

Research Repository

Forecasting Oil Volatility Direction with News: a Language Model and Regime-Aware SHAP Approach

Accepted for publication in Finance Research Letters

Research Repository link: <https://repository.essex.ac.uk/43805/>

Please note:

Changes made as a result of publishing processes such as copy-editing, formatting and page numbers may not be reflected in this version. For the definitive version of this publication, please refer to the published source. You are advised to consult the published version if you wish to cite this paper.

<https://doi.org/10.1016/j.frl.2026.110673>

Forecasting Oil Volatility Direction with News: a Language Model and Regime-Aware SHAP Approach

Felipe Maldonado*¹ and Romina Hashami¹

Abstract

This paper investigates whether financial news can forecast the direction of crude oil price volatility. While most existing studies focus on predicting volatility levels based on historical econometric dynamics, we examine the economically relevant task of anticipating directional shifts. Using approximately 590,000 Reuters oil-market headlines from 2014–2023, we construct daily textual predictors based on news volume and language-model embeddings and benchmark them against a directional heterogeneous autoregressive (HAR) model. We find that news volume and subword-level FastText embeddings significantly outperform the benchmark in out-of-sample forecasts. In contrast, complex transformer-based models (e.g., Gemini) do not yield statistically significant improvements, suggesting that lexical intensity is more informative than deep semantic context in short financial headlines. Structural breaks are identified using the Pruned Exact Linear Time (PELT) algorithm, and SHAP (SHapley Additive exPlanations) values are anchored to these regimes to reveal time-varying volatility drivers. Overall, our results show that news-based indicators provide forward-looking signals of volatility direction and that the importance of these signals varies across market regimes.

Keywords: Crude oil volatility, News-based forecasting, Directional classification, SHAP explanations, HAR model

1. Introduction

Crude oil volatility plays a central role in global financial stability, portfolio allocation, and risk management. As the dominant international benchmark, Brent crude oil significantly influences derivative pricing, investment decisions, and risk management across energy and financial sectors. For mar-

ket participants and policymakers alike, anticipating changes in volatility is therefore as important as forecasting price levels themselves (Herrera et al., 2018; Liu et al., 2019). In practice, the direction of volatility, i.e., whether realized volatility increases or decreases relative to the previous period, is often associated with hedging intensity and capital allocation decisions. Even when the precise magnitude of future volatility remains uncertain, directional signals can still influence risk management decisions (He et al., 2022).

In this paper, we address a set of open questions in this literature by developing a regime-aware forecasting framework that combines language-model embeddings with SHAP (SHapley Additive exPlanations) (Lundberg and Lee, 2017) to predict directional changes in volatility based purely on textual data. Using Reuters oil-market headlines, we construct daily textual representations using language-model embeddings, forecast the daily direction of realized volatility, and benchmark these forecasts against a directional HAR model. To ensure that interpretability reflects genuine structural shifts in volatility dynamics, we determine structural breaks using the Pruned Exact Linear Time (PELT) (Killick et al., 2012) algorithm and anchor SHAP to the resulting regimes. Our contributions to the literature are as follows: (i) we provide systematic evidence that textual news alone can forecast the direction of oil-market volatility; (ii) we show that subword-based representations outperform state-of-the-art transformer models for this specific forecasting task; (iii) we introduce a regime-aware interpretability framework that combines PELT break detection with SHAP attributions; (iv) we establish an interpretable framework for real-time volatility risk monitoring based on high-frequency textual information.

Most existing studies in the literature model oil volatility using econometric frameworks that exploit persistence in historical returns, including GARCH-type models, realized-volatility regressions, and heterogeneous autoregressive (HAR) specifications (Corsi, 2009). These models remain strong empirical benchmarks due to the well-documented persistence of volatility. However, recent evidence suggests that oil-volatility dynamics may become nonlinear and regime-dependent during periods characterized by geopolitical stress and structural disruptions (Liu et al., 2019; Zhang and Guo, 2024). Recent extensions have sought to enhance these baseline models by incorporating external signals such as energy-related uncertainty indexes within GARCH-MIDAS and mixed-frequency frameworks (Zhang and Guo, 2024). However, such approaches remain fundamentally backward-looking, relying heavily on lagged volatility realizations, as well as structured macroeconomic

inputs, which may limit their responsiveness to rapidly unfolding information shocks (Gkillas et al., 2020).

In response to these limitations, a growing body of research has turned toward machine learning (ML) techniques. Recent empirical work shows that ML variants of econometric models can outperform baseline HAR specifications for forecasting tasks, capturing nonlinear interactions and higher-order dependencies that linear models may overlook (Kim et al., 2026). Additional studies highlight that hybrid and deep learning architectures improve predictive accuracy in commodity markets, particularly under regime shifts and crisis episodes (Song et al., 2022; Ma et al., 2025). Building on this methodological shift, the literature has increasingly explored the informational content of textual data. News-based uncertainty measures and media sentiment indicators have been shown to enhance oil-volatility forecasting performance, especially during turbulent market conditions (Li et al., 2021, 2022). More recently, transformer-based language models have been employed to extract richer semantic representations from financial text and incorporate investor sentiment into volatility prediction frameworks (Song et al., 2025).

