Do Energy Sentiment Predict Oil Price Shocks? A Machine Learning Analysis

Mohammad Abdullah[a], Lei Pan[b,c], Sahib Hossain[d],

Aviral Kumar Tiwari[e *], Freeman Brobbey Owusu[f]

[a] *University of Southampton Malaysia, Malaysia*

[b] *School of Accounting, Economics and Finance, Curtin University, Australia*

[c] *Centre for Development Economics and Sustainability, Monash University, Australia*

[d] *University of Scholars, Bangladesh*

[e] *Indian Institute of Management Bodh Gaya, India*

[f] *Loughborough University, United Kingdom*

Abstract

This study explores the use of energy sentiment as a predictive tool for forecasting oil price shocks by developing a Twitter-based energy sentiment index derived from 1,911,631 tweets. The sentiment index is utilized to predict three types of oil price shocks: demand, supply, and risk shocks, through various machine learning algorithms. Among these, the XGBoost model is found to outperform other models, achieving prediction accuracies of 60.20%, 62.00%, and 92.60% for demand shock, supply shock, and risk shock, respectively. Further model interpretation using Explainable AI reveals that the developed energy sentiment indicators contribute 29.33% to the oil price shock prediction, demonstrating the significant role of sentiment data in forecasting oil price fluctuations. These findings highlight the potential of leveraging real-time social media sentiment for improving oil price prediction models.

Keywords: Energy sentiment, oil price shock, machine learning, XGBoost, Explainable AI

JEL Classifications: Q40, Q41

* Corresponding author.

E-mail address: M.Abdullah@soton.ac.uk (M. Abdullah), lei.pan@curtin.edu.au (L. Pan), sahibhossain@ius.edu.bd (S. Hossain), aviral.eco@gmail.com (A.K.. Tiwari), f.b.owusu@lboro.ac.uk (F.B. Owusu),

1. Introduction

The dynamics of oil markets play a pivotal role in shaping global economic landscapes, influencing inflation rates, stock market performance, and overall macroeconomic stability. Oil price shocks, in particular, featured by abrupt and significant fluctuations in oil prices can trigger cascading effects throughout the global economy. For instance, the ongoing conflict between Russia and Ukraine has had profound influences on global energy markets, contributing to heightened volatility and uncertainty in oil prices, results in the rise of production cost, leading to further inflationary pressures, and affect consumer confidence, and investment strategies. Understanding and predicting oil price shocks—especially during periods of market volatility—are therefore of paramount importance for policymakers, investors, and industry stakeholders alike. Traditional econometric models and time-series analyses have been employed to forecast oil prices and oil price shocks, with mixed success. While models based on macroeconomic fundamentals or supply-demand balances provide useful insights into long-term price trends, they often struggle to account for the high degree of volatility and sudden, unpredictable shifts that characterize oil markets. Moreover, these models are limited in their ability to process large volumes of unstructured data, such as social media posts, which can reflect real-time shifts in market sentiment and expectations. In recent years, there has been a growing interest in leveraging alternative data sources and using sophisticated computational methods to enhance the accuracy and timeliness of oil price forecasts.

One such emerging approach is the utilization of sentiment analysis, particularly sentiment derived from news articles, social media discussions and financial reports, to predict oil price shocks. The rationale behind employing sentiment analysis for oil price prediction lies in the recognition that market sentiments and investor psychology play a crucial role in driving short-term price fluctuations, especially in volatile commodity markets. Energy sentiment, in particular, encapsulating the collective perceptions, beliefs, and attitudes towards the energy sector, serves as a powerful barometer of market sentiment, reflecting the prevailing sentiment among investors and analysts, and ultimately, the trajectory of oil prices. By harnessing the predictive power of energy sentiment, stakeholders can gain valuable insights into the factors driving oil price movements and anticipate impending shocks before they materialize.

Previous studies have demonstrated the relevance of online-based text in predicting oil price shocks. For instance, by adopting the structural vector autoregression (SVAR) model, and the principle component analysis (PCA) method to combine the google search index, Yao, Zhang and Ma (2017) analysis the impact of investor attention on international crude oil prices. Their research highlight a significance negative influence of investor attention on crude oil prices. Using news on the oil market, Li *et al.* (2021) document that shocks in news sentiment can lead to volatility across the future prices of oil. Zhe *et al.* (2022) use the comments on an online financial forum (i.e., Eastmoney forum), and show the predictive power of sentiment for China's crude oil price. A notable fact, however, is that these studies rely solely on counting words for textual analysis and simple extraction of quantity in news which fail to provide a sentiment analysis based on the meaning of text. Hence, considering the emotional tendency of web texts often cause fluctuations of investor sentiment, which further lead to energy price movements, and also recommended by Shiller (2017) that semantic sentiment shall be used for examining narratives across financial markets since it can highlight the psychological significance of sentiment. There is a stream of literature focuses on using emotion analysis from texts and mining deeper information to aid prediction. For instance, Tetlock (2007) investigates the predictive power of daily content from Wall Street Journal column for stock market prices using vector autoregressive (VAR) framework. Although the study primarily focuses on stock market prices, it indirectly underscores the interconnectedness between emotional sentiment analysis and broader financial markets. The research reveals that media pessimism could potentially predict stock market prices. Li *et al.* (2022) conducts a comprehensive review of the literature on the relationship between oil price and investor sentiment. Their review provides valuable insights into various methodologies used in investigating the connection between oil price and sentiment analysis. The authors discuss the challenges and limitations associated with different approaches and highlight the potential of sentiment analysis as a tool in predicting oil price shocks.

Although existing studies have shown promising results in using sentiment analysis to forecast oil price shocks, there are lack of studies using sentiment data from other sources, such as energy sector, to enhance the accuracy and robustness of predictive models. We, in this paper, aim to fill this gap in the literature. Our study strives to answer several key questions aimed at exploring the potential of energy sentiment as a predictor of oil price shocks. First, we investigate whether energy sentiment can reliably forecast oil price shocks, hypothesizing that shifts in public and market sentiment, driven by news and geopolitical events, can serve as leading indicators of such shocks. Second, we aim to determine whether energy sentiment can distinguish between different types of oil price shocks—specifically demand, supply, and risk shocks—each of which has distinct drivers and implications for the oil market. Understanding the unique sentiment patterns associated with these shock types is crucial for enhancing the precision of oil price forecasts. Finally, we assess which machine learning techniques are most effective in utilizing sentiment data to predict oil price shocks. By applying these algorithms to the constructed energy sentiment index, we seek to identify the most accurate and robust predictive models, thus advancing the integration of sentiment analysis into oil price forecasting methodologies.

This paper makes three makes three important contributions to the literature on oil price forecasting and sentiment analysis: First, to the best of our knowledge, our paper is the first to explore the potential feasibility and effectiveness of using energy sentiment as a predictive tool for anticipating oil price shocks. We adopt various machine learning algorithms such as XGBoost, Random Forest, and SVM to evaluate the effectiveness of sentiment-based predictions. These methods are more flexible and accurate than traditional methods like linear regression or ARIMA.

Furthermore, they are better equipped to handle complex, non-linear relationships and high-dimensional data, making them particularly suitable for capturing the intricacies of sentiment-driven oil price shocks. By comparing the performance of these algorithms, we aim to determine which approaches are most suitable for forecasting oil price shocks. Moreover, by using machine learning algorithms, we provide a more comprehensive and predictive framework that improves upon traditional econometric models and offers a more nuanced understanding of the factors driving oil price fluctuations. This methodological innovation helps bridge the gap between traditional economic theories and modern data-driven forecasting approaches. Second, we develop a unique energy sentiment index by specifically selecting Twitter feeds related to energy markets, a methodological approach that distinguishes our study from existing literature. This approach adds value to prior work, as argued by Abdollahi (2023), who emphasizes the importance of incorporating real-time social media data in understanding market sentiment. By aggregating Twitter sentiment data, we capture public opinion and emotional responses in a timely manner, offering a fresh perspective on factors influencing oil price movements. Our third contribution is that while previous research has focused on general oil price forecasting, this study aims to distinguish between different types of oil price shocks (demand, supply, and risk). By identifying the specific drivers of each shock, we can develop more precise and actionable predictions.

Foreshadowing the main results, we find that ensemble models, particularly XGBoost, outperform traditional methods such as Logistic Regression (LR), Support Vector Machines (SVM), and K-Nearest Neighbors (KNN). XGBoost consistently delivered the highest accuracy, precision, recall, F1-score, and ROC_AUC across multiple evaluation metrics, highlighting its ability to capture complex non-linear relationships and interactions within the data. Additionally, Random Forest (RF) and LightGBM (LGBM) showed competitive performance, reinforcing the strength of tree-based models in oil price forecasting. Through robustness tests with varying train-test splits, we confirmed the stability and reliability of the XGBoost model, which maintained its superior performance across different training data sizes. Furthermore, the Explainable AI analysis using SHAP values provided valuable insights into the factors driving oil price predictions. Sentiment indicators, particularly negative and positive sentiment, emerged as the most influential features, followed by global risk aversion (RAI) and oil-specific volatility (OVX). These findings emphasize the importance of market sentiment, risk perceptions, and volatility in predicting oil price shocks, alongside traditional supply and demand factors. Our paper has important policy implications for incorporating sentiment analysis into risk management strategies and investment decision-making processes within the energy sector. By harnessing the insights derived from energy sentiment data, stakeholders can better navigate the complexities of the oil market and proactively respond to emerging trends and sentiment shifts.

The remainder of this paper is organized as follows: Section 2 reviews relevant literature. Section 3 explains the data, section 4 outlines the empirical strategies adopted in this study. Section 5 reports and discusses the empirical findings and section 6 concludes with policy recommendations provided.

