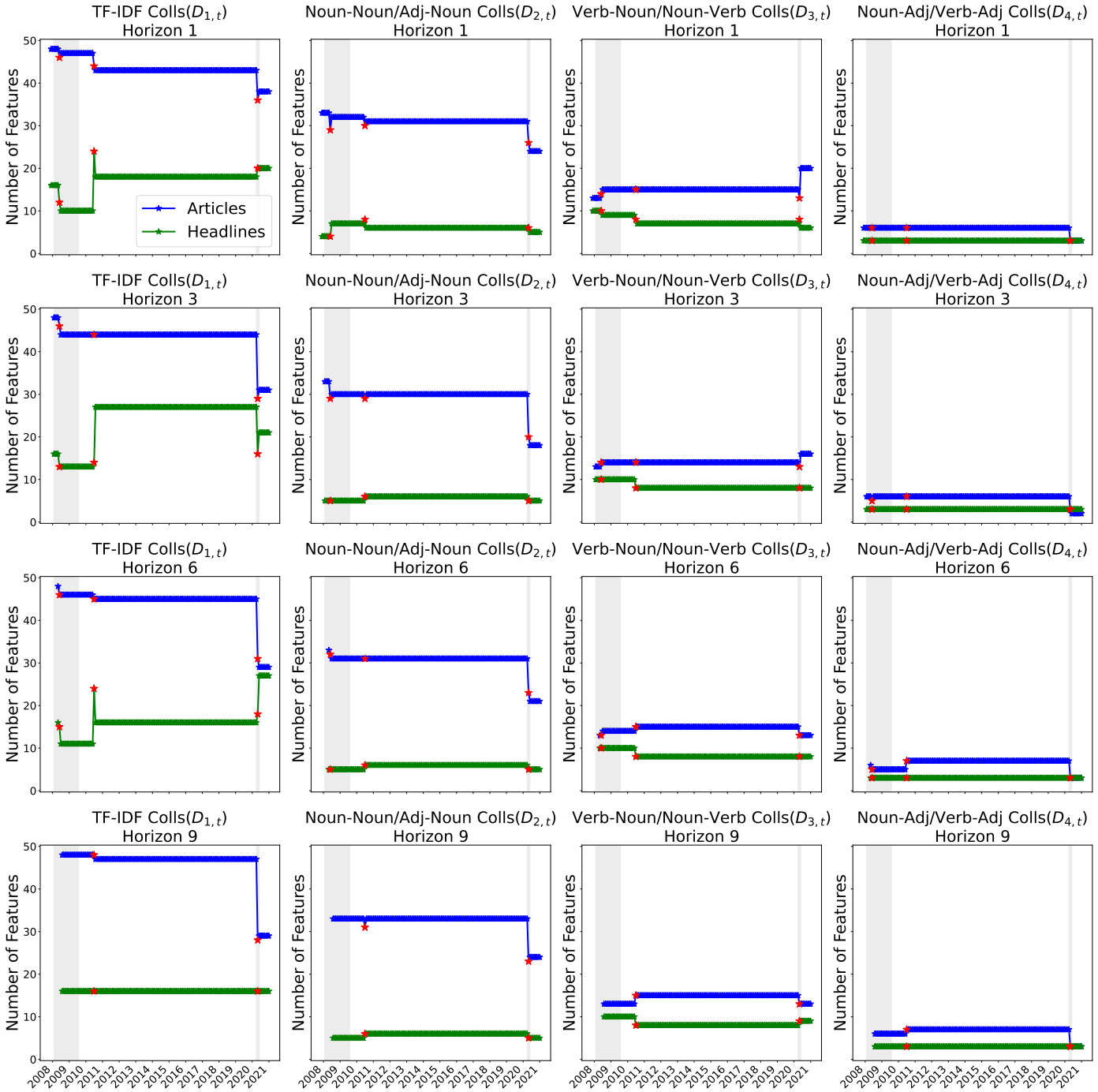

Figure 4: Number of features selected from the feature set obtained from **news articles (blue)** and **news headlines (green)** for each forecast horizon under the **Sahm rule updating scheme**. The red marker indicates the point where the Sahm rule indicator signals the model, and the shaded region represents NBER-defined recessions.

over time. It illustrates how the original large set of predictive features is reduced to a smaller, more manageable number that still retains the key information necessary for forecasting.[14]

[14] See the online appendix for figures showing the number of features selected across time, by model and forecast horizon.

4. Data – Target Variable

In this study, the West Texas Intermediate (WTI) crude oil daily spot price, obtained from the U.S. Energy Information Administration (EIA), serves as the measure of crude oil price used in this study. Daily prices are aggregated into monthly averages for analysis. The monthly log return is computed as $r_t = \ln\left(\frac{R_t}{R_{t-1}}\right)$, where R_t denotes the oil price at time t. The

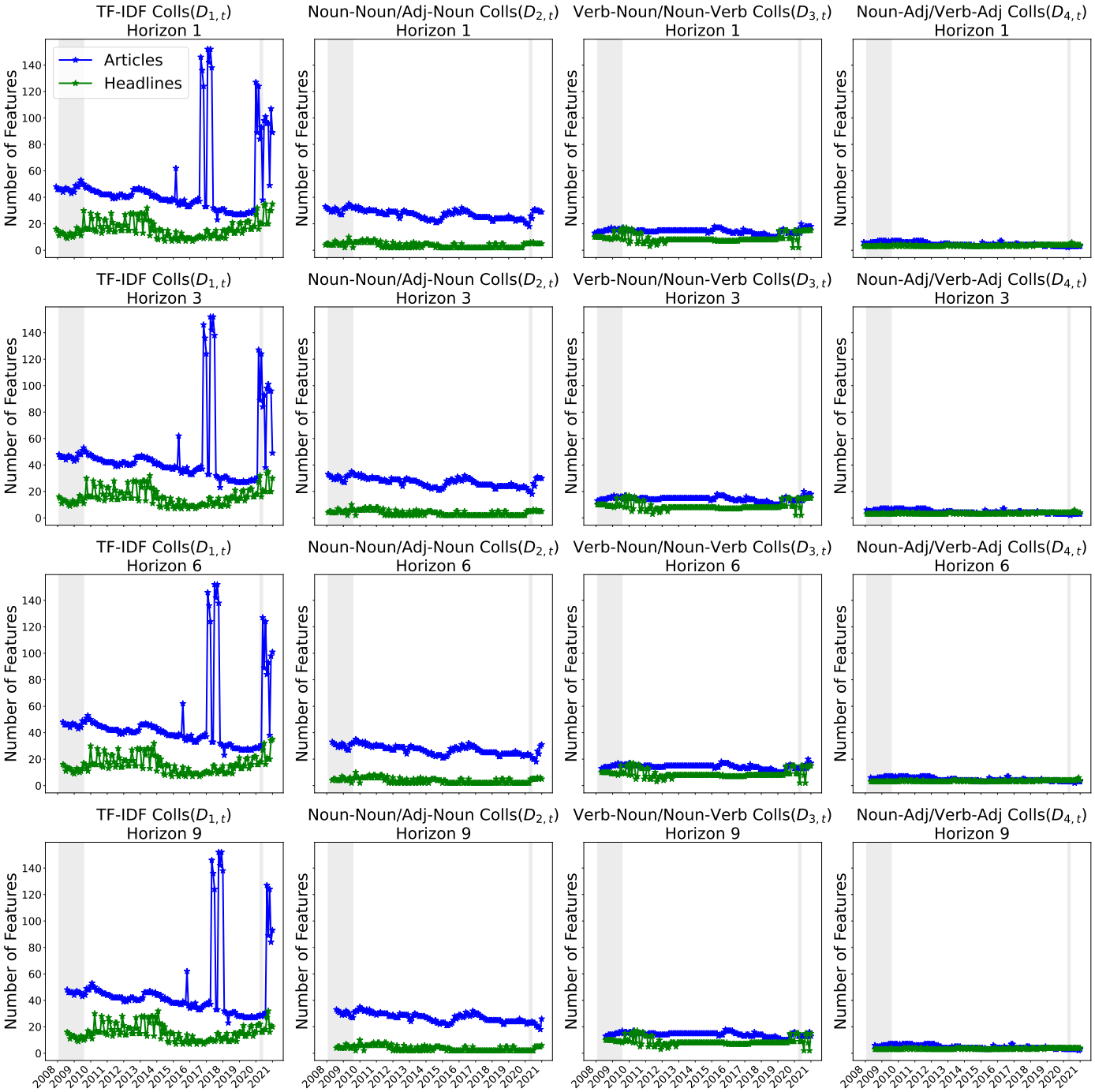

Figure 5: Number of features selected from the feature set obtained from **news articles (blue)** and **news headlines (green)** for each forecast horizon under the **continuous updating scheme**. The shaded region represents NBER-defined recessions.

sample spans January 1986 to December 2020, yielding 420 monthly observations.[15] The in-sample period extends from January 1986 to October 2007, while the out-of-sample (OOS) period covers November 2007 to December 2020.

As shown in Figure 6, WTI crude oil prices exhibit substantial fluctuations over time. Pronounced declines often align with NBER-defined recessions, underscoring the sensitivity of oil prices to macroeconomic conditions. These downturns also tend to coincide with signals from the Sahm Rule recession indicator, which relies on the three-month moving average of the national unemployment rate. In this context, the Sahm Rule provides a valuable real-time tool for updating the dictionary of terms in the forecasting model, particularly during volatile economic periods identified by both the NBER and the Sahm Rule.

[15] U.S. Energy Information Administration, *West Texas Intermediate (WTI) – Cushing, Oklahoma* [DCOILWTICO], retrieved from FRED, Federal Reserve Bank of St. Louis; https://fred.stlouisfed.org/series/DCOILWTICO.