Despite these advances, several important questions remain unresolved. First, while textual indicators are frequently combined with historical volatility measures, it remains unclear whether news-based features alone provide independent predictive information. Second, although recent work increasingly adopts complex transformer-based models, there is limited empirical evidence on whether they systematically outperform simpler representations when applied to short, information-dense financial headlines. Third, while ML methods are now widely used for volatility forecasting, existing studies provide limited interpretability and rarely explain which specific informational drivers underpin volatility predictions or how these drivers evolve across market regimes.

The remainder of the paper is organized as follows. Section 2 describes the data and methodological framework. Section 3 presents the empirical results, including forecasting performance and regime-based SHAP interpretation. Section 4 provides a combined discussion and conclusions and outlines directions for future research.

2. Methodology

This study evaluates whether financial news can forecast the direction of realized volatility in Brent crude oil futures. We compare traditional econo-

metric forecasts based on the HAR model with models that rely solely on news-based features. The goal is to assess whether textual information provides incremental predictive value beyond past volatility dynamics.

2.1. Data collection

We collect Thomson Reuters (Eikon) news headlines from January 2014 to December 2023, resulting in approximately 590,000 unique items. Headlines are selected because they provide concise representations of market-relevant developments. The daily number of news items is recorded as a proxy for media attention and information flow (see Figure 1).

Daily headlines are preprocessed to remove duplicates, hyperlinks, and very short entries, and only the final version of each story published during the day is retained to avoid overweighting repeated updates. Additional details on aggregation of textual features are reported in Appendix A.

Figure 1 shows the monthly distribution of collected headlines, which aligns with major oil-market events such as OPEC meetings, the 2014–2016 price collapse, the COVID-19 pandemic, and the Russia–Ukraine conflict. To illustrate the lexical composition, Figure 2 displays the most frequent terms in the corpus, highlighting keywords such as “oil” and “prices”, which dominate the news coverage during high-volatility periods.

High-frequency Brent oil futures data are obtained from `barchart.com`. Figure 3 (left) depicts the daily closing price trajectory over the sample period, highlighted by the key geopolitical and economic events that drove market trends. To compute realized volatility, we follow Andersen and Bollerslev (1998), and Song et al. (2025) and define it as the sum of all available intraday high-frequency squared returns,

$$RV_t = \sum_{j=1}^M \left[\ln \frac{P_{t,j}}{P_{t,j-1}} \right]^2,$$

where $\ln \frac{P_{t,j}}{P_{t,j-1}}$ denotes the log-return between two consecutive intraday price observations (j and $j - 1$) at day t . M denotes the total number of observations. Figure 3 (right) illustrates the realized-volatility series, showing distinct spikes around major global events. This measure serves as a model-free benchmark for true price variability.

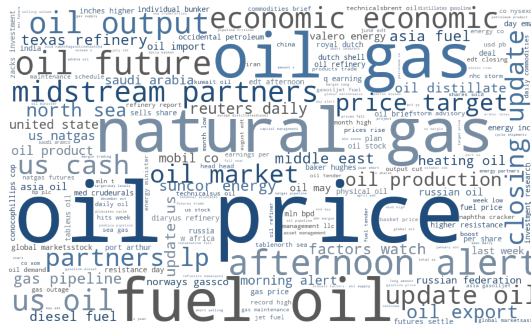
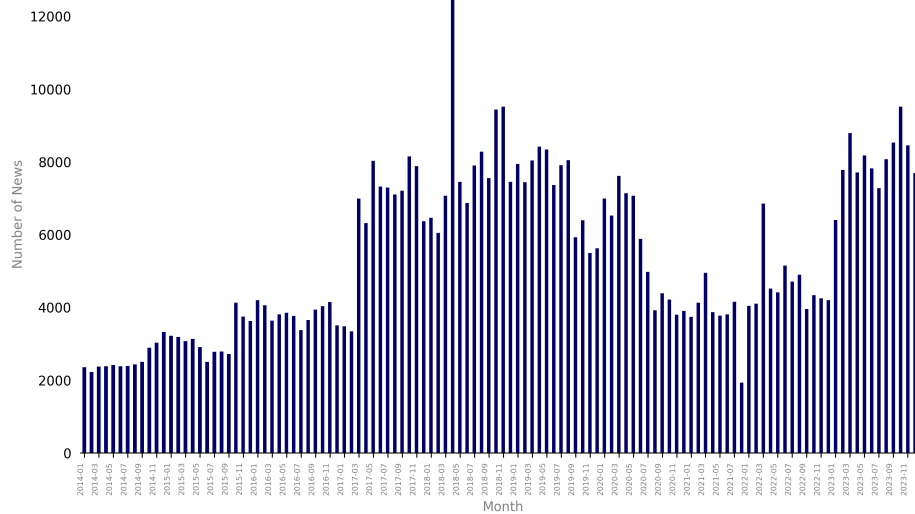


Figure 2: Word cloud of the most frequent terms in the headline corpus. Larger words indicate higher frequency and thematic importance.