2. Review of Related Literature

Due to the pivotal role of crude oil plays in the global economy, accurately forecasting crude oil prices has attracted significant attentions among researchers, leading to the implementation of various models. Early research rely on theory-based models for establishing linkages between oil price and other variables to predict prices. Three basic methods are used for the theory-based models: i) models employ futures, ii) models employ spot prices, and iii) models employ economic indicators. The seminal study of Verleger (1982) develops a theory-driven model by utilizing a barrel of oil price can be predicted by the weighted sum of the prices of the products made out of oil. Although the model has its merits such as simplicity, Knetsch (2007) argues that oil futures are not a good predictor for oil prices. Hence, a number of models have been established to predict energy futures price instead of directly forecasting oil price (e.g., Date *et al.*, 2013; Lautier and Galli, 2004). However, since our focus is the crude oil price, we do not consider these futures-based models in our analysis. Baumeister et al. (2018) argues that the futures-focused models can be improved by using spot prices instead of future prices. Alquist *et al.* (2013) propose that some economic indicators can be used for oil price forecasting. They therefore develop an economic indicators based model which assumes that oil price changes simultaneously with some economic indicators such as industrial raw materials.

Early empirical studies tend to employ traditional time series techniques such as autoregressive (AR), autoregressive moving average (ARMA), and vector autoregression (VAR) models for predicting oil prices (see e.g., Baumeister and Kilian, 2012; Park and Ratti, 2008). The AR and ARMA models are shown to be performed well for oil price forecasting. Their weakness however is that they only consider one-variable (the previous values of oil price) prediction scenarios. VAR models are regarded as more successful approaches for oil price prediction because they allow multiple variables and model the interlinkages between the variables. Using oil future spreads, Alquist and Kilian (2010) further demonstrate the reliability of VAR models for forecasting oil prices. Although VAR models improve forecast accuracy, they are not able to explain why prices change.

Due to the limitations of theory-based models and classical time series techniques, machine learning (ML) is considered as alternative approaches. Many studies have applied ML techniques in the field of energy economics (see e.g., Ghoddusi *et al.*, 2019; Lin *et al.*, 2020; Luo *et al.*, 2018). There are two advantages of ML models: i) they include information that are likely to be related to oil prices, even if such relationship is difficult to quantify through theoretical or regression equations; ii) they can investigate complex and highly nonlinear equations, hence improve predictions. A number of studies have adopted simple ML models for oil price forecasting (Ramyar and Kianfar, 2019; Zhao *et al.*, 2017). Other studies adopt more complex ML models to provide more complicated features that are useful for predicting oil prices (Hu *et al.*, 2012; Xu and Niu, 2022). The advantage of simple ML models is that they are easier to be trained and configured. In contrast, if the ML models are more complex, then more difficult to have a good estimation for all model parameters and to avoid the issue of overfitting. Hence, in this paper, we adopt ML models for oil price shock prediction.

A stream of literature on oil price prediction employs wavelet analysis (Jammazi and Aloui, 2012; Lin *et al.*, 2020). This method is appealing for dividing the information in price series into smaller pieces that makes it easier to sort out linkages within the data. The pitfall of wavelet analysis is that the future trends of oil price are deducted based on historical data. To deal with this issue, researchers realize that big data such as textual data can be a novel data source for predicting oil prices. Therefore, a growing body of literature have explored online text mining for market predictions (Gong *et al.*, 2022; Pagolu *et al.*, 2016). Among them, the widely accepted approach for processing textual data is sentiment analysis.

Sentiment analysis can be broadly divided into two types: i) dictionary-based approach, and ii) ML-based approach. Dictionary-based approach usually counts the number of positive and negative words in the textual data and then utilize these counts for computing a sentiment score (Medhat *et al.*, 2014). For instance, Das and Chen (2007) propose a method for extracting small investor sentiment from a stock message board. Nandwani and Verma (2021) document that dictionary-based approach performs well in sentiment analysis. The weakness of dictionary-based approach however is that words are only counted as fully positive or negative, while in reality some words are stronger sentiment than others. Moreover, further challenge is that one word may have different meanings in different scenarios. Such challenge, according to Liu (2012), is difficult to overcome with just using dictionary-based approach. Therefore, a body of literature adopts ML-based approach (see e.g., Lakatos *et al.*, 2022; Zhao *et al.*, 2017) as an alternative to dictionary-based analysis have emerged to overcome some of their issues. Sudhir and Suresh (2021) show that ML-based approaches are often superior because their outstanding accuracy and exceptional results. Furthermore, ML-based approach can recognize more intricate textual patterns and excel when applied to big datasets. Hence, in this paper, we use machine learning techniques as our main empirical strategies.

Overall, so far there are no studies have considered using energy sentiment for crude oil forecasting. This study aims to fill this void. Among the most promising new data sources are social media platforms, particularly Twitter, where real-time discussions surrounding energy markets offer rich, untapped insights into market sentiment. Twitter allows users to disseminate and access information in real time, making it an ideal platform for monitoring market-relevant events and discussions. As a result, financial analysts and researchers have increasingly turned to Twitter data to extract sentiment signals that can be used for forecasting asset prices, including stocks, bonds, and commodities. Therefore, we develop a unique energy sentiment index using Twitter feeds. By doing so, our index improves crude oil prediction performances significantly.

3. Data and Variables

3.1 Energy sentiment index

To construct the energy sentiment index, we selected energy-related keywords following studies by Bouteska *et al.* (2024), Corbett and Savarimuthu (2022), and Polyzos and Wang (2022). The chosen keywords included: "green energy," "renewable energy," "solar energy," "wind energy," "hydropower," "thermal power," "energy price," "energy policy," "energy poverty," "energy resource," "household energy," "industrial energy," "energy cost," "energy commodity price," "fuel cost," "oil price," "oil supply," "oil production," "oil demand," "fossil fuel," "gasoline," "gas," "natural gas," "electricity price," and "electricity cost." Using the Twitter Academic API (renamed X), we gathered 2,654,274 tweets containing these keywords.

The collected tweets were then subjected to a rigorous cleaning process using Python. This process involved removing emoticons, digits, retweets, white spaces, URLs, punctuation, correcting spelling errors, and eliminating special characters. Additionally, we used the Natural Language Toolkit (NLTK) module's corpus stop words (Bird *et al.*, 2009) to remove common stop words. Tweets with less than three words were also deleted from the dataset due to their lack of meaningful content (Abdullah *et al.*, 2024). After the data cleaning process, we were left with 1,911,631 tweets relevant to the energy market. Figure 1 illustrates the word cloud of the cleaned dataset, highlighting the prominence of terms like "gas" and "oil."

[Insert Figure 1 Here]

Next, we performed sentiment analysis to measure the sentiment of each tweet. This study employed lexical analysis for sentiment scoring. By tokenizing each tweet, we utilized the Python NLTK and TextBlob libraries to evaluate sentiment based on the polarity and subjectivity of the tweets (Hutto and Gilbert, 2014). The polarity score, ranging from -1.0 to 1.0, indicates the sentiment's positivity or negativity. Subjectivity scores, ranging from 0.0 to 1.0, reflect the degree of objectivity, with 0.0 being extremely objective and 1.0 being highly subjective. We computed a compound score by summing all items in the lexicon and normalizing it between -1 and 1. Tweets were then categorized as Positive, Neutral, or Negative based on their compound scores: Positive (compound score ≥ 0.05), Neutral ($-0.05 <$ compound score < 0.05), and Negative (compound score ≤ -0.05) (Hutto and Gilbert, 2014). Figure 2 illustrates the sentiment distribution.

[Insert Figure 2 Here]

After classifying each tweet into Positive, Neutral, or Negative sentiment, we developed the energy sentiment index based on the methodology of Abdullah *et al.* (2024). This index was calculated using the following equation

$$ESent_t = \frac{\sum_t positive - \sum_t negative}{\sum_t positive + \sum_t neutral + \sum_t negative} \tag{1}$$

Here, *ESent* denotes the energy sentiment index for day t, positive, negative, and neutral denotes each tweet's sentiment.

Figure 3 illustrates the constructed sentiment indicators, which include *ESent*, Polarity, Subjectivity, and Compound scores. The *ESent* indicator captures the overall energy sentiment by aggregating the classified sentiments of individual tweets. The Polarity score measures the positivity or negativity of the sentiment on a scale from -1.0 to 1.0, while the Subjectivity score ranges from 0.0 to 1.0, indicating the degree of objectivity versus subjectivity in the tweets. The Compound score is a normalized aggregate sentiment score ranging from -1 to 1, summarizing the overall sentiment expressed in the tweets. These sentiment indicators exhibit similar time-varying patterns, reflecting the dynamic nature of public opinion and its influence on the energy market.

[Insert Figure 3 Here]

3.2 Measurement of oil price shock

To measure oil price shocks, we employ the framework proposed by Ready (2018), which identifies three types of structural shocks: supply shocks (Supply_Shock), demand shocks (Demand_Shock), and risk shocks (Risk_Shock). This methodology utilizes three critical variables: an index of oil-producing firms, measures of oil price fluctuations, and a proxy for expected return changes. Specifically, the World Integrated Oil and Gas Producer Index was selected to provide a comprehensive representation of the oil industry. To analyze oil price changes, we used the one-month returns on the second nearest maturity of the NYMEX Crude-Light Sweet Oil contract, as these short-term futures contracts effectively reflect oil price movements. Additionally, the volatility index (VIX) from the Chicago Board Options Exchange (CBOE) was used as a proxy for investor risk attitudes, with daily VIX data modeling unexpected volatility through an ARMA(1,1) framework. In this study, oil demand shocks (Demand_Shock) reflect the influence of changes in global real economic activity on oil prices. Oil risk shocks (Risk_Shock) capture the impact of changes in stock market volatility and market expectations. Oil supply shocks (Supply_Shock) account for production-related disturbances in the crude oil market, measuring factors beyond Demand_Shock and Risk_Shock that influence oil prices. Figure 4 illustrates the oil price shock variables, showing similar patterns of shocks during different periods of economic turmoil.

[Insert Figure 4 Here]

Finally, for shock prediction we constructed binary variables based on the identified shocks: Demand_Shock is set to 0 for a negative shock and 1 for a positive shock; Supply_Shock is set to 0 for a negative supply shock and 1 for a positive supply shock; and Risk_Shock is set to 0 for a negative risk shock and 1 for a positive risk shock.