Table 5: Average Number of Extracted Common Factors

Sahm Rule	$h=1$	$h=3$	$h=6$	$h=9$
News Articles				
$D_{1,t}$	1.78	1.94	1.27	1.87
$D_{2,t}$	2.09	1.78	1.84	1.94
$D_{3,t}$	1.00	1.00	1.04	1.03
$D_{4,t}$	3.46	3.40	2.03	2.14
News Headlines				
$D_{1,t}$	1.74	1.02	1.77	1.91
$D_{2,t}$	1.16	1.37	1.34	1.31
$D_{3,t}$	2.54	5.35	5.56	5.69
$D_{4,t}$	1.06	1.06	1.06	1.06
Continuous				
News Articles				
$D_{1,t}$	1.71	1.71	1.71	1.71
$D_{2,t}$	1.52	1.53	1.54	1.55
$D_{3,t}$	1.00	1.00	1.00	1.00
$D_{4,t}$	2.42	2.44	2.47	2.50
News Headlines				
$D_{1,t}$	2.44	2.46	2.49	2.52
$D_{2,t}$	1.23	1.24	1.24	1.22
$D_{3,t}$	4.47	4.52	4.59	4.66
$D_{4,t}$	1.63	1.64	1.65	1.67

This table reports the average number of common factors extracted for each forecast horizon ($h = 1, 3, 6, 9$) under different updating schemes using either full news articles or headlines. The models are grouped into two categories: **Sahm Rule** models, which apply a recession indicator based on the Sahm Rule to guide feature extraction, and **Continuous** models, which update feature sets continuously over time without conditioning on a recession signal.

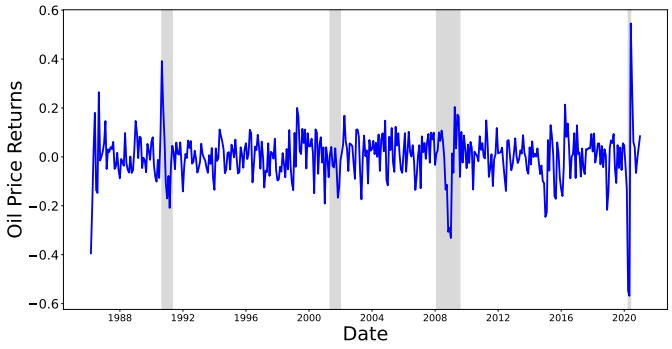

Figure 6: Monthly log returns of WTI crude oil prices between January 1986 and December 2020. Shaded areas indicate NBER-defined recessions.

5. Forecast Evaluation

The benchmark model for this study combines the predictive power of our lagged target variable with text-based features derived from a pre-established dictionary from Loughran et al. (2019). This dictionary categorizes words into positive and negative sentiments related to oil price movements, containing 291 positive and 539 negative terms.

To incorporate these sentiment indicators, I first construct a document-term frequency matrix using full news articles from our corpus, focusing on the occurrence of words from the Loughran et al. (2019) dictionary within our corpus. Given the high dimensionality of this matrix totaling 830 sentiment-related features, I adopt the traditional approach by applying PCA to extract the most significant common factors without incorporating any CPV process or attention mechanism. The forecasting equation for the benchmark model is formulated as follows:

$$r_{t+h} = \alpha + \eta r_t + \Pi' L_t + \varepsilon_{t+h}, \quad h = 1, 3, 6, 9 \tag{8}$$

where r_{t+h} denotes the return at time $t+h$, h represents the forecast horizon (1, 3, 6, or 9 months ahead). The intercept of the model is denoted by α, while η captures the influence of the lagged target variable r_t. The term L_t represents the document-term frequency matrix derived from the text data, with Π as its associated coefficient matrix. and ε_{t+h} represents the forecast error at time $t+h$.

The primary objective of the forecast evaluation in this study is to assess the out-of-sample performance of the models that incorporate the CPV process together with the attention mechanism against the benchmark model which uses a pre-established dictionary, and does not incorporate the attention mechanism. Additionally, I assess the performance of models that extract text features from actual news articles as against those from news headlines. It is expected that if the models contain predictive features, their out-of-sample forecasts would outperform those from the benchmark model. Furthermore, if the validation and attention mechanism improve the forecast accuracy, their out-of-sample forecasts should perform better than the benchmark model.

To assess performance, two key metrics are employed: Out-of-Sample R^2_{oos} as defined in Campbell and Thompson (2008), and Root Mean Squared Forecast Error (RMSFE). The formal representations of these metrics are as follows:

$$R^2_{i,oos} = 100 \times \left[1 - \frac{\sum_{t=1}^{P} \left(r_{t+h} - \hat{r}^i_{t+h|t} \right)^2}{\sum_{t=1}^{P} \left(r_{t+h} - \hat{r}^{bmk}_{t+h|t} \right)^2} \right], \quad i = 1, \ldots, 4 \tag{9}$$

$$RMSFE_i = \frac{\sqrt{\sum_{t=1}^{P} \left(r_{t+h} - \hat{r}^i_{t+h|t} \right)^2}}{\sqrt{\sum_{t=1}^{P} \left(r_{t+h} - \hat{r}^{bmk}_{t+h|t} \right)^2}}, \quad i = 1, \ldots, 4 \tag{10}$$

In these equations, P is the number of h-step-ahead OOS forecasts, with $h = 1, 3, 6, 9$, $\hat{r}^i_{t+h|t}$ is the forecast of r_{t+h} for model i using information up to period t, and $\hat{r}^{bmk}_{t+h|t}$ is the respective benchmark forecast.

A positive (negative) value for the $R^2_{i,oos}$ statistic implies that the forecast $\hat{r}^i_{t+h|t}$ outperforms (is outperformed by) the benchmark $\hat{r}^{bmk}_{t+h|t}$. For the $RMSFE_i$ metric, if its value is less (greater) than one, then model i outperforms (is outperformed by) the

benchmark model.

To provide a comprehensive assessment, this study employs the Clark and West (2007) test for predictive accuracy, which is well-suited for comparing nested models where one model may be a restricted version of another. This statistical test allows for an additional objective measure of the relative forecast accuracy of different models. The test is given by:

$$CW_i = \left(\hat{r}^{bmk}_{t+h|t} - r_{t+h}\right)^2 - \left(\hat{r}^{i}_{t+h|t} - r_{t+h}\right)^2 + \left(\hat{r}^{bmk}_{t+h|t} - \hat{r}^{i}_{t+h|t}\right)^2 \tag{11}$$

The null hypothesis of the CW test is that the mean squared error of the target model i is larger than or equal to that of the benchmark and the alternative hypothesis of the CW test is that the mean squared error of the target model i is smaller than that of the benchmark.

6. Results

The findings of this study are organized into three distinct sections. Section 6.1 presents the core results, emphasizing the differences in forecast performance between models that extract features from full news articles versus those that utilize only news headlines. This section also contrasts the performance of models that integrate the Collocation Pattern Validation (CPV) process and an attention mechanism with those that rely solely on TF-IDF thresholding and the attention mechanism, and those that rely on a pre-defined dictionary.

Section 6.2 delves deeper into the predictive power of specific collocations, including examples of news extracts that feature these influential collocations. This analysis helps illustrate how certain phrases within the news text correlate with shifts in oil price forecasts.

Finally, Section 6.3 compares the performance of a model that incorporates macroeconomic data alone with one that uses both macroeconomic and textual data. This comparison aims to highlight the added value of integrating textual analysis into traditional economic forecasting models.

6.1. Value of News Articles, CPV, and Attention Mechanism

Table 6 reports the out-of-sample forecasting performance of all models against the benchmark model. Four major findings emerge from the table. First, the standout performer among all models is the one employing text features from news articles that underwent a verb–noun/noun–verb CPV process, augmented by an attention mechanism ($D_{3,t}$). This model's superior performance can be attributed to the ability of these part-of-speech (POS) patterns to capture complex market dynamics. Verbs typically convey directional sentiment, serving as indicators of market trends. Examples of these collocation patterns, crucial in text-based crude oil price forecasting, include phrases such as "prices fell," "prices surged," "spending rose," "cut costs," and "prices tumbled."

Second, models using verb–noun and noun–verb collocation features from full news articles – combined with an attention mechanism – demonstrate superior performance under both updating schemes. This advantage reflects not only the linguistic role of verbs in expressing direction and sentiment, but also the simple fact that full news articles provide a much larger volume of text than headlines. The greater volume increases the likelihood that informative collocations appear more frequently, thereby producing a sharper and more precise predictive signal. Thus, the contrast between articles and headlines is consistent with both a richer linguistic structure and a volume effect.

Another interesting finding from the study is that the Sahm rule updating scheme, despite updating the dictionary only three times throughout the OOS period—twice at $h = 9$—produces results comparable to those of the continuously updating scheme. Particularly noteworthy is model $D_{3,t}$, which employs a verb-noun/noun-verb collocation pattern validation scheme. This model selects a maximum of 20 features during the feature selection process, updates the dictionary as needed, and typically extracts one factor from the selected features. These characteristics make model $D_{3,t}$ notably effective for forecasting: it is simpler and less computationally demanding than the continuous updating approach, as well as being more parsimonious and efficient.

Third, at all horizons, models that extract features from news articles demonstrate superior performance over models that extract features from news headlines. This can be attributed to the more comprehensive content found in full articles, which provides a richer and more detailed context for predicting changes in oil prices. News articles often contain extended discussions of factors influencing the markets, such as geopolitical events, supply chain updates, or financial results, offering deeper insights than the typically concise and less detailed news headlines.

These results further validate the central thesis of the research, affirming that the integration of validated collocations and improved attention mechanisms can significantly enhance the forecasting power of text-based models, allowing practitioners to move "Beyond News Headlines and TF-IDF".

To assess the cumulative performance of the models over time, I constructed time-series plots to illustrate the differences in the cumulative squared prediction error (CSPE) between the benchmark forecast and the CSPEs for the forecasts generated by each model. This visual representation offers insights into how consistently each forecasting model outperforms or underperforms relative to the benchmark model throughout the observed period.

Figures 7 and 8 display the CSPE for models using features from news articles and headlines, contrasting the Sahm rule and continuously varying updating schemes respectively. In each panel of these figures, the curve for model $D_{3,t}$ remains above the zero line across all forecast horizons, peaking during the COVID-19 recession. This indicates superior performance of $D_{3,t}$ compared to the benchmark model.

6.2. Identifying Key Verb-Noun/Noun-Verb Collocations with Predictive Power

In this section, I examine verb-noun and noun-verb collocations that exhibit strong predictive power across the out-of-sample (OOS) period, the business cycle, and specifically during the COVID-19 recession. From the feature selection analysis in Section 3.1.1, the maximum number of features selected for these collocations is 20. To illustrate this, Figure 9 presents bar plots of the top 20 features for both the Sahm rule and continuously updating schemes. In addition, Figure 10 provides excerpts from news articles that contain these collocations.