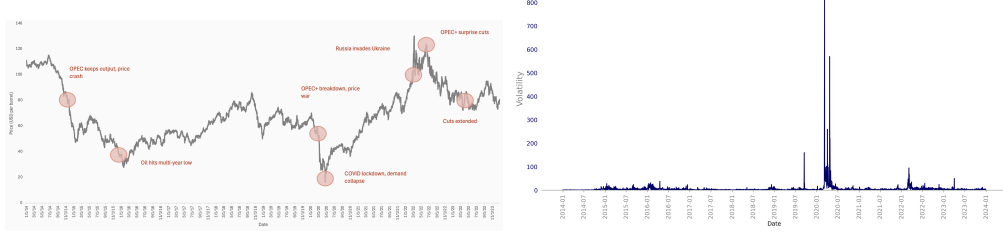


Figure 3: Brent crude oil price trend (2014–2023) annotated with structural breaks (left) and Realized Volatility computed from 5-minute returns (right).

2.2. Structural break detection using PELT

To formally identify structural breaks in oil market volatility, we employ the Pruned Exact Linear Time (PELT) algorithm of Killick et al. (2012). PELT is a multiple change-point detection method that partitions a time series into homogeneous segments by minimizing a penalized cost function, thereby balancing model fit against segmentation complexity. Compared to ad-hoc regime definitions or rolling-window approaches, PELT provides a data-driven and statistically principled framework for detecting regime shifts in financial time series.

We apply PELT to the daily log-realized volatility series of Brent crude oil futures. Following Killick et al. (2012), the algorithm minimizes the objective function:

$$\sum_{m=0}^L C(RV_{(\tau_m+1):\tau_{m+1}}) + \beta L,$$

where $C(\cdot)$ denotes the within-segment cost, τ_m are the estimated breakpoints, L is the number of detected breaks, and β is a penalty parameter to avoid overfitting.

We employ a radial basis function (RBF) cost, which allows for flexible detection of nonlinear changes in the distribution of volatility. We set $\beta = 20$, which yields a parsimonious segmentation of the sample into a small number of economically interpretable regimes and avoids over-fragmentation into short-lived volatility states. This choice produces stable break patterns that coincide with major oil-market episodes, including the 2014–2015 oil price collapse and the COVID-19 demand shock (Table 4). The resulting breakpoints are used to define volatility regimes that underpin the regime-dependent SHAP analysis in Section 3.3.

2.3. Text transformation framework

Each headline is transformed into numerical representations via an organized pipeline in order to model the informative content of news. Figure 4 summarizes the overall process: daily headlines are cleaned and pre-processed, transformed into numerical features via embedding models, and then used as predictors in ensemble classifiers that forecast volatility direction. Model interpretability is achieved through SHAP-based explanations of feature importance.

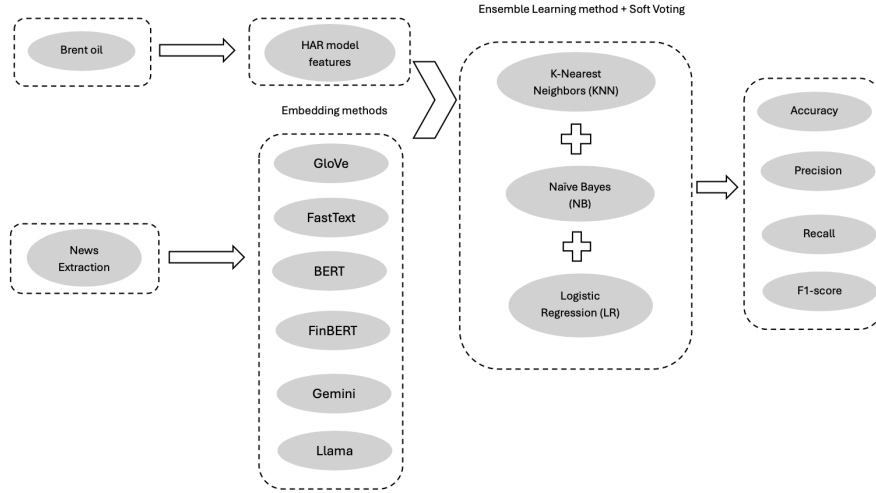


Figure 4: Workflow of the proposed news-based forecasting framework. Headlines are pre-processed, embedded using language models, and applied to ensemble classifiers. Model interpretability is obtained through SHAP analysis.