3.3 Forecasting dataset

In addition to sentiment indicators, we also consider other related predictors of oil price shocks, following earlier studies (Yang *et al.* 2024; Kumar and Mallick, 2024; Yang *et al.*, 2023; Al-Fayoumi *et al.*, 2023; Sehgal and Pandey, 2025; Tiwari *et al.*, 2023). The additional predictors include the OVX (Oil Volatility Index), which measures the market's expectations of volatility in crude oil prices; the USEPU (USA Economic Policy Uncertainty Index) and UKEPU (UK Economic Policy Uncertainty Index), which reflect the uncertainty regarding economic policies in the USA and UK, respectively; the RAI (Global Risk Aversion Index), which gauges global risk aversion; the GPRI (Geopolitical Risk Index), which assesses geopolitical risks; the BDI (Baltic Dry Index), which is an economic indicator issued daily by the London-based Baltic Exchange that measures the cost of shipping goods worldwide; and the USSSR (US Monetary Policy), which captures the stance of US monetary policy. Data for these variables were collected for the period from 01 August 2008 to 30 June 2022 to align with the availability of the energy sentiment index data. Table A1 provides a detailed description of these additional variables, along with their respective data sources.

Table 1 presents the summary statistics for the variables used in this study. For instance, the mean of the energy sentiment index (*ESent*) is 0.546, with a low variance of 0.005, suggesting relatively stable sentiment levels over time. Skewness values reveal the asymmetry of the distributions. A negative skewness for *ESent* (-0.734) indicates that the distribution is tilted towards more negative sentiment, while positive skewness is observed in some of the sentiment sub-categories, such as ESent_Neg and ESent_Pos. Kurtosis values describe the "tailedness" of the data distribution. High positive kurtosis values, like those observed for ESent_Neg (16.656) and ESent_Pos (20.541), suggest heavy tails and potential outliers in these series. The Jarque-Bera (JB) test statistic for normality strongly rejects the null hypothesis of normality across all variables, with extremely high values indicating that most variables deviate significantly from a normal distribution. The results from the Elliott, Rothenberg, and Stock (ERS) unit root test show that all variables exhibit negative test statistics (Elliott *et al.*, 1992), suggesting the rejection of the null hypothesis of a unit root and indicating that the time series data are stationary. Figure 5 illustrates the correlation matrix.

[Insert Table 1 & Figure 5 Here]

4. Methodology

Using a binary classification technique, our work focuses on predicting oil price shock. Thus, we have selected nine popular models for our research from the comprehensive literature survey. In particular, we have chosen eight machine learning models, and logistic regression as the conventional benchmark. Table 2 lists the chosen models that have been hyperparameter tuned. We use the Python scikit-learn library to create the benchmark and machine learning models. Description of each models are described as follows:

[Insert Table 2 Here]

4.1 Logistic Regression (LR)

Logistic Regression (LR) is a statistical method used for binary classification, modeling the relationship between a dependent binary variable (0 or 1) and one or more independent variables (Kleinbaum *et al.*, 2002). It estimates the probability of a binary response based on one or more predictor variables.

The logistic regression model predicts the probability of the binary outcome $p(y=1|X)$ using the following equation:

$$p(y=1|X) = \frac{1}{1+e^{-(\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_n X_n)}} \tag{2}$$

here, $p(y=1|X)$ is the probability that the target variable y is 1, given the features X. β_0 is the intercept term. $\beta_1, \dots, \beta_n$ are the coefficients for the predictor variables $X_1, \dots, X_n$. e is the base of the natural logarithm.

4.2 Random Forest (RF)

Random Forest (RF) is an ensemble learning method that constructs a multitude of decision trees during training (Breiman, 2001). The output is the majority vote of all trees (for classification) or average prediction (for regression). It is known for its robustness and high accuracy in many applications, including binary classification. The final prediction $\hat{y}$ is made by aggregating the predictions from multiple decision trees $T_1, T_2, \dots, T_M$:

$$\hat{y} = \text{majority vote}\big(T_1(X), T_2(X), \dots, T_M(X)\big) \tag{2}$$

Each decision tree $T_m(X)$ produces a binary output (0 or 1) based on the features X.

4.3 Support Vector Machine (SVM)

Support Vector Machine (SVM) is a supervised learning algorithm that is used for classification (Hearst *et al.*, 1998). It works by finding a hyperplane that best separates data points of different classes in a high-dimensional space. SVM seeks the optimal hyperplane that maximizes the margin between the two classes. The SVM optimization problem can be formulated as:

$$\min \frac{1}{2} \parallel \mathbf{w} \parallel^2 \tag{4}$$

subject to:

$$y_i(\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1, \ \forall i = 1, \dots, N \tag{5}$$

where $\mathbf{x}_i$ represents the feature vector for sample i. y_i is the true class label (either +1 or -1) for sample i. $\mathbf{w}$ is the weight vector. b is the bias term. $\parallel \mathbf{w} \parallel$ is the norm of the weight vector, which determines the margin.

4.4 Naïve Bayes (NB)

Naïve Bayes is a probabilistic classifier based on Bayes' Theorem, assuming conditional independence between features (Murphy, 2006). It calculates the probability of a class label given the features and selects the class with the highest probability. The probability of class y given features $X = (X_1, X_2, \dots, X_n)$ is calculated using Bayes' Theorem:

$$P(y|X) = \frac{P(X|y)P(y)}{P(X)} \tag{6}$$

where $P(y|X)$ is the posterior probability of class y given the features X. $P(X|y)$ is the likelihood of features X given the class y. $P(y)$ is the prior probability of class y. $P(X)$ is the evidence, the total probability of X across all classes.

4.5 Extra Trees

Extra Trees (Extremely Randomized Trees) is an ensemble method similar to Random Forest, but with more randomness (Geurts *et al.*, 2006). It builds a collection of decision trees, where each tree is trained by selecting random splits at each node, making the model less prone to overfitting. The final prediction is a majority vote (for classification) from all trees:

$$\hat{y} = \text{majority vote}\big(T_1(X), T_2(X), \dots, T_M(X)\big) \tag{7}$$

where each $T_m(X)$ is a decision tree trained on random subsets of features and samples.

4.6 AdaBoost

AdaBoost (Adaptive Boosting) is an ensemble method that combines multiple weak learners (often decision trees) to form a strong classifier (Schapire, 2013). It adjusts the weight of each weak learner based on its performance, focusing more on samples that are incorrectly classified. The AdaBoost algorithm aggregates weak learners $T_m(X)$ as follows:

$$\hat{y} = \text{sign}\left(\sum_{m=1}^{M} \alpha_m T_m(X)\right) \tag{8}$$

where α_m is the weight assigned to the weak learner $T_m(X)$, based on its accuracy. sign$(\cdot)$ is the function that converts the output to binary values (+1 or -1).

4.7 XGBoost (XGB)

XGBoost (Extreme Gradient Boosting) is an efficient and scalable implementation of gradient boosting (Chen and Guestrin, 2016). It uses decision trees as base learners and builds an ensemble by iteratively fitting trees to the residuals (errors) of previous trees. XGBoost optimizes the model using gradient descent. The prediction is made by summing the contributions of each tree $T_m(X)$:

$$\hat{y} = \sum_{m=1}^{M} T_m\,(X) \tag{9}$$

where each $T_m(X)$ is a decision tree, and the model minimizes the loss function using gradient descent.

4.8 K-Nearest Neighbors (KNN)

K-Nearest Neighbors (KNN) is a non-parametric, instance-based learning algorithm (Zhang, 2016). It classifies a sample based on the majority class of its K nearest neighbors in the feature space. The distance metric, such as Euclidean distance, is used to determine the "closeness" between data points. The prediction for a new sample X_{extnew} is:

$$\hat{y}(X_{extnew}) = \text{majority vote}(y_1, y_2, \dots, y_K) \tag{20}$$

where $y_1, y_2, \dots, y_K$ are the class labels of the K nearest neighbors, and the distance function (e.g., Euclidean) is used to identify these neighbors.

4.9 LightGBM (LGBM)

LightGBM is a gradient boosting framework that uses decision trees as base learners. It is optimized for speed and efficiency and is particularly suited for large datasets (Ke *et al.*, 2017). It uses histogram-based algorithms to speed up the training process. Like XGBoost, the prediction is the sum of contributions from each tree:

$$\hat{y} = \sum_{m=1}^{M} T_m(X) \tag{31}$$

where $T_m(X)$ is the decision tree, and the algorithm uses gradient-based optimization to minimize the loss function.

4.10 Hyperparameters

Table 2 presents the hyperparameters used for tuning the machine learning models in this study. Hyperparameter optimization is a critical step in improving model performance, as it determines the configuration of algorithms for better generalization and predictive accuracy. The selected hyperparameters for each model include parameters like learning rate, maximum depth, number of estimators, and regularization terms. These hyperparameters were carefully chosen and tuned using cross-validation to ensure optimal performance across various machine learning models.

4.11 Shapley Adaptive Explanation

Shapley values are a concept from cooperative game theory used to fairly allocate the contribution of each feature towards the prediction (Lundberg, 2017). In the context of binary classification, Shapley values help explain how each feature influences the model's output. The Shapley value for a feature is the average marginal contribution of that feature across all possible subsets of features. In the context of predicting oil price shocks, Shapley values provide an explanation of how each feature (such as past oil prices, geopolitical events, supply-demand factors, etc.) influences the model's prediction (either 0 or 1, indicating a shock or no shock).

For a binary classification task, the Shapley value for feature i can be computed as:

$$\phi_i(f) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!\,(|N| - |S| - 1)!}{|N|!} [f(S \cup \{i\}) - f(S)] \tag{42}$$

where $\phi_i(f)$ is the Shapley value of feature i for the model f, N is the set of all features, $S \subseteq N \setminus \{i\}$ represents a subset of features that does not include feature i, $f(S)$ is the model's prediction using the subset of features S, the summation is over all possible subsets S of features excluding feature i, $\frac{|S|!(|N|-|S|-1)!}{|N|!}$ is the weight assigned to each subset, which reflects the importance of the order in which features are added to the model.