Examples of these impactful collocations include *drop oil*, *falling oil*, *rose barrel*, *price fell*, *prices decline*, *prices fall*, and *prices surged*. These phrases are directly related to crude oil prices, making it unsurprising that they hold strong predictive power. Notably, the top seven collocations were selected consistently across the 617 out-of-sample periods, underscoring their robustness regardless of economic conditions.[16]

From Table 5, the average number of common factors extracted from the top 20 collocations is one. The collocations that load heavily on the first principal component (PC1) largely consist of terms directly

[16] The 617 OOS periods consist of 158 at $h = 1$, 156 at $h = 3$, 153 at $h = 6$, and 150 at $h = 9$.

Table 6: Out-of-Sample Forecast Evaluation

	$h=1$			$h=3$			$h=6$			$h=9$		
Sahm Rule	$R^2_{oos}(\%)$	$RMSFE$	CW	$R^2_{oos}(\%)$	$RMSFE$	CW	$R^2_{oos}(\%)$	$RMSFE$	CW	$R^2_{oos}(\%)$	$RMSFE$	CW
News Articles												
$D_{1,t}$	2.03	0.990	2.203**	5.77	0.971	1.297*	4.68	0.976	1.132	5.64	0.971	1.363*
$D_{2,t}$	2.66	0.987	1.437*	-1.44	1.007	0.432	-0.23	1.001	0.795	-0.84	1.004	0.827
$D_{3,t}$	**11.59**	**0.940**	2.856***	**13.19**	**0.932**	2.784***	**11.90**	**0.939**	2.716***	**11.75**	**0.939**	2.732***
$D_{4,t}$	3.30	0.983	3.315***	3.34	0.983	2.945***	2.13	0.989	2.705***	2.86	0.986	2.569***
News Headlines												
$D_{1,t}$	1.80	0.991	1.923**	-1.11	1.006	-0.269	-0.81	1.004	0.229	1.28	0.994	0.748
$D_{2,t}$	0.66	0.997	1.278	-0.03	1.000	0.941	0.15	0.999	1.049	0.10	1.000	1.054
$D_{3,t}$	3.59	0.982	1.616*	1.30	0.993	1.385*	1.13	0.994	1.415*	0.12	0.999	1.121
$D_{4,t}$	2.19	0.989	1.889**	1.87	0.991	1.551*	1.76	0.991	1.513*	1.42	0.993	1.380*
Continuous												
News Articles												
$D_{1,t}$	3.04	0.985	2.154**	3.79	0.981	1.600*	4.05	0.980	2.173**	8.30	0.958	1.693**
$D_{2,t}$	-0.35	1.002	0.749	-6.70	1.033	-0.808	-4.97	1.025	-0.666	-3.87	1.019	-0.208
$D_{3,t}$	**10.55**	**0.946**	3.094***	**11.10**	**0.943**	2.807***	**10.93**	**0.944**	2.871***	**11.43**	**0.941**	3.003***
$D_{4,t}$	1.44	0.993	2.030**	1.51	0.992	2.014**	2.17	0.989	2.618***	1.81	0.991	1.961**
News Headlines												
$D_{1,t}$	0.76	0.996	1.106	-0.18	1.001	0.862	-2.78	1.014	-0.199	3.59	0.982	1.409*
$D_{2,t}$	-0.68	1.003	0.699	0.21	0.999	1.305*	0.63	0.997	1.725**	0.88	0.996	1.555*
$D_{3,t}$	3.75	0.981	2.203**	0.70	0.996	1.250	-0.22	1.001	1.007	1.59	0.992	1.528*
$D_{4,t}$	2.87	0.986	2.653***	2.23	0.989	2.334**	2.39	0.988	2.380***	2.18	0.989	2.252**

This table presents the out-of-sample forecast performance of models, $D_{1,t}$ to $D_{4,t}$, across different horizons, $h = 1, 3, 6, 9$. The metrics $R^2_{oos}(\%)$, $RMSFE$, and CW represent the percentage of out-of-sample R^2, the root mean squared forecast error, and the Clark and West (2007) test statistic, respectively. Bold figures indicate the best-performing model in terms of R^2_{oos} and $RMSFE$ for each horizon. Asterisks denote significance levels: * for 10%, ** for 5%, and *** for 1%.

associated with movements in oil prices, including explicit mentions of price changes (e.g., *drop oil*, *prices surged*, *fell barrel*) as well as broader economic indicators (e.g., *cut costs*, *spending rose*). These collocations are inherently tied to market activity and sentiment, which are crucial for predicting commodity prices such as oil.

The analysis in this section examines the impact of specific verb-noun and noun-verb collocations on the volatility of crude oil prices. These collocations predominantly describe past market events, such as *prices fell*, *prices tumbled*, and *prices dropped*, which are commonly found in retrospective news reports. To evaluate their predictive influence, I estimate a regression model that incorporates both the lagged change in oil prices and the first principal component derived from these collocations.

The regression results, detailed in Table 7, indicate a significant negative relationship between the principal component of these collocations and changes in crude oil prices. Specifically, the presence of collocations such as "prices fallen," "prices fall," and "prices dropped" in news discourse is strongly associated with a decrease in oil price returns. This relationship should be interpreted as associative rather than causal: the collocations reflect how market participants react to news, which in turn is linked to price movements, rather than the phrases themselves directly causing the changes. The factor loadings of the first principal component, which measure the influence of each collocation, confirm that these terms contribute substantially to the component associated with declining oil price returns. As shown in Figure A.12 in Appendix A, these collocations have high positive loadings, and the scatterplot in the same figure visually depicts the negative correlation between the PC1 scores and the monthly log difference of oil prices. Together, these results underscore how negative sentiment in oil-related news is closely tied to actual decreases in oil price returns.

Table 7: Relationship Between Verb-Noun/Noun-Verb Collocations and Oil Price Changes

	$r_t = \ln\left(\frac{R_t}{R_{t-1}}\right)$
Intercept	0.002
	(0.004)
r_{t-1}	0.174***
	(0.045)
PC_1	-0.025***
	(0.003)
N	418
R^2	0.189
Adjusted R^2	0.185

Standard errors in parentheses.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Note: This table reports associations, not causal effects. The link between collocations and oil price changes likely reflects market responses to news, rather than collocations directly driving price movements.

6.3. Integrating Macroeconomic Data with Text

Numerous studies have traditionally relied on macroeconomic indicators to forecast crude oil prices(Zhang et al. (2019, 2022, 2021); Miao et al. (2017); Nonejad (2020)). However, with advancements in data processing and natural language processing technologies, there is a growing interest in exploring how textual data from news articles and reports can complement these traditional models.

This section assesses the effectiveness of combining macroeconomic data with verb-noun/noun-verb collocations extracted from news text to enhance the predictive accuracy of crude oil price models. A comprehensive dataset from the Federal Reserve Economic Data (FRED-MD) (McCracken and Ng (2016)) serves as our base of traditional macroeconomic variables.[17]

[17] The FRED-MD data is available at FRED-MD

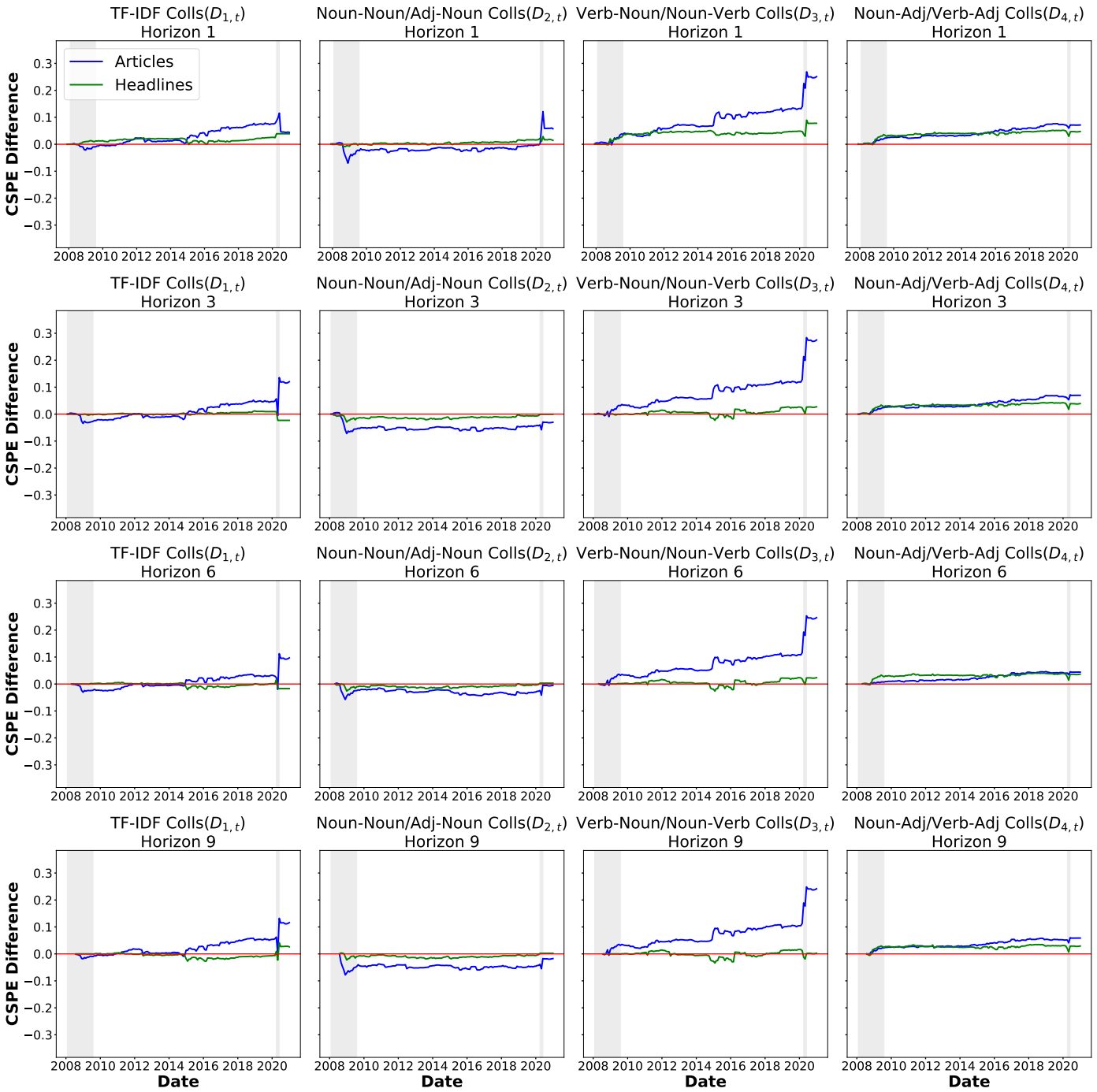

Figure 7: This graph shows the difference between the CSPE of the benchmark model and that of selected forecasting models under the **Sahm rule updating scheme**, from November 2007 to December 2020. The blue line corresponds to features from **news articles**, while the green line corresponds to **news headlines**. A positive slope indicates the superior performance of the selected model compared to the benchmark. Conversely, a downward trajectory suggests that the benchmark model is outperforming the selected forecasting model. If the curve remains above the zero mark as the period concludes, it signifies that the selected model yielded a smaller error throughout the OOS period. Note: The shaded regions represent NBER-defined recessions.