2.4. News-based features

Two types of text-based predictors are considered:

- **News volume:** number of articles published each day and how far it deviates from a moving average. Proxy for media attention.
- **Word embeddings:** dense vector representations from pre-trained models such as FastText (Bojanowski et al., 2017), GloVe (Pennington et al., 2014), FinText (<https://fintext.ai>), and Transformer-based encoders: FinBERT (Araci, 2019), LLaMA (Touvron et al., 2023), and Gemini (Anil et al., 2023).

Table 1 summarizes the key embedding models used to convert textual data into numerical representations.

Table 1: Embedding models used for news representation.

Model	Type	Vector dimension
GloVe	Static word embeddings	300
FastText	Subword-level embeddings	300
FinText	Financial-domain embeddings	300
FinBERT	Transformer-based contextual model	768
LLaMA	Large language model	4096
Gemini	Transformer-based embedding model	768

Notes: This table summarizes the embedding models used to represent Reuters oil-market headlines. “Vector dimension” denotes the dimensionality of the embedding returned by each model. For each trading day t , embeddings from all headlines published on day t are aggregated by simple averaging to obtain a daily feature vector $\mathbf{X}_t$.

2.5. Forecasting framework

The forecasting task is formulated as a binary classification of the direction of realized-volatility change – upward ($\Delta RV_t > 0$) or downward ($\Delta RV_t < 0$). Since the sign of volatility movements frequently contains more useful information than the exact magnitude does. Each news-based feature set is used to train an ensemble of standard classifiers, namely Logistic Regression, Naïve Bayes, and K-Nearest Neighbors. Model hyperparameters are optimized using time-series cross-validation within the training sample only, preventing data leakage. Forecasts are generated strictly in a rolling out-of-sample setting. The benchmark for comparison is a directional HAR model (Corsi, 2009), which forecasts volatility direction using past daily, weekly, and monthly realized-volatility components. Comparing news-only models with this benchmark isolates the incremental predictive content of textual information.

2.6. Evaluation and robustness checks

Accuracy, precision, recall, and F1-score are used to evaluate the model’s performance. When comparing out-of-sample accuracy differences between two models, robustness is evaluated using McNemar’s non-parametric test (McNemar, 1947). This test determines if performance gains are statistically significant, and focuses on inconsistent predictions. To assess whether

the predictive content of news merely proxies for aggregate macroeconomic uncertainty, we conduct an additional robustness check using the Global Economic Policy Uncertainty (EPU) index of Baker et al. (2016), for details we refer to Appendix B.

2.7. Model interpretation

We use SHAP values (Lundberg and Lee, 2017) to interpret the top-performing models in order to improve transparency. SHAP provides a framework for attributing the marginal contribution of each predictor to the model’s output, thereby offering transparent and locally consistent explanations to ML forecasts. By breaking down each prediction into additive feature contributions, SHAP measures the impact of each word or embedding dimension on the likelihood of an increase in volatility. By grouping influential tokens and expressions by market regime (e.g., pre/post COVID) we identify how news themes evolve as volatility drivers across time. This interpretability links text-based forecasts to economically meaningful narratives.

3. Results

3.1. Out-of-sample predictive performance

Table 2 reports the out-of-sample directional forecasting performance of all models, allowing a direct comparison between the traditional HAR benchmark and news-based textual predictors derived from different embedding methods. In line with earlier volatility-forecasting research that relied on lagged realized volatility, the HAR benchmark achieves a directional accuracy of 64.8%. News-based models offer a clear additive predictive value. The best text representations are the **Count** (news volume) and **FastText** models, which achieve accuracies of 69.7% and 71.1%, respectively. While GloVe (67.2%) and FinText (66.5%) produce modest gains over HAR, transformer-based encoders (FinBERT, LLaMA, Gemini) do not provide consistent improvements, indicating that increased model complexity does not necessarily translate into better predictive performance in this setting.

To evaluate whether the observed differences in predictive accuracy are statistically significant, we apply McNemar’s non-parametric test to compare discordant classification outcomes between models. Table 3 summarizes the corresponding p-values for the main pairwise model comparisons. Each “A vs. B” entry measures whether model A’s predictions outperform those from

model B. Low p-values imply statistically significant differences, while high p-values indicate comparable performance. We find that both *FastText* and *Count* significantly outperform the HAR benchmark at the 5% level, whereas other models do not. Additionally, FastText performs significantly better than FinText, GloVe, and FinBERT embeddings, demonstrating the strength of its predictive advantage.

Table 2: Out-of-sample directional forecasting performance. Best value in bold.