Shapley values calculate the average contribution of each feature to the prediction by considering all possible ways in which the feature could be added to subsets of other features. This provides a fair and unbiased method to explain the importance of each feature in binary classification models used for predicting oil price shocks.

4.12 Model performance evaluation

For our oil price shock prediction, we have used several performance metrices to evaluate models. Accuracy measures the overall correctness of the model by calculating the proportion of correct predictions. Precision indicates the percentage of true positive predictions among all predicted positives, helping to assess how well the model avoids false positives. Recall (or sensitivity) focuses on the percentage of true positives detected by the model, indicating its ability to capture all relevant instances. Log-Loss evaluates the probability output of the model, penalizing false classifications more heavily as the confidence in those predictions increases. F1-score provides a harmonic mean of precision and recall, balancing the trade-off between the two. Jaccard index measures the intersection of predicted and actual positives relative to their union, giving an indication of overlap. Finally, ROC_AUC (Receiver Operating Characteristic - Area Under Curve) summarizes the model's ability to distinguish between positive and negative classes across all thresholds, with a higher score reflecting better model performance.

5. Empirical Results

5.1 Forecasting results

For our empirical analysis we use 70% training and 30% in testing sample split. All nine models are trained and tested. Figure 6 shows the confusion matrices for each model provide valuable insights into their classification performance in predicting oil price shocks. The confusion matrix for LR shows that it correctly predicted 33.05% of negative shocks and 22.48% of positive shocks. The model's performance in identifying negative shocks is relatively better, but it still struggles with positive shocks, with 24.98% false negatives. This indicates that LR has a moderate ability to classify oil price shocks but could benefit from improvements in handling positive shocks. RF performed slightly better than LR, with 29.59% true negative and 26.61% true positive predictions. However, it still experienced considerable misclassification, with 20.85% of positive shocks misclassified as negative, and 22.96% of negative shocks misclassified as positive. These results suggest RF is more balanced in its predictions but still has room for improvement in reducing misclassifications.

[Insert Figure 6 Here]

The SVM model produced mixed results, with 31.12% of negative shocks and 23.05% of positive shocks correctly classified. It had a relatively high false positive rate of 24.40% and a moderate false negative rate of 23.05%. While SVM is competent at distinguishing negative shocks, its performance is less consistent for positive shocks. Naïve Bayes showed relatively weaker performance compared to the other models, with 14.31% true negatives and 36.31% true positives for negative and positive shocks, respectively. It had high false positive and false negative rates, suggesting that Naïve Bayes struggles with correctly classifying both negative and positive shocks, particularly with positive shocks. ExtraTrees demonstrated a moderate performance, with 31.32% true negatives and 23.54%

true positives. However, its false positive and false negative rates were somewhat high, particularly with positive shocks. The results suggest that ExtraTrees can provide reasonable classifications, but there is significant room for improvement, especially in handling misclassifications of positive shocks.

AdaBoost performed similarly to ExtraTrees, with 27.67% true negatives and 26.80% true positives. Although its performance was fairly balanced, the model still exhibited considerable misclassifications, particularly false positives (20.65%). This suggests that while AdaBoost is decent, it could be more precise in classifying both types of shocks. XGBoost clearly outperforms all other models in terms of accuracy, with 30.84% true negatives and 24.88% true positives. The misclassification rates are relatively lower, with false positives at 22.57% and false negatives at 22.57%, which is better than most other models. This indicates that XGBoost is not only good at identifying both negative and positive shocks but also more consistent in its predictions across both categories. KNN showed a moderate performance, with 27.09% true negatives and 24.50% true positives. However, its false positive and false negative rates were still significant, at 22.96% and 22.36%, respectively. These results indicate that while KNN is a decent model, it does not perform as well as XGBoost in terms of precision and recall. LightGBM's performance was relatively balanced, with 34.68% true negatives and 23.05% true positives. However, it had a relatively higher false negative rate of 24.40% compared to other models. The performance of LGBM was decent but not as strong as XGBoost.

The superior performance of XGBoost is further confirmed by the ROC AUC plot shown in Figure 7. The ROC curve illustrates the model's ability to distinguish between positive and negative oil price shocks, with the AUC (Area Under the Curve) providing a measure of overall model performance. XGBoost exhibits the highest AUC value, indicating its strong discriminatory power and superior classification ability compared to the other models.

[Insert Figure 7 Here]

The performance metrics of the models presented in Table 3. Accuracy is highest for XGBoost (60.2%), followed by RF (59.2%) and LGBM (57.8%). This indicates that XGBoost correctly classified the most instances of oil price shocks overall. When considering precision, XGBoost again leads with a score of 0.622, followed by RF (0.620) and LGBM (0.611), which means that XGBoost had the highest proportion of true positive predictions among all models. In terms of recall, XGBoost (0.556) outperforms most models, particularly with better sensitivity in detecting positive shocks compared to RF (0.509) and Naïve Bayes (0.733), though the latter has an unusually high recall but low precision. Log-Loss, a measure of model uncertainty, is lowest for XGBoost (14.334), signifying the model's more confident and accurate predictions. F1-score, which balances precision and recall, is also highest for XGBoost (0.587), indicating a well-rounded performance. The Jaccard index for XGBoost is 0.415, demonstrating good similarity between predicted and actual positive shock instances. Lastly, the ROC AUC score is highest for XGBoost (0.603), confirming its superior ability to distinguish between positive and negative oil price shocks. While models like RF and Naïve Bayes show competitive performance, XGBoost consistently outperforms across nearly all metrics, highlighting its robustness and reliability for forecasting oil price shocks. Earlier studies by Gumus and Kiran (2017), Tissaoui *et al.* (2023), Jabeur *et al.* (2023), and Simsek *et al.* (2024) have also documented the superior forecasting accuracy of the XGBoost model for oil price forecasting.

[Insert Table 3 Here]

5.2 Robustness test results

We have conducted robustness test based on different train test split, robustness test based on supply shock prediction, and risk shock prediction.

Table 4 presents the results of the robustness test based on different train-test splits (60:40 and 80:20) to assess the stability of model performance. In the baseline analysis, we used a 70:30 train-test split, and the robustness test explores the effects of varying the training data proportion. For the 60:40 split, the models' performance metrics indicate that XGBoost continues to outperform other models with an accuracy of 58.6%, an F1-score of 0.563, and a ROC AUC of 0.585, maintaining its strong performance across all metrics. Random Forest (RF) follows closely with an accuracy of 57.9%, an F1-score of 0.556, and a ROC AUC of 0.578, while Naïve Bayes also shows competitive recall with an F1-score of 0.615, though its overall performance remains weaker in terms of accuracy and ROC AUC. The 80:20 split results show a slight drop in model performance across the board, likely due to the reduced training data available for the models. XGBoost still remains the top performer with an accuracy of 56.3%, an F1-score of 0.512, and a ROC AUC of 0.562, albeit with a decrease in these values when compared to the 60:40 split. Similarly, RF and Naïve Bayes exhibit a decline in performance, particularly in accuracy and ROC AUC, indicating that they are more sensitive to smaller training sets. Overall, XGBoost remains the most robust model across both split configurations, with relatively stable performance across all metrics, affirming its reliability in different scenarios. KNN consistently underperforms, with the lowest scores in both splits, further suggesting that its predictive power is weaker in this context. These results validate that while some models exhibit slight fluctuations in performance depending on the train-test split, XGBoost remains the most stable model for forecasting oil price shocks. The superior predicting accuracy of the XGBoost model for oil price forecasting has also been proven in previous studies by Gumus and Kiran (2017), Tissaoui *et al.* (2023), Jabeur *et al.* (2023), and Simsek *et al.* (2024).

[Insert Table 4 Here]

Table 5 presents the results of the robustness test based on supply shock prediction, contrasting with the baseline analysis that used demand shock prediction as the proxy for oil price shocks. In this robustness test, all models were evaluated on their ability to predict supply shocks, providing insight into how well the models adapt to a different aspect of oil price movements. The results demonstrate that XGBoost continues to outperform the other models across multiple performance metrics, achieving an accuracy of 62.0%, precision of 0.624, recall of 0.647, F1-score of 0.635, Jaccard index of 0.466, and ROC AUC of 0.619. These results suggest that XGBoost is not only the most accurate model but also excels in its ability to correctly identify both positive and negative supply shocks, as indicated by its higher recall and precision values compared to other models. Random Forest (RF) follows closely with an accuracy of 60.4%, precision of 0.611, and recall of 0.627, showcasing its strong performance in predicting supply shocks. Similarly, AdaBoost (0.595 accuracy, 0.644 recall) and LGBM (0.609 accuracy, 0.632 recall) also deliver competitive results, particularly in terms of recall, suggesting they are effective in capturing a substantial proportion of true supply shock instances. On the other hand, models such as Naïve Bayes and KNN underperform, with KNN achieving the lowest performance (accuracy of 0.510, recall of 0.512), indicating its weakness in predicting supply shocks. The Log-Loss values across all models are relatively consistent, with XGBoost showing the lowest (13.711), indicating the model's more confident predictions. Overall, the robustness test based on supply shock prediction reaffirms that XGBoost is the most robust model, performing exceptionally well across various metrics, which highlights its strong predictive capabilities.