Initially, all variables containing missing values are removed, resulting in 113 variables for the forecasting exercise. These variables are then transformed according to the codes provided by McCracken and Ng (2016) to ensure stationarity.

Subsequently, common factors are extracted from these variables, employing an attention mechanism to enhance their relevance for forecasting. This approach is benchmarked against a model that augments the factors obtained from the macroeconomic data with common factors from the $D_{3,t}$ model, which incorporates verb-noun and noun-verb collocations.

The results, as presented in Table 8, show a notable improvement in out-of-sample predictive performance when text data is integrated with macroeconomic data. Particularly, the augmentation with $D_{3,t}$ values across all forecast horizons compared to models relying solely

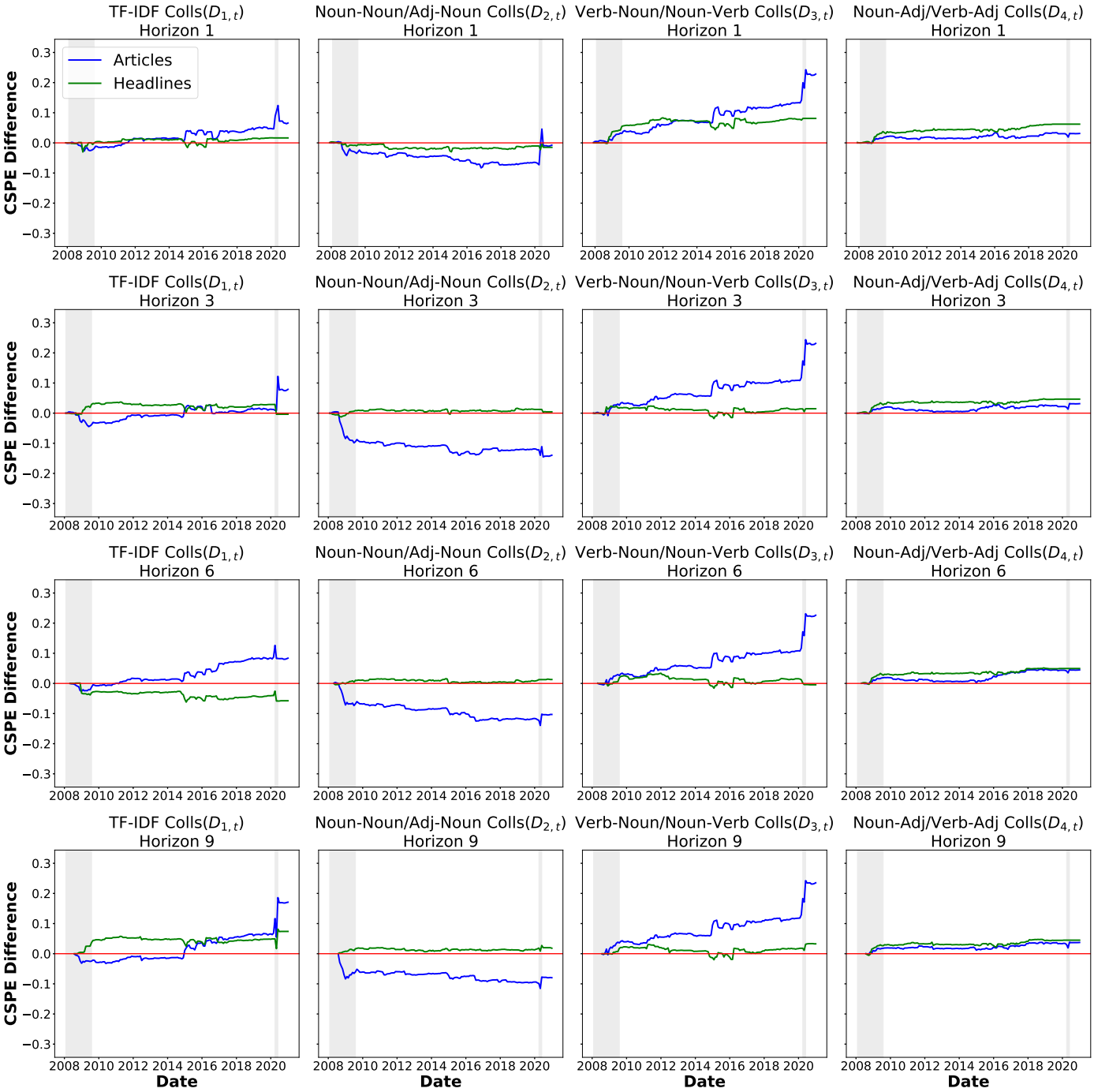

Figure 8: This graph shows the difference between the CSPE of the benchmark model and that of selected forecasting models under the **continuous updating scheme**, from November 2007 to December 2020. The blue line corresponds to features from **news articles**, while the green line corresponds to **news headlines**. A positive slope indicates the superior performance of the selected model compared to the benchmark. Conversely, a downward trajectory suggests that the benchmark model is outperforming the selected forecasting model. If the curve remains above the zero mark as the period concludes, it signifies that the selected model yielded a smaller error throughout the OOS period. Note: The shaded regions represent NBER-defined recessions.

on FRED-MD data.

These findings, consistent with that of Zheng et al. (2024) and Ellingsen et al. (2022) suggest that while traditional macroeconomic data remains a robust predictor of oil price movements, its predictive power can be substantially enhanced by incorporating real-time, context-rich information from news text. This integration captures underlying market sentiments and emerging trends that hard data alone may not fully detect, thereby offering a more comprehensive approach to forecasting in volatile markets like crude oil.

7. Robustness Checks

To evaluate the stability and reliability of the main results, this section presents several robustness exercises. First, I compare the predic-

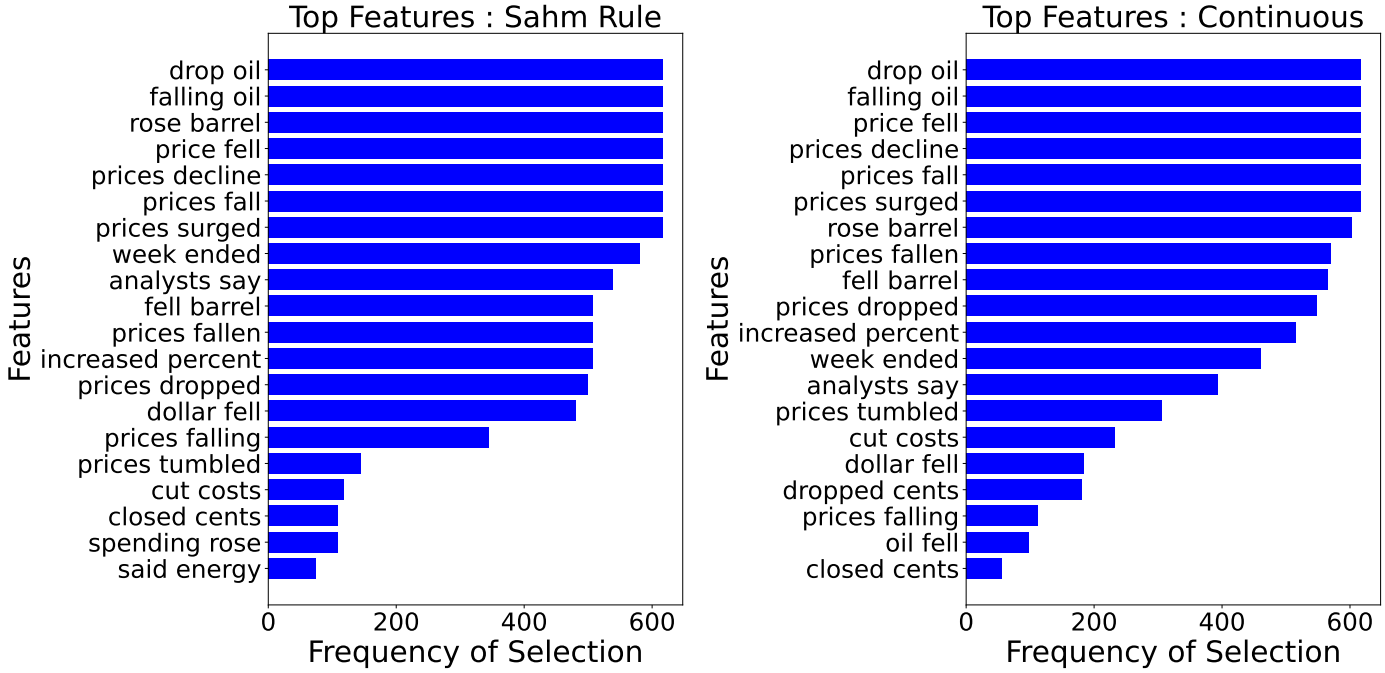

Figure 9: Top 20 V-N/N-V collocations in Sahm rule (left) and continuous (right) updating schemes.

Table 8: Out-of-Sample Forecast Evaluation: Macroeconomic Data with Textual Data

	$h=1$			$h=3$			$h=6$			$h=9$		
	$R^2_{oos}(\%)$	$RMSFE$	CW	$R^2_{oos}(\%)$	$RMSFE$	CW	$R^2_{oos}(\%)$	$RMSFE$	CW	$R^2_{oos}(\%)$	$RMSFE$	CW
Sahm rule												
FRED-MD	1.39	0.993	2.097**	0.69	0.997	1.582*	0.45	0.998	1.417*	3.24	0.984	2.779***
FRED-MD + $D_{3,t}$	**11.73**	**0.940**	2.905***	**13.22**	**0.932**	2.805***	**11.91**	**0.939**	2.720***	**12.81**	**0.934**	2.898***
Continuous												
FRED-MD	1.51	0.992	1.846**	0.12	0.999	0.945	0.48	0.998	1.445*	0.32	0.998	1.282
FRED-MD + $D_{3,t}$	**11.11**	**0.943**	3.197***	**10.77**	**0.945**	2.740***	**10.92**	**0.944**	2.878***	**11.24**	**0.942**	2.969***

This table presents the out-of-sample forecast performance across different horizons, $h = 1, 3, 6, 9$. The metrics $R^2_{oos}(\%)$, $RMSFE$, and CW represent the percentage of out-of-sample R^2, the root mean squared forecast error, and the Clark and West (2007) test statistic, respectively. Bold figures indicate the best-performing model in terms of R^2_{oos} and $RMSFE$ for each horizon. Asterisks denote significance levels: * for 10%, ** for 5%, and *** for 1%.

tive performance of the proposed model against an alternative benchmark specification.