Model	Accuracy (%)	Precision	Recall	F1-score
HAR (benchmark)	64.8	0.63	0.65	0.64
Count	69.7	0.68	0.70	0.69
GloVe	67.2	0.66	0.67	0.66
FastText	71.1	0.70	0.72	0.71
FinText	66.5	0.66	0.66	0.66
FinBERT	65.1	0.64	0.65	0.64
LLaMA	64.3	0.63	0.64	0.63
Gemini	63.7	0.63	0.63	0.63

Notes: This table reports out-of-sample performance for daily forecasts of volatility *direction*, defined as an increase or decrease in realized volatility relative to the previous trading day. The evaluation sample covers the period 2014–2023. All models are estimated and evaluated using a rolling-window design. Accuracy is reported in percent, while precision, recall, and F1-score are computed for the “up” (volatility increase) class.

Table 3: McNemar test p -values for key model comparisons (out-of-sample forecasts).

Model comparison	p-value
GloVe vs. HAR	0.5782
FinText vs. HAR	0.0562
Count vs. HAR	0.0392
FastText vs. HAR	0.0127
FastText vs. FinText	0.0008
FastText vs. GloVe	0.0032
FastText vs. FinBERT	0.0104

Notes: This table reports p -values from McNemar’s test comparing pairs of classifiers based on their out-of-sample directional volatility forecasts over the period 2014–2023. The null hypothesis is equal predictive accuracy between the two models. Low p -values (for A vs. B) indicate statistically significant differences in classification performance (i.e. A better than B).

3.2. Structural breaks and volatility regimes

Figure 5 and Table 4 report structural breaks in daily log-realized volatility detected using the PELT algorithm (Killick et al., 2012) over the sample period 2014–2023. The procedure identifies six breakpoints that partition the sample into distinct volatility regimes, and their dates correspond closely to major oil-market episodes (Table 4). These statistically identified regimes provide a formal foundation for our regime-dependent SHAP analysis and support the interpretation that the predictive role of news varies systematically across volatility states.

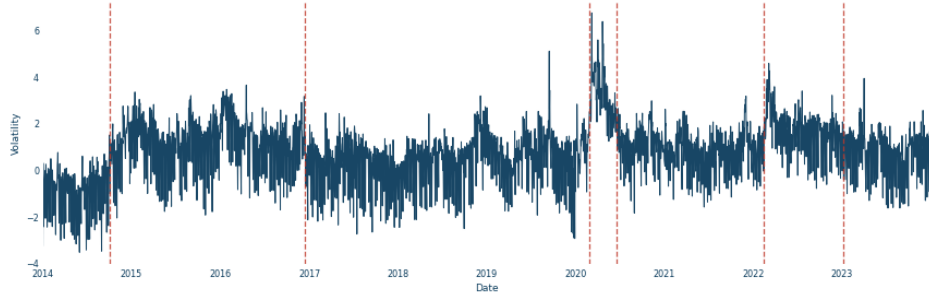


Figure 5: Structural breaks in log-realized volatility detected by PELT. This figure plots daily log-realized volatility of Brent crude oil futures over the sample period 2014–2023. Vertical dashed lines denote breakpoints detected by the PELT algorithm (Killick et al., 2012) using an RBF cost function. The estimated breaks coincide with major oil-market episodes, including the 2014–2015 oil price collapse, the COVID-19 demand shock, and the Russia–Ukraine conflict, indicating the presence of distinct volatility regimes.

Table 4: Structural break dates and associated market episodes detected by PELT. This table lists the breakpoints in daily log-realized volatility of Brent crude oil futures detected by PELT (Killick et al., 2012) using an RBF cost function over the sample period 2014–2023, together with their “Associated market episode” which provides a qualitative interpretation of each statistically detected regime shift.

Break date	Associated market episode
2014-10-02	2014–2015 oil price collapse
2016-12-14	OPEC-led market stabilization phase
2020-02-26	Onset of the COVID-19 demand shock
2020-06-17	COVID-19 stabilization and policy response period
2022-02-11	Russia–Ukraine conflict shock
2023-01-08	Post-war normalization and monetary tightening phase

3.3. SHAP-based interpretation

For expositional clarity and economic interpretability, using the PELT breakpoints, we aggregate subperiods into four broad macroeconomic regimes (see Figure 6). Then we computed the SHAP values for FastText model under those distinct market regimes. As illustrated in Figure 6, the most influential tokens identified by SHAP cluster around economic themes, including supply and production conditions, demand-side factors, and references to major oil-producing or consuming geographical regions. Importantly, the

composition and relative importance of these themes vary across volatility regimes (panels a–d), indicating that the information emphasized by the model adapts to changing market conditions. For example, earlier regimes place greater weight on terms associated with production and petroleum products, whereas later regimes emphasize geographically-related terminology. This regime-dependent structure would not be observable from predictive accuracy alone, and highlights the value of SHAP for understanding how textual information contributes to volatility forecasts.