[Insert Table 5 Here]

Table 6 presents the results of the robustness test based on risk shock prediction. In this analysis, the models were evaluated on their ability to predict risk shocks, a different aspect of oil price shocks than the demand shock used in the baseline analysis. The results highlight significant performance differences across models. XGBoost stands out as the top performer with an accuracy of 92.6%, precision of 0.925, recall of 0.897, F1-score of 0.911, Jaccard index of 0.836, and ROC AUC of 0.922, indicating its strong ability to correctly identify both risk shocks and non-risk shocks while maintaining low levels of prediction uncertainty. Similarly, Random Forest (RF) and AdaBoost show exceptional performance, with accuracy values of 92.3% and identical ROC AUC of 0.92, reflecting their excellent capacity to predict risk shocks. These models also exhibit high precision and recall, with RF achieving a recall of 0.9 and AdaBoost matching this with an F1-score of 0.908, demonstrating that they accurately capture the key risk shock instances. In contrast, SVM and KNN perform poorly in this context, with SVM showing the lowest performance across all metrics, including an accuracy of just 54.7% and recall of 0.441. Naïve Bayes also struggles, with a recall of only 0.388 and lower accuracy compared to the tree-based models, indicating that it is less effective in capturing risk shocks. ExtraTrees and LGBM also show strong performance, though slightly behind XGBoost, with LGBM achieving an accuracy of 92.4% and precision of 0.92. Overall, the robustness test based on risk shock prediction further reinforces the superior performance of XGBoost, with this model consistently leading in all key performance metrics. The XGBoost model's higher predicting accuracy for oil price forecasting has also been proven in previous studies by Gumus and Kiran (2017), Tissaoui *et al.* (2023), Jabeur *et al.* (2023), and Simsek *et al.* (2024).

[Insert Table 6 Here]

5.3 Explainable AI results

From the results of the robustness tests and performance metrics, it is evident that XGBoost emerges as the best-performing model. To gain a deeper understanding of how different features influence the model's predictions, we further explore the results using SHAP (SHapley Additive exPlanations) values. SHAP values provide a transparent method to interpret the contribution of each feature to the model's output, highlighting the individual impact of variables on the model's predictions. Figure 8 presents the SHAP values for each model and each feature used in the analysis. For XGBoost, the SHAP values suggest that certain features have a more significant influence on the model's predictions. RAI has the highest SHAP value of 21.77%, indicating its dominant role in shaping the model's output (Xiao *et al.*, 2023). Other sentiment-related features, such as ESent_Neg and ESent_Pos, also show notable SHAP values, reflecting their impact on identifying negative or positive oil price shocks. Among the economic and financial predictors, OVX shows a strong influence with a SHAP value of 16.85%, suggesting that oil market volatility is a crucial predictor for oil price movements. Similarly, USSSR with a SHAP value of 9.55% further confirms the importance of monetary policy in forecasting oil price shocks (Castillo *et al.*, 2020).

[Insert Figure 8 Here]

Figure 9 illustrates the SHAP summary plot for the XGBoost model. From the plot and Figure 8, it is evident that sentiment indicators as a group contribute 29.33% to the model's predictions, making them a crucial factor in determining oil price shocks. This underscores the importance of sentiment analysis in forecasting oil price shocks, as the model relies significantly on the emotional tone and market mood captured by these indicators. Among the sentiment variables, ESent, ESent_Neg, and ESent_Pos stand out in terms of their influence, further emphasizing how shifts in sentiment, both negative and positive, can predict oil price shocks (Li *et al.*, 2021; Zhu *et al.*, 2020; Fang *et al.*, 2023).

[Insert Figure 9 Here]

In addition to sentiment, RAI and OVX emerge as dominant factors in the model's predictions. RAI, with its strong link to global investor risk preferences, plays a key role in understanding how broader market risk impacts oil prices. Similarly, OVX, which measures oil market volatility, reflects the direct influence of market fluctuations on oil prices. The contribution of these variables aligns with the broader understanding that both economic sentiment and market volatility are critical in predicting oil price shocks. These results can be explained by the complex dynamics of the oil market, where sentiment and risk aversion often lead to significant price movements. Positive or negative shifts in sentiment can trigger reactions from investors, influencing oil demand expectations and geopolitical risks (Li *et al.*, 2021; Zhu *et al.*, 2020; Fang *et al.*, 2023). Moreover, volatility indicators such as OVX highlight how oil market fluctuations drive changes in oil prices, making these features essential.

6. Conclusion and Policy Recommendations

This study has explored the effectiveness of various machine learning models in forecasting oil price shocks, with a focus on sentiment indicators and related economic variables. The results of the forecasting analysis demonstrate the effectiveness of various machine learning models in predicting oil price shocks based on sentiment indicators and other related economic factors. The models evaluated include Logistic Regression (LR), Random Forest (RF), Support Vector Machine (SVM), Naïve Bayes, ExtraTrees, AdaBoost, XGBoost, K-Nearest Neighbors (KNN), and LightGBM (LGBM). Among these, XGBoost consistently outperforms other models across multiple evaluation metrics such as accuracy, precision, recall, F1-score, Jaccard index, and ROC_AUC. XGBoost achieved an accuracy of 60.2%, an F1-score of 0.587, and a ROC_AUC of 0.603, making it the top performer. This model's superior performance is attributed to its ability to capture complex interactions and non-linear relationships within the data, leading to improved predictions. To assess the stability of the models, robustness tests were performed using different train-test split ratios (60:40 and 80:20). The results from the robustness test show that XGBoost continue to lead in performance, with XGBoost consistently yielding the highest accuracy, F1-score, and ROC_AUC across all test splits.

The Explainable AI results further enhance our understanding of the model's decision-making process, especially for the top-performing model, XGBoost. By using SHAP (Shapley Additive Explanations) values, we gain insights into how each feature contributes to the model's predictions. From the SHAP analysis, sentiment indicators emerge as the most influential features in predicting oil price shocks, with a combined contribution of 29.33% to the XGBoost model's predictions. This demonstrates the significant impact that shifts in market sentiment—captured by sentiment indicators such as ESent (sentiment score), ESent_Neg (negative sentiment), and ESent_Pos (positive sentiment)—can have on oil price fluctuations. These findings align with existing literature, which suggests that market sentiment often drives investor behavior and thus impacts oil price movements. Moreover, RAI (Risk Aversion Index) and OVX (Oil Volatility Index) are identified as the next most important features, with RAI reflecting global risk aversion and OVX capturing oil-specific volatility. The contribution of these factors further supports the view that oil price shocks are not solely driven by supply and demand fundamentals but also by broader market dynamics such as investor risk perception and volatility.

This study provides several important policy implications for investors, regulators, and policymakers in the context of forecasting oil price shocks. For Investors, the results of this study underscore the critical role of sentiment indicators in forecasting oil price movements. Investors can benefit from integrating sentiment analysis into their decision-making processes, as it significantly influences oil price shocks. Given that sentiment, global risk aversion (RAI), and oil-specific volatility (OVX) are dominant factors, investors should closely monitor these metrics alongside traditional supply-demand indicators. Additionally, machine learning models, particularly XGBoost, which consistently outperformed other models, can be leveraged to build more robust trading strategies. Investors may also consider hedging strategies based on sentiment shifts, as positive or negative changes in sentiment can predict significant market movements. For Regulators: Regulators can use the findings of this study to better understand the dynamics of oil price volatility and the potential impact of sentiment-driven market behavior. Given that sentiment and economic uncertainty play significant roles in price fluctuations, regulators may consider implementing policies that promote transparency and stability in global markets, particularly in response to shifts in sentiment. The results also highlight the importance of closely monitoring global risk aversion and economic policy uncertainty, as these factors can exacerbate oil price volatility. Regulators might explore mechanisms to reduce excessive speculation or manage investor panic during periods of high volatility, thereby contributing to more stable oil markets.

Policymakers can use the insights from this study to design better frameworks for managing oil price fluctuations and their effects on the economy. As sentiment indicators and economic uncertainty (e.g., USEPU, UKEPU) strongly correlate with oil price shocks, policymakers should take these factors into account when crafting policies related to energy security and pricing. For instance, during periods of heightened economic policy uncertainty or risk aversion, governments could consider implementing counter-cyclical policies, such as strategic oil reserves or targeted subsidies, to buffer the impact of oil price volatility on consumers and industries. Additionally, the use of sentiment analysis tools can be integrated into decision-making processes to anticipate and respond proactively to shifts in global oil markets. As oil price shocks often have ripple effects on inflation, trade balances, and economic growth, monitoring these key sentiment and risk metrics could support more effective and timely interventions. Future studies may extend our findings by incorporating high-frequency data to capture more granular market dynamics and improve the accuracy of predictions. Additionally, exploring the application of sentiment indicators to predict broader energy prices, beyond oil, could offer valuable insights into other segments of the energy market.

References

Abdollahi, H. (2023). Oil price volatility and new evidence from news and Twitter. *Energy Economics*, 122, 106711.

Abdullah, M., Sulong, Z. and Chowdhury, M. A. F. (2024). Explainable deep learning model for stock price forecasting using textual analysis. *Expert Systems with Applications*, *249*, 123740.

Al-Fayoumi, N., Bouri, E. and Abuzayed, B. (2023). Decomposed oil price shocks and GCC stock market sector returns and volatility. *Energy Economics*, *126*, 106930.

Alquist, R. and Kilian, L. (2010). What do we learn from the price of crude oil futures? *Journal of Applied Econometrics*, 25(4), 539-573.

Alquist, R., Kilian, L. and Vigfusson, R.J. (2013). Forecasting the price of oil. In: Elliott, G. and Timmermann, A. (Eds.), Handbook of Economic Forecasting. Vol. 2, North Holland, 427-507.

Baker, S. R., Bloom, N. and Davis, S. J. (2016). Measuring economic policy uncertainty. *The quarterly journal of economics*, *131*(4), 1593-1636.

Baumeister, C. and Kilian, L. (2012). Real-time forecasts of the real price of oil. *Journal of Business & Economic Statistics*, 30(2), 326-336.

Baumeister, C., Kilian, L. and Zhou, X. (2018). Are product spreads useful for forecasting oil prices? An empirical evaluation of the verleger hypothesis. *Macroeconomic Dynamics*, 22(3), 562-580.

Bekaert, G., Engstrom, E. C. and Xu, N. R. (2022). The time variation in risk appetite and uncertainty. *Management Science*, *68*(6), 3975-4004.

Bird, S., Klein, E. and Loper, E. (2009). *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

Bouteska, A., Hassan, M. K. and Safa, M. F. (2024). Riding the waves of investor sentiment: Cryptocurrency price and renewable energy volatility during the pandemic-war era. *Journal of Behavioral and Experimental Finance*, *44*, 101001.