Additional robustness checks covering an alternative updating scheme, performance across different phases of the business cycle, and forecast accuracy during the COVID-19 recession are provided in Appendix A.

7.1. Alternative Benchmark Model

This section introduces an alternative benchmark model that uses the predictive power of the lagged target variable alongside monthly data from the news-based equity market volatility (EMV) indices. These indices, constructed by Baker et al. (2019), utilize scaled frequency counts of articles from eleven major US newspapers[18] that mention terms associated with economic, market, and volatility themes: E: economic, economy, financial, M: "stock market", equity, equities, "Standard and Poors" (and variants), and V: volatility, volatile, uncertain, uncertainty, risk, risky.

[18]The eleven major US newspapers include: 'the Boston Globe, Chicago Tribune, Dallas Morning News, Houston Chronicle, Los Angeles Times, Miami Herald, New York Times, San Francisco Chronicle, USA Today, Wall Street Journal, and Washington Post'.

For this analysis, I use all 45 variables under the general-economic and policy-related categories of the EMV indices to generate forecasts. The forecasting model is specified as follows:

$$r_{t+h} = \alpha + \eta r_t + \Psi' E_t + \varepsilon_{t+h}, \quad h = 1, 3, 6, 9 \tag{12}$$

where r_{t+h} denotes the return at time $t+h$, h represents the forecast horizon (1, 3, 6, or 9 months ahead). The intercept of the model is denoted by α, while η captures the influence of the lagged target variable r_t. The term E_t denotes the features extracted from the equity market volatility indices, with Ψ as its associated coefficient matrix. and ε_{t+h} represents the forecast error at time $t+h$.

Employing the attention mechanism alongside the Sahm rule updating scheme, I generate common factors and corresponding forecasts. Table 9 reports the corresponding out-of-sample results. The reported R^2_{oos}, $RMSFE$, and CW values align with the interpretation provided in Table 6.

This additional analysis confirms that the results obtained in the primary investigation are robust and not merely an artifact of a particular choice of benchmark. It adds further weight to the central findings of the study and underscores the efficacy of the attention mechanism and pattern validation in enhancing text-based forecasting models.

1. U.S. oil prices tumbled below $ 15 a barrel after Saudi Arabia indicated it was unwilling to cut output or resume its role as the oil cartel's swing producer.
By Michael Siconolfi – *Wall Street Journal, Nov 25, 1986*

2. OPEC fears further drop in oil price unless 2nd-quarter output is curbed. OPEC's oil ministers opened crucial talks on their production levels with a near consensus that the price of crude oil will fall further if they don't reduce second-quarter output.
By James Tanner – *Wall Street Journal, Mar 25, 1994*

3. Crude oil prices dropped after reports that Israel and Hamas militants were nearing a cease-fire deal amid worries the conflict would block supplies from the region.
By David Bird – *Wall Street Journal, Nov 21, 2012*

4. Futures finished Wednesday $ 1.43 or 1.5% higher on the day at $ 98.67 a barrel on the New York Mercantile Exchange. In the minutes after the report's release, oil prices surged as high as $ 99.25.
By Jerry A. Dicolo – *Wall Street Journal, Dec 22, 2011*

Figure 10: Illustrative news excerpts featuring key verb-noun and noun-verb collocations.

Table 9: Out-of-Sample Forecast Evaluation with an Alternative Benchmark

	$h=1$			$h=3$			$h=6$			$h=9$		
	$R^2_{oos}(\%)$	$RMSFE$	CW	$R^2_{oos}(\%)$	$RMSFE$	CW	$R^2_{oos}(\%)$	$RMSFE$	CW	$R^2_{oos}(\%)$	$RMSFE$	CW
Sahm Rule												
News Articles												
$D_{1,t}$	0.71	0.996	1.290*	2.61	0.987	0.832	6.93	0.965	0.853	8.71	0.955	1.081
$D_{2,t}$	1.36	0.993	1.003	-4.84	1.024	0.044	2.17	0.989	0.723	2.43	0.988	0.818
$D_{3,t}$	**10.41**	**0.947**	2.577***	**10.29**	**0.947**	2.027**	**14.01**	**0.927**	1.700**	**14.62**	**0.924**	1.797**
$D_{4,t}$	2.01	0.990	0.907	0.01	1.000	0.456	4.48	0.977	0.781	6.02	0.969	0.954
News Headlines												
$D_{1,t}$	0.48	0.998	0.563	-4.49	1.022	-0.345	1.56	0.992	0.525	4.48	0.977	0.660
$D_{2,t}$	-0.67	1.003	0.200	-3.38	1.017	-0.159	2.54	0.987	0.618	3.34	0.983	0.748
$D_{3,t}$	2.30	0.988	0.731	-2.00	1.010	0.429	3.50	0.982	0.853	3.37	0.983	0.913
$D_{4,t}$	0.88	0.996	0.662	-1.51	1.008	0.232	4.12	0.979	0.753	4.63	0.977	0.847
Continuous												
News Articles												
$D_{1,t}$	1.74	0.991	1.727**	0.57	0.997	0.859	6.30	0.968	1.452*	11.28	0.942	1.239
$D_{2,t}$	-1.69	1.008	0.502	-10.27	1.050	-0.559	-2.46	1.012	0.365	-0.50	1.002	0.565
$D_{3,t}$	**9.35**	**0.952**	2.472***	**8.12**	**0.959**	1.813**	**13.06**	**0.932**	1.671**	**14.31**	**0.926**	1.827**
$D_{4,t}$	0.13	0.999	0.562	-1.89	1.009	0.247	4.51	0.977	0.866	5.01	0.975	0.909
News Headlines												
$D_{1,t}$	-0.57	1.003	0.461	-3.53	1.017	0.093	-0.36	1.002	0.421	6.73	0.966	0.915
$D_{2,t}$	-2.03	1.010	0.061	-3.13	1.016	-0.105	3.01	0.985	0.680	4.10	0.979	0.768
$D_{3,t}$	2.46	0.988	1.306*	-2.62	1.013	0.311	2.18	0.989	0.746	4.79	0.976	0.993
$D_{4,t}$	1.57	0.992	0.879	-1.14	1.006	0.257	4.73	0.976	0.855	5.36	0.973	0.965

This table presents the out-of-sample forecast performance of models, $D_{1,t}$ to $D_{4,t}$, across different horizons, $h = 1, 3, 6, 9$. The metrics $R^2_{oos}(\%)$, $RMSFE$, and CW represent the percentage of out-of-sample R^2, the root mean squared forecast error, and the Clark and West (2007) test statistic, respectively. Bold figures indicate the best-performing model in terms of R^2_{oos} and $RMSFE$ for each horizon. Asterisks denote significance levels: * for 10%, ** for 5%, and *** for 1%.

8. Conclusion

This paper proposes a method to improve the interpretability and predictive performance of text data in forecasting crude oil prices using prominent historical new articles related to crude oil. Two methods are employed, namely CPV and attention mechanism. A summary of the findings obtained from the research is as follows.

Models augmented with text-based features, specifically those incorporating validated collocations of verb-noun and noun-verb patterns, consistently outperformed baseline models that lacked these textual insights. The findings were robust across different forecasting horizons and economic conditions, emphasizing the dynamic utility of textual data in capturing market sentiments and responses to global events.

Incorporating attention mechanisms and CPV processes proved instrumental in refining the selection and weighting of features, thereby enhancing the model's predictive accuracy. These techniques allowed for a more detailed interpretation and utilization of the information embedded in news text, leading to more precise and reliable forecasts.

The combination of macroeconomic indicators with text-based features led to a noticeable improvement in forecasting performance, suggesting that the hybrid approach captures a broader spectrum of influencing factors. This integrative approach leverages both structured economic data and unstructured textual information, providing a comprehensive view of the market dynamics.

While this study has shown that CPV and the attention mechanism are ways of enhancing text-based forecasting models, there remains a

vast potential for extensions. Future research can explore other forms of CPV, such as adjective-adjective combinations like "volatile unpredictable", "light sweet", or verb-verb combinations like "help stabilize", that may possess deeper insights that can unravel market dynamics. While this study focuses on bigrams, exploring higher order n-grams could further enhance model performance. Moreover, the scope of this study is limited to the English language; hence, extending the research to investigate pattern validation in other languages could significantly increase its impact and reach, offering a more global perspective. Furthermore, integrating other data sources, such as social media sentiments or expert opinions, to further augment forecasting accuracy. Future work could explore hybrid frameworks that combine the broad semantic understanding of large language models with our syntactic validation and attention mechanisms. By showcasing the benefits of integrating pattern validation for collocations and attention mechanism, this research highlights the need to evolve and innovate in the domain of text-based forecasting.

Acknowledgements

I am grateful to the Editor-in-Chief, Pierre Pinson, the Handling Editor, George Kapetanios, and the referees for their insightful comments and suggestions, which substantially improved the paper. I also thank Andrew S. Hanson, Luiz R. Lima, and Mohammed Mohsin for valuable comments and discussions. Any remaining errors are my own.

Data and code availability

The reproducibility package (code, processed data, and instructions) is available at: Github. The news articles used in this study are accessed via ProQuest and are subject to license restrictions; users will need institutional or personal access to ProQuest to retrieve the raw text.

References

Ahn, Seung C and Alex R Horenstein (2013) "Eigenvalue ratio test for the number of factors," *Econometrica*, 81 (3), 1203–1227.

An, Wuyue, Lin Wang, and Yu-Rong Zeng (2023) "Text-based soybean futures price forecasting: A two-stage deep learning approach," *Journal of Forecasting*, 42 (2), 312–330.

Aprigliano, Valentina, Simone Emiliozzi, Gabriele Guaitoli, Andrea Luciani, Juri Marcucci, and Libero Monteforte (2023) "The power of text-based indicators in forecasting Italian economic activity," *International Journal of Forecasting*, 39 (2), 791–808.