4. Discussion and Conclusions

This study provides new evidence that financial news contains economically meaningful information for forecasting the *direction* of crude oil volatility, even in the absence of historical volatility inputs. Unlike most of the existing literature, which focuses on predicting volatility levels, our results demonstrate that textual information alone can anticipate whether volatility will rise or fall. This distinction is economically important because volatility direction influences hedging decisions, risk exposure, and responses to changing uncertainty in energy markets. This complements recent work demonstrating the usefulness of alternative information channels, such as financial stress indices (Gkillas et al., 2020), investor fear gauges (Gong and Lin, 2018), and news-based predictors (Song et al., 2022).

A central empirical finding is that simple textual representations (Fast-Text) significantly outperform the directional HAR benchmark in out-of-sample forecasts. Traditional volatility models assume that information is gradually incorporated into prices through persistence in past volatility. However, our results suggest that news conveys forward-looking information about market stress, supply disruptions, and geopolitical risks that is not fully reflected in past volatility realizations. This aligns with recent evidence highlighting the growing role of alternative information sources in commodity markets (Li et al., 2021, 2022; Song et al., 2022)

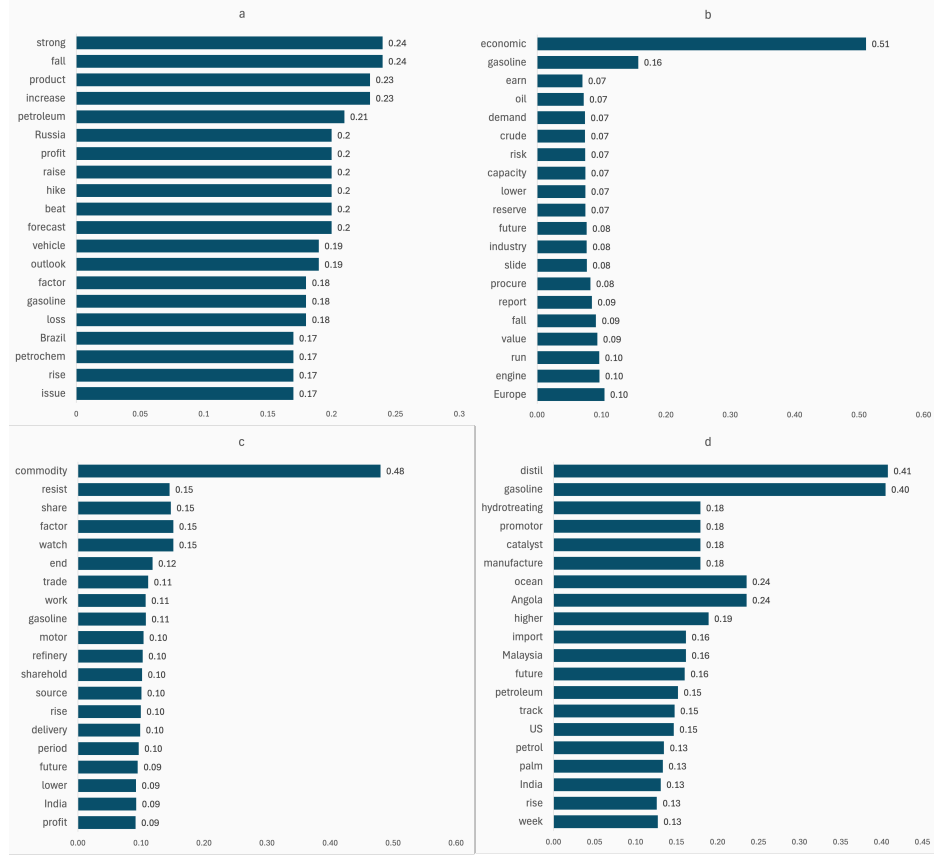


Figure 6: Top 20 Most Influential Words by SHAP Impact Across Four Time Periods. For expositional clarity and economic interpretability, and in line with recent oil-volatility studies that adopt event-based regime aggregation (Xiao et al., 2023), adjacent PELT-detected regimes are aggregated into four broad macroeconomic subperiods: (a) Before the Coronavirus Outbreak, (b) Epidemic Shock Period, (c) Epidemic Stabilization Period, (d) Russia-Ukraine conflict.

A striking result is the superior performance of FastText relative to transformer-based embeddings. While recent studies (e.g., Song et al., 2025) emphasize the potential of deep contextual language models for financial prediction, our findings indicate that subword-level information plays a more important role than model complexity when extracting predictive signals from short financial headlines. This is in line with the forecasting literature that suggests that well-regularized predictors often outperform high-capacity models in small-signal environments (Herrera et al., 2018), especially when the target variable is noisy and weakly predictable.