Breiman, L. (2001). Random forests. *Machine learning*, *45*, 5-32.

Caldara, D. and Iacoviello, M. (2022). Measuring geopolitical risk. *American Economic Review*, *112*(4), 1194-1225.

Castillo, P., Montoro, C. and Tuesta, V. (2020). Inflation, oil price volatility and monetary policy. *Journal of Macroeconomics*, *66*, 103259.

Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785-794).

Corbett, J. and Savarimuthu, B. T. R. (2022). From tweets to insights: A social media analysis of the emotion discourse of sustainable energy in the United States. *Energy Research & Social Science*, *89*, 102515.

Das, S.R. and Chen, M.Y. (2007). Yahoo! For Amazon: sentiment extraction from small talk on the web. *Management Science*, 53(9), 1375-1388.

Date, P., Mamon, R. and Tenyakov, A. (2013). Filtering and forecasting commodity futures prices under an HMM framework. *Energy Economics*, 40, 1001-1013.

Earlier studies by Gumus and Kiran (2017), Tissaoui et al. (2023), Jabeur et al. (2023), Simsek et al. (2024) also documented superior forecasting accuracy with XGBoost model for oil price forecasting.

Elliott, G., Rothenberg, T. J. and Stock, J. H. (1992). Efficient tests for an autoregressive unit root.

Fang, Y., Wang, W., Wu, P. and Zhao, Y. (2023). A sentiment-enhanced hybrid model for crude oil price forecasting. *Expert Systems with Applications*, *215*, 119329.

Geurts, P., Ernst, D. and Wehenkel, L. (2006). Extremely randomized trees. *Machine learning*, *63*, 3-42.

Ghoddusi, H., Creamer, G.G. and Rafizadeh, N. (2019). Machine learning in energy economics and finance: a review. *Energy Economics*, 81, 709-727.

Gong, X., Guan, K. and Chen, Q. (2022). The role of textual analysis in oil futures price forecasting based on machine learning approach. *Journal of Futures Markets*, 42, 1987-2017.

Gumus, M. and Kiran, M. S. (2017). Crude oil price forecasting using XGBoost. In *2017 International conference on computer science and engineering (UBMK)* (pp. 1100-1103). IEEE.

Hearst, M. A., Dumais, S. T., Osuna, E., Platt, J. and Scholkopf, B. (1998). Support vector machines. *IEEE Intelligent Systems and their applications*, *13*(4), 18-28.

Hu, J.W.-S., Hu, Y.-C. and Lin, R.R.-W. (2012). Applying neural networks to prices prediction of crude oil futures. *Mathematical Problems in Engineering*, 2012.

Hutto, C. and Gilbert, E. (2014, May). Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international AAAI conference on web and social media* (Vol. 8, No. 1, pp. 216-225).

Jabeur, S. B., Mefteh-Wali, S. and Viviani, J. L. (2024). Forecasting gold price with the XGBoost algorithm and SHAP interaction values. *Annals of Operations Research*, *334*(1), 679-699.

Jammazi, R. and Aloui, C. (2012). Crude oil price forecasting: experimental evidence from wavelet decomposition and neural network modeling. *Energy Economics*, 34(3), 828-841.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., ... and Liu, T. Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, *30*.

Kleinbaum, D. G., Dietz, K., Gail, M., Klein, M. and Klein, M. (2002). *Logistic regression* (p. 536). New York: Springer-Verlag.

Knetsch, T.A. (2007). Forecasting the price of crude oil via convenience yield predictions. *Journal of Forecasting*, 26(7), 527-549.

Krippner, L. (2013). Measuring the stance of monetary policy in zero lower bound environments. *Economics Letters*, *118*(1), 135-138.

Kumar, A. and Mallick, S. (2024). Oil price dynamics in times of uncertainty: Revisiting the role of demand and supply shocks. *Energy Economics*, *129*, 107152.

Lakatos, R., Bogacsovics, G. and Hajdu, A. (2022). Predicting the direction of the oil price trend using sentiment analysis, 2022 IEEE 2nd Conference on Information Technology and Data Science (CITDS), Debrecen, Hungary, 177-182.

Lautier, D. and Galli, A. (2004). Simple and extended Kalman filters: an application to term structures of commodity prices. *Applied Financial Economics*, 14(13), 963-973.

Li, Y., Jiang, S., Li, X. and 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.

Li, Z., Huang, Z. and Failler, P. (2022). Dynamic correlation between crude oil price and investor sentiment in China: heterogeneous and asymmetric effect. *Energies*, 15(3), 687.

Lin, L., Jiang, Y., Xiao, H. and Zhou, Z. (2020). Crude oil price forecasting based on a novel hybrid long memory GARCH-M and wavelet analysis model. *Physica A*, 123532.

Liu, B. (2012). Sentiment Analysis and Opinion Mining. Morgan & Claypool Publishers.

Lundberg, S. (2017). A unified approach to interpreting model predictions. *arXiv preprint arXiv:1705.07874*.

Luo, Z., Chen, J., Cai, X.J., Tanaka, K., Takiguchi, T., Kinkyo, T. and Hamori, S. (2018). Oil price forecasting using supervised GANs with continuous wavelet transform features. In: 2018 24th International Conference on Pattern Recognition, ICPR, IEEE, 830-835.

Medhat, W., Hassan, A. and Korashy, H. (2014). Sentiment analysis algorithms and applications: a survey. *Ain Shams Engineering Journal*, 5(4), 1093-1113.

Murphy, K. P. (2006). Naive bayes classifiers. *University of British Columbia*, *18*(60), 1-8.

Nandwani, P. and Verma, R. (2021). A review on sentiment analysis and emotion detection from text. *Social Network Analysis and Mining*, 11, 81.

Pagolu, V.S., Reddy, K.N., Panda, G. and Majhi, B. (2016). Sentiment analysis of Twitter data for predicting stock market movements. In Proceedings of the 2016 International Conference on Signal Processing, Communication, Power and Embedded System (SCOPES), Paralakhemundi, India, 3-5 October 2016, 1345-1350.

Park, J. and Ratti, R.A. (2008). Oil price shocks and stock markets in the US and 13 European countries. *Energy Economics*, 30(5), 2587-2608.

Polyzos, E. and Wang, F. (2022). Twitter and market efficiency in energy markets: Evidence using LDA clustered topic extraction. *Energy Economics*, *114*, 106264.

Ramyar, S. and Kianfar, F. (2019). Forecasting crude oil prices: a comparison between artificial neural networks and vector autoregressive models. *Computational Economics*, 53(2), 743-761.

Ready, R. C. (2018). Oil prices and the stock market. *Review of Finance*, *22*(1), 155-176.

Schapire, R. E. (2013). Explaining adaboost. In *Empirical inference: festschrift in honor of vladimir N. Vapnik* (pp. 37-52). Berlin, Heidelberg: Springer Berlin Heidelberg.

Sehgal, N. and Pandey, K. K. (2015). Artificial intelligence methods for oil price forecasting: a review and evaluation. *Energy Systems*, *6*, 479-506.

Shiller, R.J. (2017). Narrative economics. *American Economic Review*, 107(4), 967-1004.

Simsek, A. I., Bulut, E., Gur, Y. E. and Tarla, E. G. (2024). A novel approach to Predict WTI crude spot oil price: LSTM-based feature extraction with Xgboost Regressor. *Energy*, *309*, 133102.

Sudhir, P. and Suresh, V.D. (2021). Comparative study of various approaches and classifiers for sentiment analysis. *Global Transitions Proceedings*, 2, 205-211.

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

Tissaoui, K., Zaghdoudi, T., Hakimi, A. and Nsaibi, M. (2023). Do gas price and uncertainty indices forecast crude oil prices? Fresh evidence through XGBoost modeling. *Computational Economics*, *62*(2), 663-687.

Tiwari, A. K., Abakah, E. J. A., Abdullah, M. and Sulong, Z. (2023). What investors need to know about forecasting stock market return volatility using artificial intelligence.

Verleger, P.K. (1982). The determinants of official OPEC crude prices. *Review of Economics and Statistics*, 177-183.

Xiao, J., Wang, Y. and Wen, D. (2023). The predictive effect of risk aversion on oil returns under different market conditions. *Energy Economics*, *126*, 106969.

Xu, K. and Niu, H. (2022). Do EEMD based decomposition-ensemble models indeed improve prediction for crude oil futures prices? *Technological Forecasting and Social Change*, 184, 121967.

Yang, J., Geng, J. B. and Liang, Z. (2024). Time-varying effects of structural oil price shocks on financial market uncertainty. *Energy Economics*, *139*, 107910.

Yang, T., Dong, Q., Du, M. and Du, Q. (2023). Geopolitical risks, oil price shocks and inflation: Evidence from a TVP–SV–VAR approach. *Energy Economics*, *127*, 107099.

Yao, T., Zhang, Y.J. and Ma, C.Q. (2017). How does investor attention affect international crude oil prices? *Applied Energy*, 205, 336-344.

Zhang, Z. (2016). Introduction to machine learning: k-nearest neighbors. *Annals of translational medicine*, *4*(11).

Zhao, Y., Li, J. and Yu, L. (2017). A deep learning ensemble approach for crude oil price forecasting. *Energy Economics*, 66, 9-16.

Zhe, J., Lin, Z., Lingling, Z. and Bo, W. (2022). Investor sentiment and machine learning: predicting the price of China's crude oil futures market. *Energy*, 123471.

Zhou, Y., Li, T., Shi, J. and Qian, Z. (2019). A CEEMDAN and XGBOOST-based approach to forecast crude oil prices. *Complexity*, *2019*(1), 4392785.

Zhu, Z., Ji, Q., Sun, L. and Zhai, P. (2020). Oil price shocks, investor sentiment, and asset pricing anomalies in the oil and gas industry. *International Review of Financial Analysis*, *70*, 101516.