Aruoba, S Boragan and Thomas Drechsel (2022) "Identifying monetary policy shocks: A natural language approach."

Bai, Jushan and Serena Ng (2005) "Tests for skewness, kurtosis, and normality for time series data," *Journal of Business & Economic Statistics*, 23 (1), 49–60.

——— (2008) "Forecasting economic time series using targeted predictors," *Journal of Econometrics*, 146 (2), 304–317.

Bai, Yun, Xixi Li, Hao Yu, and Suling Jia (2022) "Crude oil price forecasting incorporating news text," *International Journal of Forecasting*, 38 (1), 367–383.

Bailliu, Jeannine, Xinfen Han, Mark Kruger, Yu-Hsien Liu, and Sri Thanabalasingam (2019) "Can media and text analytics provide insights into labour market conditions in China?" *International Journal of Forecasting*, 35 (3), 1118–1130.

Baker, Scott R, Nicholas Bloom, Steven J Davis, and Kyle J Kost (2019) "Policy news and stock market volatility,"Technical report, National Bureau of Economic Research.

Benamara, Farah, Carmine Cesarano, Antonio Picariello, Diego Reforgiato Recupero, and Venkatramana S Subrahmanian (2007) "Sentiment analysis: Adjectives and adverbs are better than adjectives alone.," *ICWSM*, 7, 203–206.

Calomiris, Charles W, Nida Çakır Melek, and Harry Mamaysky (2021) "Predicting the oil market,"Technical report, National Bureau of Economic Research.

Campbell, John Y and Samuel B Thompson (2008) "Predicting excess stock returns out of sample: Can anything beat the historical average?" *The Review of Financial Studies*, 21 (4), 1509–1531.

Clark, Todd E and Kenneth D West (2007) "Approximately normal tests for equal predictive accuracy in nested models," *Journal of econometrics*, 138 (1), 291–311.

Ellingsen, Jon, Vegard H Larsen, and Leif Anders Thorsrud (2022) "News media versus FRED-MD for macroeconomic forecasting," *Journal of Applied Econometrics*, 37 (1), 63–81.

Gonçalves, Sílvia, Michael W McCracken, and Benoit Perron (2017) "Tests of equal accuracy for nested models with estimated factors," *Journal of Econometrics*, 198 (2), 231–252.

Handlan, Amy (2020) "Text shocks and monetary surprises: Text analysis of fomc statements with machine learning," *Published Manuscript*.

Hoerl, Arthur E and Robert W Kennard (1970) "Ridge regression: Biased estimation for nonorthogonal problems," *Technometrics*, 12 (1), 55–67.

Jiao, Xingrui, Yuping Song, Yang Kong, and Xiaolong Tang (2022) "Volatility forecasting for crude oil based on text information and deep learning PSO-LSTM model," *Journal of Forecasting*, 41 (5), 933–944.

Justeson, John S and Slava M Katz (1995) "Technical terminology: some linguistic properties and an algorithm for identification in text," *Natural language engineering*, 1 (1), 9–27.

Kalamara, Eleni, Arthur Turrell, Chris Redl, George Kapetanios, and Sujit Kapadia (2022) "Making text count: economic forecasting using newspaper text," *Journal of Applied Econometrics*, 37 (5), 896–919.

Khong, Wai-Howe, Lay-Ki Soon, Hui-Ngo Goh, and Su-Cheng Haw (2018) "Leveraging part-of-speech tagging for sentiment analysis in short texts and regular texts," in *Semantic Technology: 8th Joint International Conference, JIST 2018, Awaji, Japan, November 26–28, 2018, Proceedings 8*, 182–197, Springer.

Lei, Bolin, Zhengdi Liu, and Yuping Song (2021) "On stock volatility forecasting based on text mining and deep learning under high-frequency data," *Journal of Forecasting*, 40 (8), 1596–1610.

Li, Xuerong, Wei Shang, and Shouyang Wang (2019) "Text-based crude oil price forecasting: A deep learning approach," *International Journal of Forecasting*, 35 (4), 1548–1560.

Li, Yelin, Hui Bu, Jiahong Li, and Junjie Wu (2020) "The role of text-extracted investor sentiment in Chinese stock price prediction with the enhancement of deep learning," *International Journal of Forecasting*, 36 (4), 1541–1562.

Lima, Luiz Renato and Lucas Lúcio Godeiro (2023) "Equity-premium prediction: Attention is all you need," *Journal of Applied Econometrics*, 38 (1), 105–122.

Lima, Luiz Renato, Lucas Lúcio Godeiro, and Mohammed Mohsin (2021) "Time-varying dictionary and the predictive power of FED minutes," *Computational Economics*, 57, 149–181.

Loughran, Tim, Bill McDonald, and Ioannis Pragidis (2019) "Assimilation of oil news into prices," *International Review of Financial Analysis*, 63, 105–118.

Manning, Christopher D (2011) "Part-of-speech tagging from 97% to 100%: is it time for some linguistics?" in *International conference on intelligent text processing and computational linguistics*, 171–189, Springer.

McCracken, Michael W and Serena Ng (2016) "FRED-MD: A monthly database for macroeconomic research," *Journal of Business & Economic Statistics*, 34 (4), 574–589.

Miao, Hong, Sanjay Ramchander, Tianyang Wang, and Dongxiao Yang (2017) "Influential factors in crude oil price forecasting," *Energy Economics*, 68, 77–88.

Morales-Arias, Leonardo and Guilherme V Moura (2013) "Adaptive forecasting of exchange rates with panel data," *International Journal of Forecasting*, 29 (3), 493–509.

Nicholls, Chris and Fei Song (2009) "Improving sentiment analysis with part-of-speech weighting," in *2009 International Conference on Machine Learning and Cybernetics*, 3, 1592–1597, IEEE.

Nonejad, Nima (2020) "A detailed look at crude oil price volatility prediction using macroeconomic variables," *Journal of Forecasting*, 39 (7), 1119–1141.

Nowak, Adam and Patrick Smith (2017) "Textual analysis in real estate," *Journal of Applied Econometrics*, 32 (4), 896–918.

Ochs, Adrian CR (2021) "A New Monetary Policy Shock with Text Analysis."

Salton, Gerard and Chung-Shu Yang (1973) "On the specification of term values in automatic indexing," *Journal of documentation*.

Schneider, Matthew J and Sachin Gupta (2016) "Forecasting sales of new and existing products using consumer reviews: A random projections approach," *International Journal of Forecasting*, 32 (2), 243–256.

Semiromi, Hamed Naderi, Stefan Lessmann, and Wiebke Peters (2020) "News will tell: Forecasting foreign exchange rates based on news story events in the economy calendar," *The North American Journal of Economics and Finance*, 52, 101181.

Song, Minchae and Kyung-shik Shin (2019) "Forecasting economic indicators using a consumer sentiment index: Survey-based versus text-based data," *Journal of forecasting*, 38 (6), 504–518.

Thorsrud, Leif Anders (2020) "Words are the new numbers: A newsy coincident index of the business cycle," *Journal of Business & Economic Statistics*, 38 (2), 393–409.

Tibshirani, Robert (1996) "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58 (1), 267–288.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin (2017) "Attention is all you need," *Advances in neural information processing systems*, 30.

Wu, Binrong, Lin Wang, Sheng-Xiang Lv, and Yu-Rong Zeng (2021) "Effective crude oil price forecasting using new text-based and big-data-driven model," *Measurement*, 168, 108468.

Yadav, Anita, CK Jha, Aditi Sharan, and Vikrant Vaish (2020) "Sentiment analysis of financial news using unsupervised approach," *Procedia Computer Science*, 167, 589–598.

Zhang, Weiguo, Xue Gong, Chao Wang, and Xin Ye (2021) "Predicting stock market volatility based on textual sentiment: A nonlinear analysis," *Journal of Forecasting*, 40 (8), 1479–1500.

Zhang, Yaojie, Feng Ma, and Yudong Wang (2019) "Forecasting crude oil prices with a large set of predictors: Can LASSO select powerful predictors?" *Journal of Empirical Finance*, 54, 97–117.

Zhang, Zhikai, Mengxi He, Yaojie Zhang, and Yudong Wang (2022) "Geopolitical risk trends and crude oil price predictability," *Energy*, 258, 124824.

Zheng, Tingguo, Xinyue Fan, Wei Jin, and Kuangnan Fang (2024) "Words or numbers? Macroeconomic nowcasting with textual and macroeconomic data," *International Journal of Forecasting*, 40 (2), 746–761.

Zou, Hui and Trevor Hastie (2005) "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 67 (2), 301–320.

Appendix A.

Table A.10: Descriptive Statistics of Parts of Speech in News Articles and Headlines

News Articles	Count	Mean	Standard Deviation	Minimum	25%	50%	75%	Maximum
Nouns	420	15,589.82	10,553.15	430	8,303.75	13,971.50	20,451	71,416
Verbs	420	6,053.56	4,109.42	202	3,330.75	5,396.50	7,840	27,491
Adjectives	420	5,667.36	3,736.97	183	3,154.25	5,141	7,241	25,074
Adverbs	420	1,381.03	934.86	44	745.50	1,231.50	1,787.50	6,476
News Headlines								
Nouns	420	209.98	143.66	6	103.75	190.50	271.25	915
Verbs	420	74.89	49.46	0	39	67	95	297
Adjectives	420	59.55	39.42	0	30.75	54	75	248
Adverbs	420	10.16	7.42	0	5	9	14	40

This table presents the descriptive statistics for parts of speech counts, separated into nouns, verbs, adjectives, and adverbs, within the corpus. The statistics showcase the mean, standard deviation, minimum and maximum counts, as well as the quartile distributions for each part of speech.