SHAP results (Figure 6) indicate that influential tokens correspond to well-documented drivers of oil-price volatility (e.g., Liu et al., 2019), including geopolitical events, and supply disruptions. Supporting the interpretation that our model learns economically coherent relationships rather than spurious textual correlations. Anchoring SHAP to regimes detected by PELT ensures that these interpretations reflect genuine structural shifts in volatility dynamics rather than arbitrary calendar partitions.

The developed framework allows the incorporation of textual information from (high-frequency) news-headlines through an efficient method (FastText), providing early indications of volatility shifts that have not been reflected in realized volatility. These valuable insights have implications for real-time risk monitoring, providing for example, timely adjustments in option hedging strategies. For instance, a risk desk observing a sustained FastText-based directional signal ahead of a confirmed HAR-based volatility shift could pre-emptively rebalance option positions or adjust margin requirements, rather than reacting only once realized volatility has already moved. PELT breakpoint detection and SHAP attributions enhance further the operational usefulness of our framework, providing a clear interpretation for drivers affecting these volatility changes.

Several limitations should be acknowledged. First, our sample is restricted to Reuters/Eikon headlines and Brent crude futures; the extent to which these results generalize to other news sources, benchmarks (e.g., WTI), or commodity markets more broadly remains an open question. Second, our forecasting task focuses on the direction of volatility change rather than its magnitude, which, while economically relevant for hedging decisions, does not capture the full risk profile relevant to position sizing. Third, the ensemble classifiers used in this study are comparatively simple relative to more complex architectures available in the literature; our results suggest that model complexity does not necessarily improve performance in this setting, but this finding is specific to the classifiers, embeddings, and headline-level textual data considered here.

Finally, there are several possible future research directions addressing current limitations of this study. First, the use of attention-based pooling to aggregate daily headlines may further improve signal extraction by emphasizing economically salient events. Second, extending the framework to hybrid models that jointly exploit textual features and realized-volatility dynamics could enhance robustness across regimes and improve performance at longer forecasting horizons, where persistence effects become more dominant.

Third, incorporating country-specific news sources may refine the geopolitical and supply-chain channels identified by SHAP and clarify cross-market spillovers between energy and financial markets.

Data Availability Statement

The data and code supporting the findings of this study are available at: https://anonymous.4open.science/r/Textual_Direction_Prediction_Oil_Volatility-1D72

References

- Andersen, T.G., Bollerslev, T., 1998. Answering the skeptics: Yes, standard volatility models do provide accurate forecasts. *International Economic Review* 39, 885–905.
- Anil, R., Borgeaud, S., Wu, Y., Alayrac, J.B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A.M., Hauth, A., Millican, K., et al., 2023. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.
- Araci, D., 2019. Finbert: Financial sentiment analysis with pre-trained language models. *arXiv preprint arXiv:1908.10063*.
- Baker, S.R., Bloom, N., Davis, S.J., 2016. Measuring economic policy uncertainty. *The Quarterly Journal of Economics* 131, 1593–1636.
- Bojanowski, P., Grave, E., Mikolov, T., Puhersch, C., Joulin, A., 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics* 5, 135–146.
- Corsi, F., 2009. A simple approximate long-memory model of realized volatility. *Journal of Financial Econometrics* 7, 174–196.
- Gkillas, K., Gupta, R., Pierdzioch, C., 2020. Forecasting realized oil-price volatility: The role of financial stress and asymmetric loss. *Journal of International Money and Finance* 104, 102137.
- Gong, X., Lin, B., 2018. The incremental information content of investor fear gauge for volatility forecasting in the crude oil futures market. *Energy Economics* 74, 370–386.

- He, H., Sun, M., Li, X., Mensah, I.A., 2022. A novel crude oil price trend prediction method: Machine learning classification algorithm based on multi-modal data features. *Energy* 244, 122706.
- Herrera, A., Hu, L., Pastor, D., 2018. Forecasting crude oil price volatility. *International Journal of Forecasting* 34, 622–635.
- Killick, R., Fearnhead, P., Eckley, I.A., 2012. Optimal detection of change-points with a linear computational cost. *Journal of the American Statistical Association* 107, 1590–1598.
- Kim, M.J., Kim, B.J., Kim, T., Jang, B.G., 2026. Forecasting realized volatility of the oil future prices via machine learning. *Applied Economics* 58, 756–779.
- Li, X., Liang, C., Chen, Z., Umar, M., 2022. Forecasting crude oil volatility with uncertainty indicators: New evidence. *Energy Economics* 108, 105895.
- Li, Y., Jiang, S., Li, X., Wang, S., 2021. The role of news sentiment in oil futures returns and volatility forecasting: Data-decomposition based deep learning approach. *Energy Economics* 95, 105140.
- Liu, J., Ma, F., Tang, Y., Zhang, Y., 2019. Geopolitical risk and oil volatility: A new insight. *Energy Economics* 84, 104548.
- Lundberg, S.M., Lee, S.I., 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems* 30.
- Ma, F., Lu, X., Zhu, B., 2025. Uncertainty and fluctuation in crude oil price: Evidence from machine learning models. *Annals of Operations Research* 345, 725–755.
- McNemar, Q., 1947. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika* 12, 153–157.
- Pennington, J., Socher, R., Manning, C., 2014. Glove: Global vectors for word representation, in: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP 2014)*, pp. 1532–1543.
- Song, Y., He, M., Wang, Y., Zhang, Y., 2022. Forecasting crude oil market volatility: A newspaper-based predictor regarding petroleum market volatility. *Resources Policy* 79, 103093.