Figures

Figure 1: Word cloud of cleaned dataset

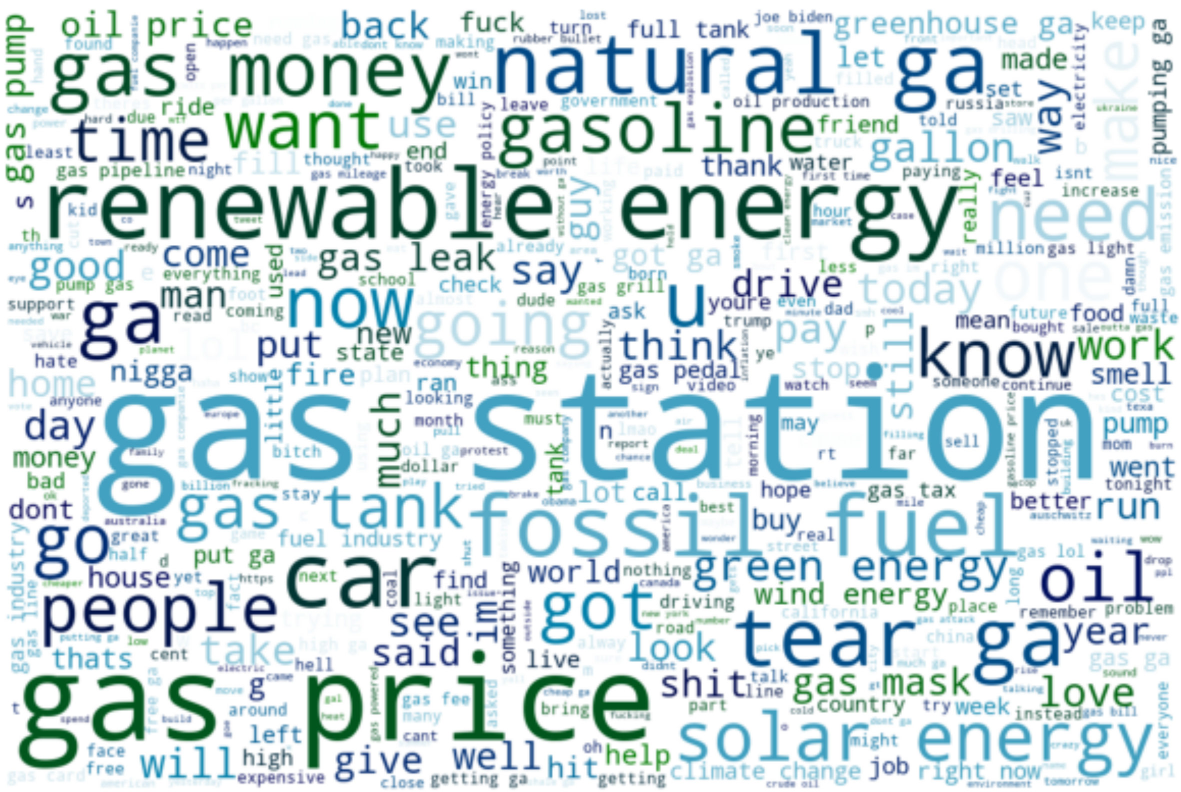


Figure 2: Sentiment distribution

Figure 3: Sentiment indicators

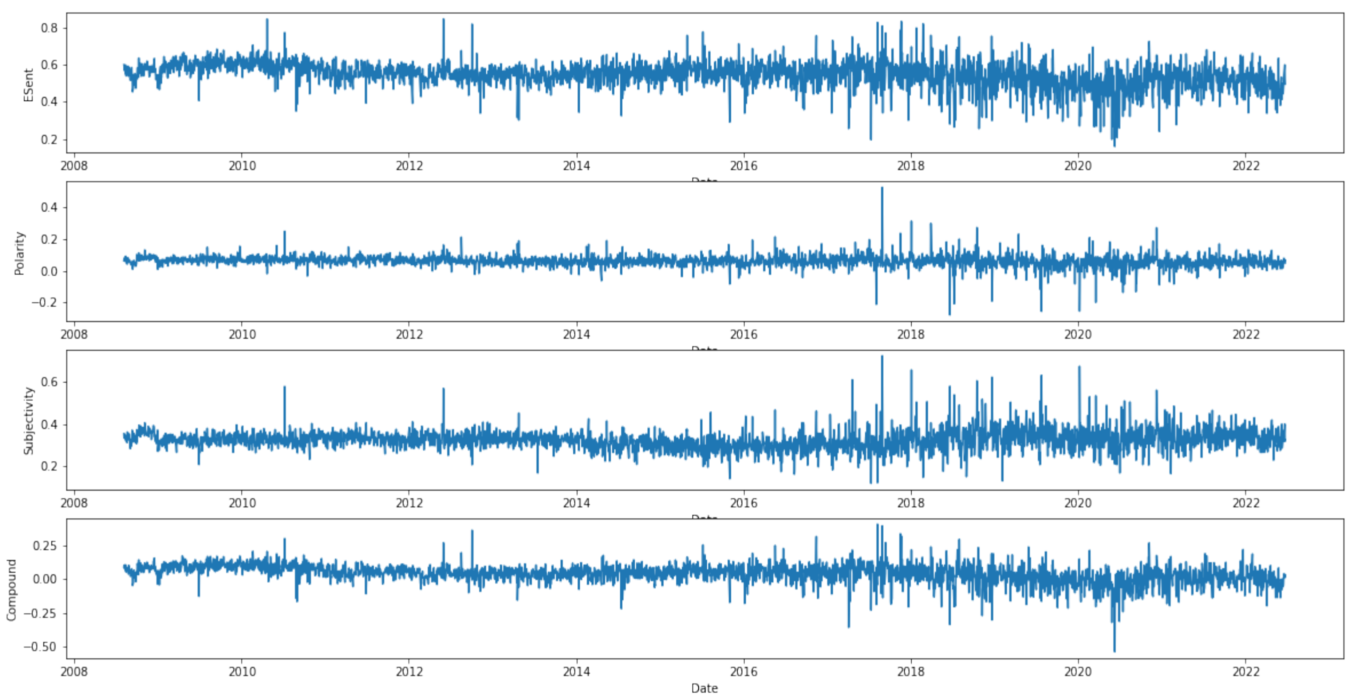


Figure 4: Oil price shock variables

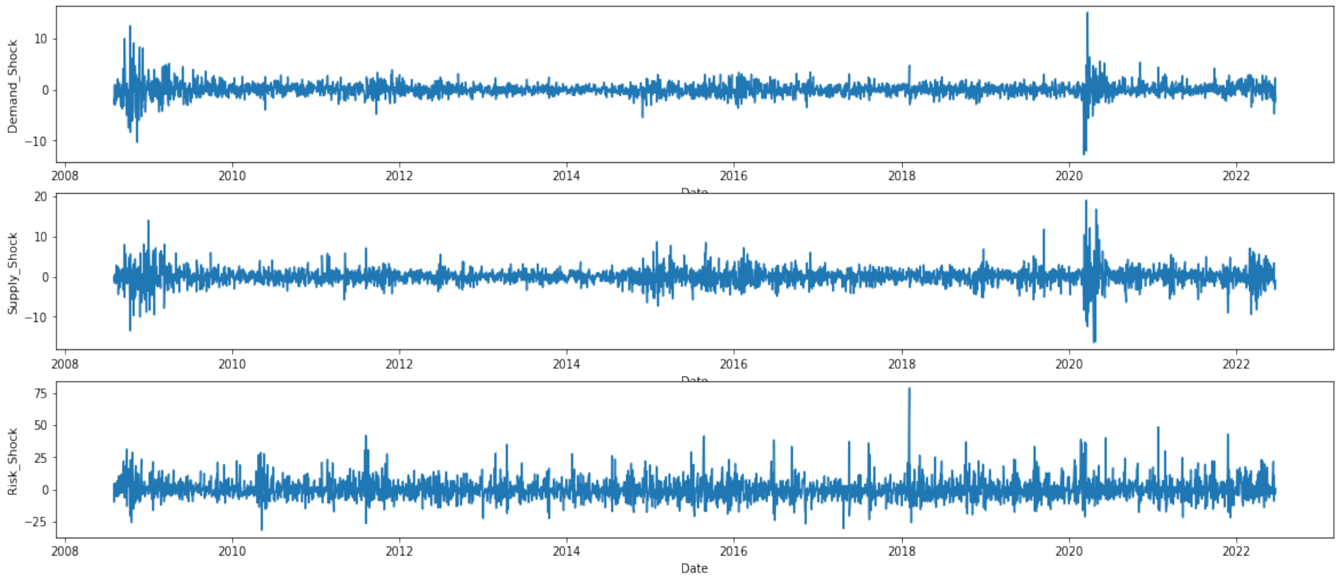