Table A.11: Descriptive Statistics of P.O.S of Collocations in News Articles and Headlines

News Articles	Count	Mean	Standard Deviation	Minimum	25%	50%	75%	Maximum
Noun-Noun	420	4,945.18	3,409.31	140	2,492.75	4,376.50	6,619	23,477
Adjective-Noun	420	4,021.77	2,670.26	123	2,253.0	3,660.50	5,197.50	18,131
Verb-Noun	420	1,843.58	1,287.28	56	961.50	1,634.50	2,409.50	8,863
Noun-Verb	420	2,000.52	1,275.63	67	1,101.0	1,783	2,549.50	8,492
Noun-Adjective	420	1,832.75	1,199.47	42	1,019.25	1,674.50	2,328.25	8,070
Verb-Adjective	420	1,908.09	1,286.07	67	1,052.0	1,701	2,457	8,743
News Headlines								
Noun-Noun	420	58.33	42.20	0	28	52	76	299
Adjective-Noun	420	45.30	30.48	0	22.75	41	58	193
Verb-Noun	420	25.26	17.70	0	13	22	34	111
Noun-Verb	420	10.21	8.01	0	4	9	14	43
Noun-Adjective	420	20.53	14.39	0	10	19	26	95
Verb-Adjective	420	21.67	14.95	0	11	19	27	96

This table presents the descriptive statistics for parts of speech counts of collocations, within the corpus. The statistics showcase the mean, standard deviation, minimum and maximum counts, as well as the quartile distributions for each part of speech.

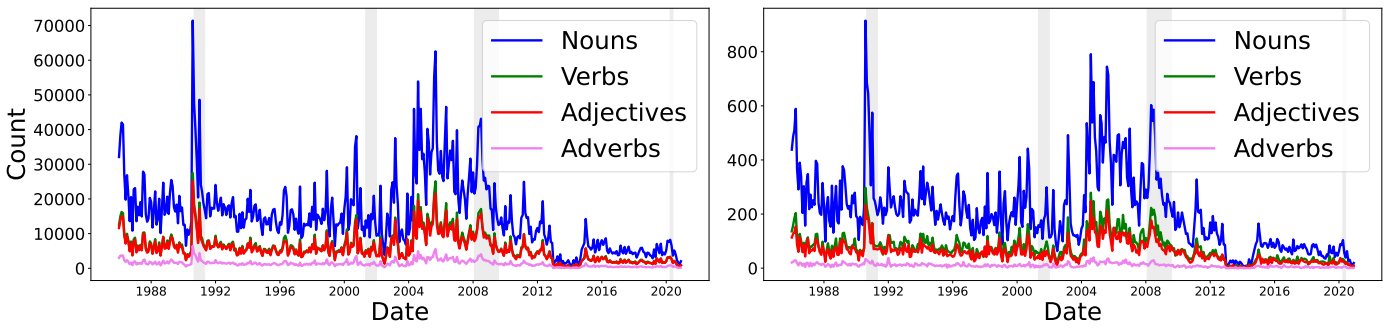

Figure A.11: The graph shows a time series plot of POS in articles (left) and headlines (right) from January 1986 to December 2020 within the corpus. The shaded regions represents NBER-defined recessions.

Algorithm 1: Text-Based Crude Oil Price Forecasting Algorithm

Input:

- Historical crude oil returns r_t for $t = 1, 2, \ldots, T$
- Raw text data (news articles/headlines) for $t = 1, 2, \ldots, T$

Output: Forecasts $r_{T+1}, r_{T+2}, \ldots, r_{T+H}$ for horizon H

Step 1: Construct Document-Term Matrix D_t

1.1 Preprocess text (clean, tokenize, POS tag)
1.2 Extract collocations (bigrams) with TF-IDF $\geqslant 0.01$
1.3 Validate collocations using POS patterns (N-N/A-N, V-N/N-V, N-A/V-A)
1.4 Create four document-term matrices $D_{1,t}$-$D_{4,t}$ (Table 4)

Step 2: Feature Selection (Time-Varying Dictionary)

2.1 For each forecast origin t:
  a. Update dictionary when Sahm indicator $\geqslant 0.5\%$ (threshold parameter)
  b. Estimate elastic net with 5-fold CV to select α, ρ (tuning parameters)
  c. Retain features with non-zero coefficients (Figs. 4-5)

Step 3: Factor Estimation

3.1 Apply PCA to selected features $D^*_{i,t}$
3.2 Determine optimal factors r^* using Bai & Ng (2002) criteria ($r_{max} = 8$)
3.3 Extract common factors F_t (Table 5)

Step 4: Factor Selection

4.1 Estimate forecasting equation with all factors
4.2 Retain factors with p-value < 0.05 (significance threshold)

Step 5: Construct Forecasts

5.1 Re-estimate model with selected factors
5.2 Generate H-month ahead forecasts $\hat{r}_{T+1}, \ldots, \hat{r}_{T+H}$
5.3 Update training window and repeat Steps 2-5 recursively

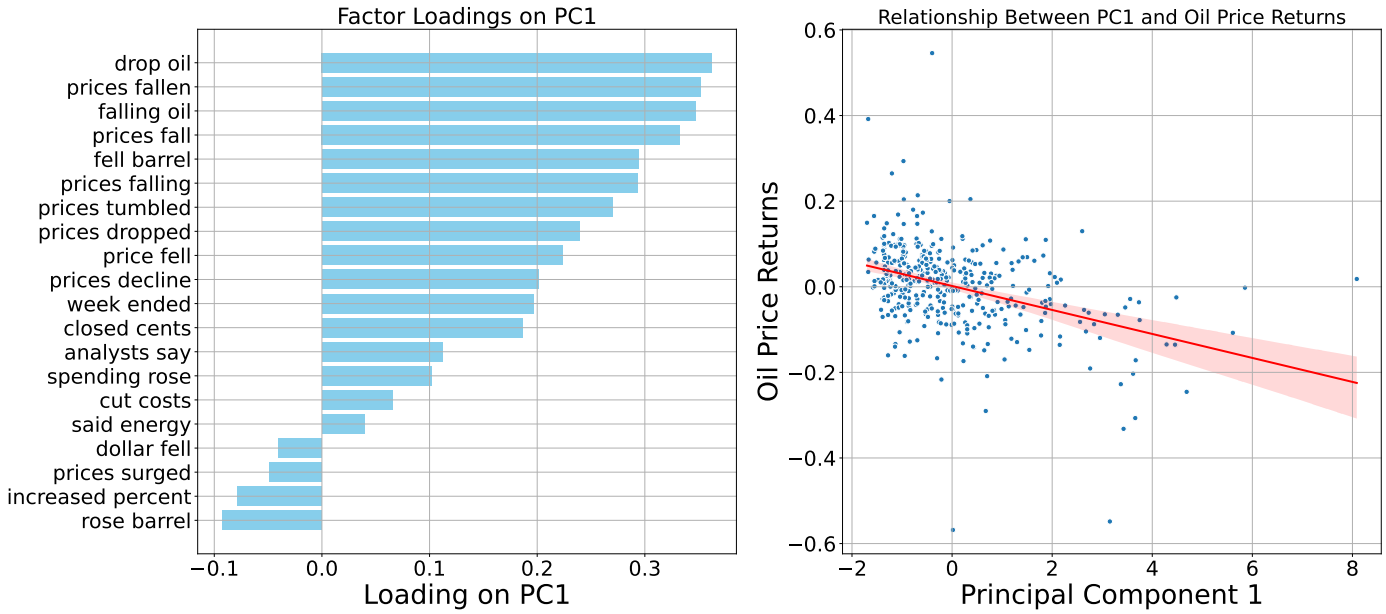

Figure A.12: The bar chart (left) shows the factor loadings on the first principal component (PC1), which primarily captures significant terms associated with declines in oil price returns, reflecting their contribution and influence on the component. The scatterplot (right) illustrates the relationship between PC1 scores and the monthly log difference of oil prices from January 1986 to December 2020, demonstrating a negative correlation where higher PC1 scores (indicative of stronger negative sentiment) are associated with decreases in oil price returns.

Table A.12: Out-of-Sample Forecast Evaluation Across Business Cycles

	$h=1$			$h=3$			$h=6$			$h=9$		
PANEL A: RECESSION	$R^2_{oos}(\%)$	$RMSFE$	CW	$R^2_{oos}(\%)$	$RMSFE$	CW	$R^2_{oos}(\%)$	$RMSFE$	CW	$R^2_{oos}(\%)$	$RMSFE$	CW
Sahm Rule												
News Articles												
$D_{1,t}$	1.34	0.993	0.678	-10.75	1.052	-1.128	-9.72	1.047	-1.539	-8.74	1.043	-0.971
$D_{2,t}$	6.55	0.967	0.950	-9.17	1.045	-1.391	-5.27	1.026	-0.633	-8.48	1.042	-1.011
$D_{3,t}$	**13.44**	**0.930**	1.214	**13.60**	**0.930**	1.255	**13.37**	**0.931**	1.289	**13.47**	**0.930**	1.294
$D_{4,t}$	1.54	0.992	0.877	1.81	0.991	1.024	0.21	0.999	0.315	1.63	0.992	0.758
News Headlines												
$D_{1,t}$	3.23	0.984	1.599*	-4.42	1.022	-1.023	-2.07	1.010	-0.374	-2.29	1.011	-0.425
$D_{2,t}$	-1.69	1.008	-0.556	-2.15	1.011	-0.559	-1.46	1.007	-0.300	-1.41	1.007	-0.252
$D_{3,t}$	1.21	0.994	0.440	-2.73	1.014	-1.405	-2.48	1.012	-1.273	-4.25	1.021	-1.264
$D_{4,t}$	1.70	0.991	0.654	1.11	0.994	0.477	1.05	0.995	0.435	0.64	0.997	0.303
Continuous												
News Articles												
$D_{1,t}$	5.11	0.974	0.973	-5.92	1.029	-1.042	-3.45	1.017	-0.299	-5.27	1.026	-0.339
$D_{2,t}$	3.38	0.983	0.606	-15.25	1.074	-2.448	-12.36	1.060	-2.206	-11.34	1.055	-1.688
$D_{3,t}$	**9.07**	**0.954**	1.210	**9.40**	**0.952**	1.279	**9.46**	**0.952**	1.250	**10.83**	**0.944**	1.411*
$D_{4,t}$	1.09	0.995	0.668	1.14	0.994	0.768	1.28	0.994	0.786	0.35	0.998	0.229
News Headlines												
$D_{1,t}$	-0.70	1.003	0.073	-1.78	1.009	0.063	-7.50	1.037	-0.994	-1.00	1.005	-0.086
$D_{2,t}$	-3.41	1.017	-1.396	0.84	0.996	0.733	0.94	0.995	0.976	0.36	0.998	0.298
$D_{3,t}$	6.41	0.967	1.895**	1.13	0.994	0.452	-1.36	1.007	-0.522	-0.02	1.000	0.030
$D_{4,t}$	4.57	0.977	1.749**	4.05	0.980	1.789**	4.01	0.980	1.707*	3.64	0.982	1.525*
PANEL B: EXPANSION												
Sahm Rule												
News Articles												
$D_{1,t}$	2.39	0.988	2.091**	15.32	0.920	1.841**	12.89	0.933	1.634*	13.72	0.929	1.859**
$D_{2,t}$	0.61	0.997	1.084	3.02	0.985	1.876**	2.65	0.987	1.526*	3.45	0.983	1.919**
$D_{3,t}$	**10.61**	**0.945**	2.856***	**12.96**	**0.933**	2.599***	**11.05**	**0.943**	2.501***	**10.79**	**0.945**	2.518***
$D_{4,t}$	4.23	0.979	3.296***	4.23	0.979	2.770***	3.23	0.984	2.869***	3.55	0.982	2.530***
News Headlines												
$D_{1,t}$	1.04	0.995	1.328*	0.80	0.996	1.455*	-0.09	1.000	0.836	3.28	0.983	1.029
$D_{2,t}$	1.90	0.990	1.818**	1.20	0.994	1.914**	1.07	0.995	1.822**	0.95	0.995	1.800**
$D_{3,t}$	4.84	0.975	1.564*	3.63	0.982	1.734**	3.19	0.984	1.724**	2.57	0.987	1.666**
$D_{4,t}$	2.44	0.988	1.917**	2.31	0.988	1.590*	2.17	0.989	1.581*	1.86	0.991	1.497*
Continuous												
News Articles												
$D_{1,t}$	1.96	0.990	1.938**	9.40	0.952	1.821**	8.31	0.958	3.409***	15.94	0.917	2.091**
$D_{2,t}$	-2.31	1.011	0.439	-1.75	1.009	0.763	-0.77	1.004	0.686	0.33	0.998	1.053
$D_{3,t}$	**11.33**	**0.942**	2.903***	**12.08**	**0.938**	2.493***	**11.76**	**0.939**	2.601***	**11.77**	**0.939**	2.661***
$D_{4,t}$	1.62	0.992	1.921**	1.72	0.991	1.854**	2.67	0.987	2.531***	2.64	0.987	2.117**
News Headlines												
$D_{1,t}$	1.52	0.992	1.328*	0.75	0.996	1.195	-0.09	1.000	0.948	6.18	0.969	1.594*
$D_{2,t}$	0.75	0.996	1.448*	-0.15	1.001	1.098	0.45	0.998	1.468*	1.18	0.994	1.566*
$D_{3,t}$	2.35	0.988	1.542*	0.45	0.998	1.166	0.43	0.998	1.211	2.49	0.987	1.541*
$D_{4,t}$	1.98	0.990	2.028**	1.17	0.994	1.581*	1.47	0.993	1.714**	1.36	0.993	1.692**