- Song, Y., Zhang, Y., Huang, J., Yang, A., 2025. Volatility and value-at-risk forecasting using bert and transformer models incorporating investors' textual sentiments. *Finance Research Letters* , 108210.
- Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al., 2023. Llama: open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.
- Xiao, J., Wen, F., He, Z., 2023. Impact of geopolitical risks on investor attention and speculation in the oil market: Evidence from nonlinear and time-varying analysis. *Energy* 267, 126564.
- Zhang, X., Guo, Q., 2024. How useful are energy-related uncertainty for oil price volatility forecasting? *Finance Research Letters* 60, 104953.

Appendix A. News preprocessing and aggregation

Reuters frequently issues updates for major news events (e.g., UPDATE–Oil prices rise on supply concerns) and because these updates frequently convey overlapping information, we intentionally retain only the final daily version of each story, to avoid overweighting near-identical content. This preprocessing strategy mitigates distortions arising from heterogeneous article length and update frequency and ensures that the resulting news signal reflects information arrival rather than document volume. Finally, because language-model embeddings map each news item to a fixed-dimensional vector without an intrinsic notion of article importance, daily embedding representations are constructed by simple averaging across retained news items.

Appendix B. Robustness to Economic Policy Uncertainty (EPU)

This appendix reports an additional robustness analysis using the Global Economic Policy Uncertainty (EPU) index of Baker et al. (2016). The aim is to assess whether the predictive performance of news-based models reflects general macroeconomic uncertainty rather than oil-market-specific information contained in high-frequency news flow. All models are estimated using the same rolling-window design, sample period, and evaluation metrics as in the main text.

Benchmark setup

As a benchmark predictor, we use the EPU index as a macro-level proxy for policy-related uncertainty. We consider (i) an *EPU-only* classifier, and (ii) a *HAR+EPU* specification that augments the HAR benchmark with EPU. We compare both against the best-performing news-based specification (FastText).

Results

Table A.1 summarizes the out-of-sample predictive performance. An EPU-only classifier exhibits limited predictive power (accuracy = 0.523), indicating that monthly macro uncertainty alone is insufficient to explain high-frequency volatility movements. Augmenting a HAR-type specification with EPU improves performance (accuracy = 0.634), reflecting the role of volatility persistence and macro uncertainty. However, the news-based model substantially outperforms both benchmarks (accuracy = 0.705). A McNemar

Table A.1: Robustness benchmark—EPU versus news-based volatility direction forecasts.

Model	Accuracy	Precision	Recall	F1-score
EPU only	0.523	0.512	0.523	0.365
HAR + EPU	0.634	0.644	0.634	0.632
News only	0.705	0.708	0.705	0.703

Notes: This table reports rolling-window classification performance for daily volatility direction forecasts. “EPU only” uses the Global Economic Policy Uncertainty index of Baker et al. (2016) as the sole predictor. “HAR + EPU” augments a HAR model with EPU. “News only” uses lagged FastText news embeddings. All models are estimated using the same rolling-window design as in the main analysis. Performance metrics are computed over the common evaluation sample spanning 2014–2023.

test confirms that the predictive accuracy of the news-based model is significantly higher than that of the EPU benchmark ($\chi^2(1) = 49.284$, $p < 0.001$). Overall, these results indicate that the proposed textual signal captures incremental information beyond macro uncertainty proxies and is not merely a proxy for EPU.

An important implication concerns the nature of the information embedded in news text. Unlike macroeconomic uncertainty indices, such as EPU index commonly used in volatility models (Zhang and Guo, 2024), news headlines reflect high-frequency, event-driven market microstructure. They capture granular developments such as OPEC announcements and geopolitical escalations that may not be immediately reflected in low-frequency macro uncertainty measures. This distinction is supported by the robustness analysis showing that EPU-based models (EPU only and EPU-augmented HAR) perform substantially worse than news-based classifiers. Together, these results indicate that the predictive content of textual news is not merely a proxy for aggregate uncertainty but reflects distinct, high-resolution information about oil-market conditions.