This table presents the out-of-sample forecast performance of models $D_{1,t}$ to $D_{4,t}$ across forecast horizons $h = 1, 3, 6, 9$. The metrics $R^2_{oos}(\%)$, $RMSFE$, and CW represent the percentage of out-of-sample R^2, the root mean squared forecast error, and the Clark and West (2007) test statistic, respectively. Bold figures indicate the best-performing model (highest R^2_{oos} and lowest $RMSFE$) within each scheme's News Articles block, per horizon. Asterisks denote significance on the CW statistic: * 10%, ** 5%, *** 1%.

Table A.13: Out-of-Sample Forecast Evaluation: Alternative Updating Scheme

	$h=1$			$h=3$			$h=6$			$h=9$		
Δ Oil Price	$R^2_{oos}(\%)$	$RMSFE$	CW	$R^2_{oos}(\%)$	$RMSFE$	CW	$R^2_{oos}(\%)$	$RMSFE$	CW	$R^2_{oos}(\%)$	$RMSFE$	CW
News Articles												
$D_{1,t}$	6.71	0.966	1.645*	6.13	0.969	1.567*	7.65	0.961	1.610*	8.39	0.957	1.767**
$D_{2,t}$	-3.55	1.018	-0.193	-5.40	1.027	-0.606	-5.58	1.028	-0.772	-2.25	1.011	0.142
$D_{3,t}$	**11.91**	**0.939**	2.809***	**10.87**	**0.944**	2.920***	**11.53**	**0.941**	2.983***	**11.31**	**0.942**	2.983***
$D_{4,t}$	0.73	0.996	1.524*	1.87	0.991	2.405***	2.32	0.988	2.614***	1.84	0.991	1.958**
News Headlines												
$D_{1,t}$	1.68	0.992	1.799**	-1.43	1.007	0.470	1.31	0.993	0.928	1.99	0.990	2.038**
$D_{2,t}$	0.49	0.998	1.488*	0.10	0.999	1.471*	-0.53	1.003	0.670	-0.81	1.004	0.383
$D_{3,t}$	0.51	0.997	1.295*	1.75	0.991	1.491*	1.03	0.995	1.458*	0.57	0.997	1.270
$D_{4,t}$	2.19	0.989	1.889**	1.87	0.991	1.551*	1.76	0.991	1.513*	1.42	0.993	1.380*
Continuous												
News Articles												
$D_{1,t}$	3.04	0.985	2.154**	3.79	0.981	1.600*	4.05	0.980	2.173**	8.30	0.958	1.693**
$D_{2,t}$	-0.35	1.002	0.749	-6.70	1.033	-0.808	-4.97	1.025	-0.666	-3.87	1.019	-0.208
$D_{3,t}$	**10.55**	**0.946**	3.094***	**11.10**	**0.943**	2.807***	**10.93**	**0.944**	2.871***	**11.43**	**0.941**	3.003***
$D_{4,t}$	1.44	0.993	2.030**	1.51	0.992	2.014**	2.17	0.989	2.618***	1.81	0.991	1.961**
News Headlines												
$D_{1,t}$	0.76	0.996	1.106	-0.18	1.001	0.862	-2.78	1.014	-0.199	3.59	0.982	1.409*
$D_{2,t}$	-0.68	1.003	0.699	0.21	0.999	1.305*	0.63	0.997	1.725**	0.88	0.996	1.555*
$D_{3,t}$	3.75	0.981	2.203**	0.70	0.996	1.250	-0.22	1.001	1.007	1.59	0.992	1.528*
$D_{4,t}$	2.87	0.986	2.653***	2.23	0.989	2.334**	2.39	0.988	2.380***	2.18	0.989	2.252**

This table presents the out-of-sample forecast performance of models, $D_{1,t}$ to $D_{4,t}$, across horizons $h = 1, 3, 6, 9$ where the model is refitted to update the dictionary when the change in oil price falls below the 10th percentile or rises above the 90th percentile. The metrics $R^2_{oos}(\%)$, $RMSFE$, and CW represent the percentage of out-of-sample R^2, the root mean squared forecast error, and the Clark and West (2007) test statistic, respectively. Bold figures indicate the best-performing model in terms of R^2_{oos} and $RMSFE$ for each horizon. Asterisks denote significance levels: * 10%, ** 5%, *** 1%.

Table A.14: COVID-19 Recession: Out-of-Sample Forecast Evaluation

	$h=1$			$h=3$			$h=6$			$h=9$		
	$R^2_{oos}(\%)$	$RMSFE$	CW	$R^2_{oos}(\%)$	$RMSFE$	CW	$R^2_{oos}(\%)$	$RMSFE$	CW	$R^2_{oos}(\%)$	$RMSFE$	CW
Sahm Rule												
News Articles												
$D_{1,t}$	5.11	0.974	2.886*	-11.72	1.057	-0.741	-11.64	1.057	-1.291	-12.31	1.060	-0.793
$D_{2,t}$	16.03	0.916	1.080	-3.73	1.019	-0.816	-3.22	1.016	-0.745	-2.80	1.014	-0.640
$D_{3,t}$	**17.78**	**0.907**	0.877	**18.36**	**0.904**	0.941	**16.55**	**0.914**	0.931	**16.66**	**0.913**	0.946
$D_{4,t}$	-2.88	1.014	-2.093	-2.48	1.012	-2.199	-1.30	1.006	-0.890	-2.36	1.012	-0.947
News Headlines												
$D_{1,t}$	3.15	0.984	0.888	-7.41	1.036	-1.000	-4.12	1.020	-0.390	-4.04	1.020	-0.387
$D_{2,t}$	-1.41	1.007	-0.372	1.39	0.993	1.412	1.41	0.993	1.413	1.44	0.993	1.404
$D_{3,t}$	-4.36	1.022	-2.537	-5.19	1.026	-2.801	-5.58	1.028	-3.611	-8.15	1.040	-1.846
$D_{4,t}$	-5.11	1.025	-2.687	-5.44	1.027	-2.557	-5.47	1.027	-2.561	-5.56	1.027	-2.566
Continuous												
News Articles												
$D_{1,t}$	**13.85**	**0.928**	1.606	-0.69	1.003	-1.000	0.34	0.998	0.092	-1.97	1.010	-0.006
$D_{2,t}$	13.45	0.930	0.864	-4.67	1.023	-1.166	-5.08	1.025	-1.521	-4.93	1.024	-1.536
$D_{3,t}$	11.49	0.941	0.807	**11.61**	**0.940**	0.890	**11.64**	**0.940**	0.920	**12.15**	**0.937**	1.006
$D_{4,t}$	-2.69	1.013	-1.894	-2.00	1.010	-1.722	-2.07	1.010	-1.834	-3.53	1.018	-1.744
News Headlines												
$D_{1,t}$	-0.07	1.000	–	-7.41	1.036	-1.000	-4.16	1.021	-0.399	-7.43	1.036	-1.000
$D_{2,t}$	-2.85	1.014	-1.000	0.00	1.000	–	0.92	0.995	1.763	-1.40	1.007	-0.487
$D_{3,t}$	0.74	0.996	0.478	-2.11	1.011	-1.000	-4.15	1.021	-2.429	-1.60	1.008	-1.000
$D_{4,t}$	0.00	1.000	–	0.00	1.000	–	0.00	1.000	–	0.00	1.000	–

This table presents the out-of-sample (OOS) forecast performance of models $D_{1,t}$ to $D_{4,t}$ across forecast horizons $h = 1, 3, 6, 9$ during the COVID-19 recession. The metrics $R^2_{oos}(\%)$, $RMSFE$, and CW represent the percentage of out-of-sample R^2, the root mean squared forecast error, and the Clark and West (2007) test statistic, respectively. Bold figures indicate the best-performing model (highest R^2_{oos} and lowest $RMSFE$) within each scheme's News Articles block, per horizon. Asterisks denote significance levels on the CW statistic: * 10%, ** 5%, *** 1%.