Figure 5: Correlation matrix

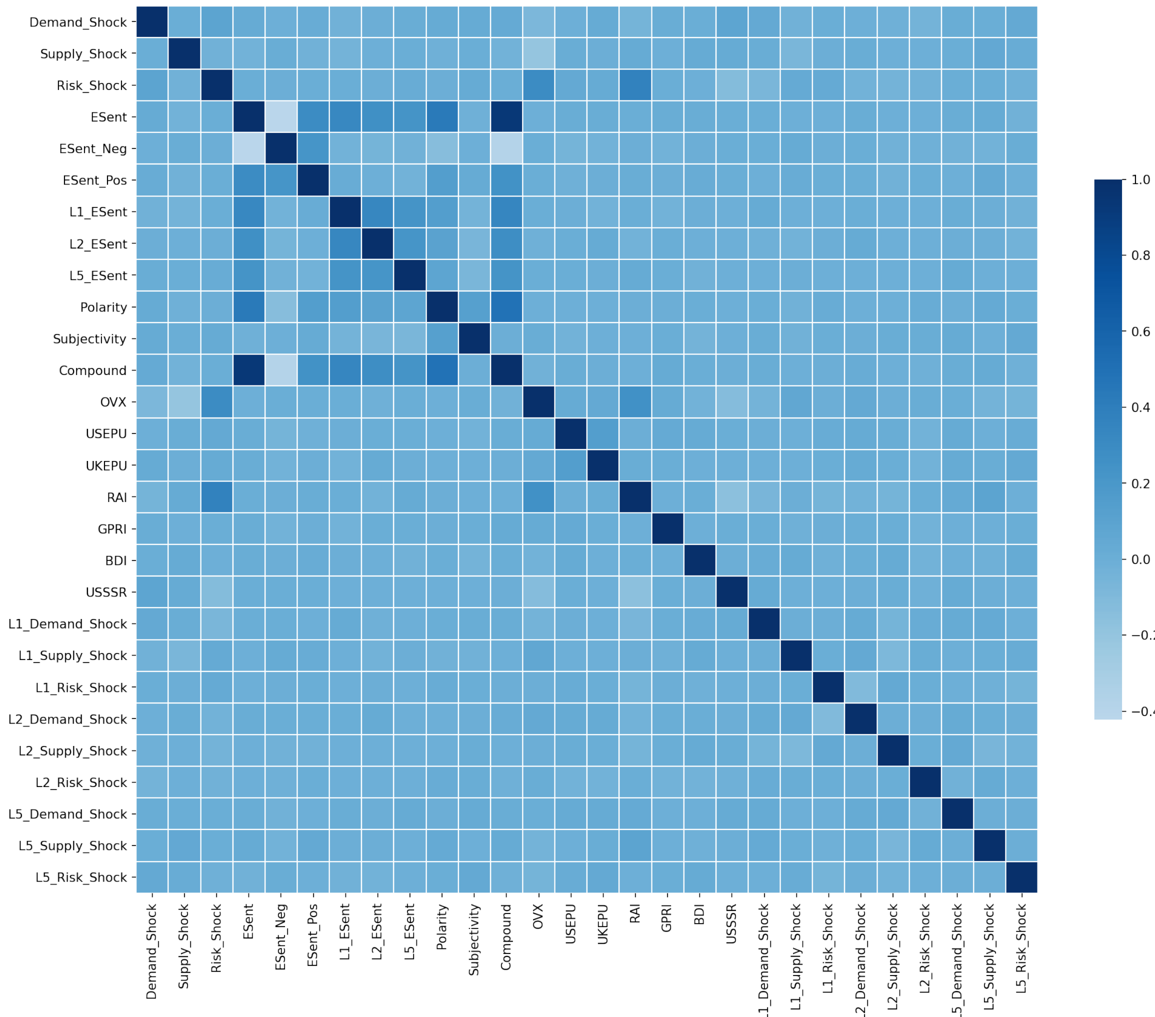


Figure 6: Confusion matrix

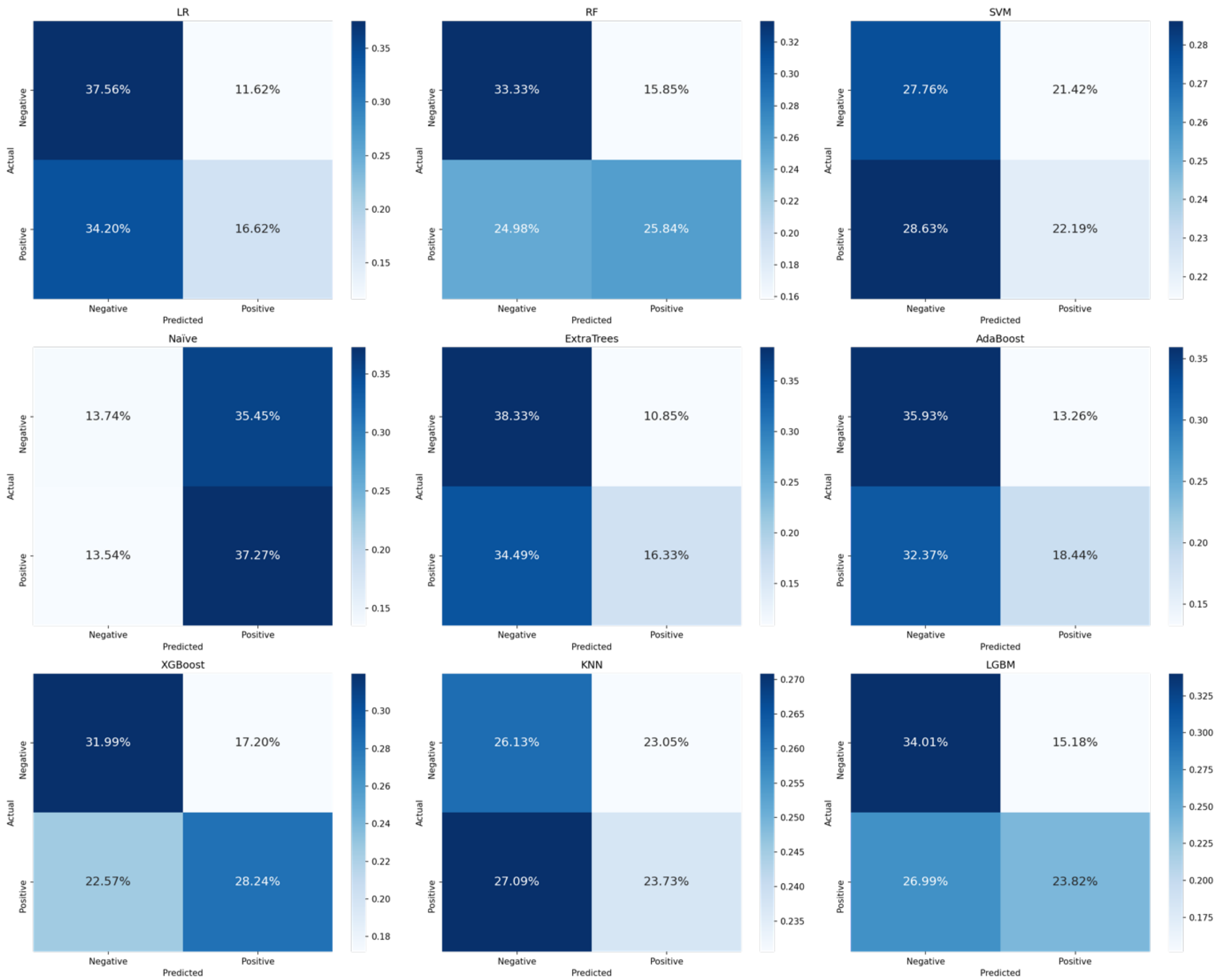


Figure 7: Receiver Operating Characteristic Curve

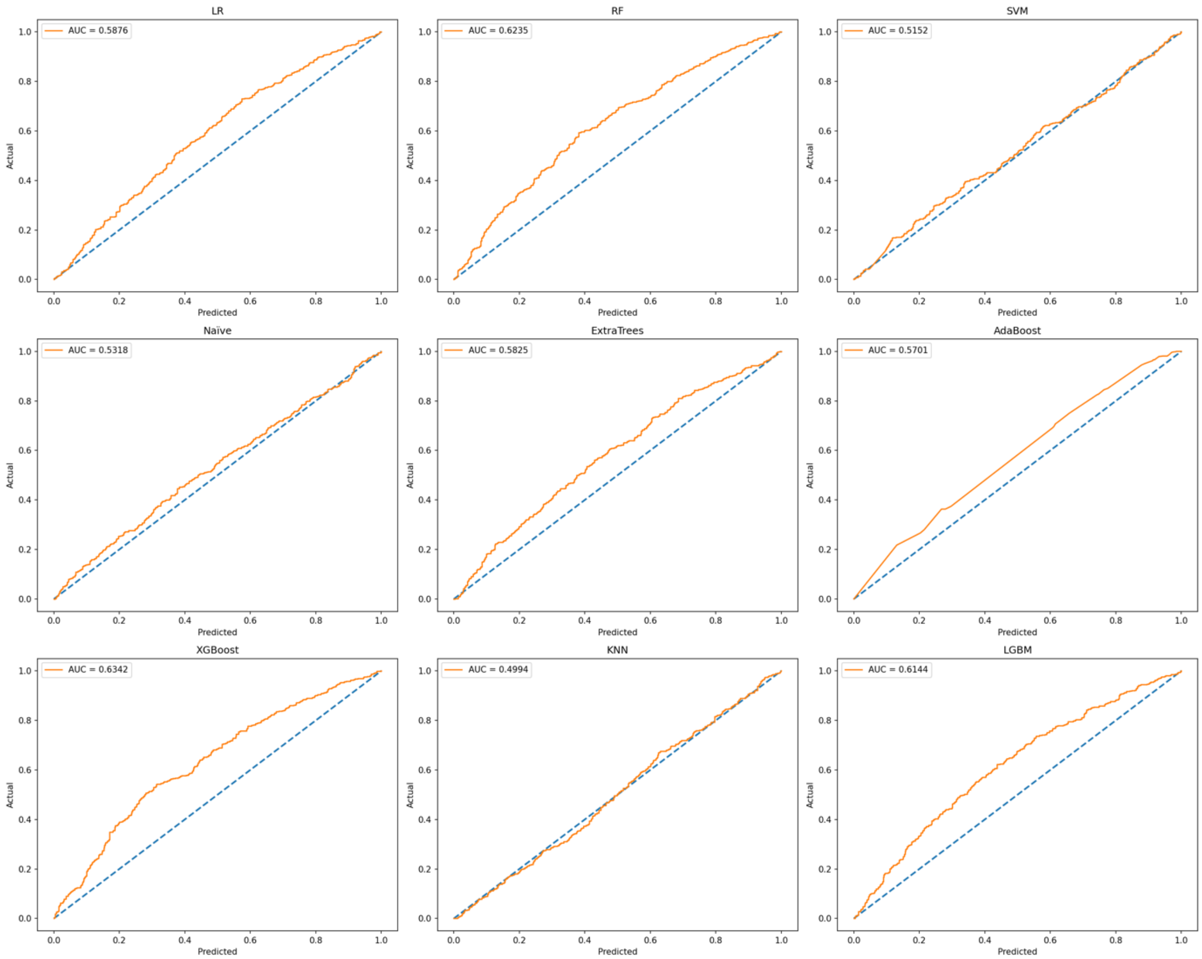


Figure 8: Variable importance

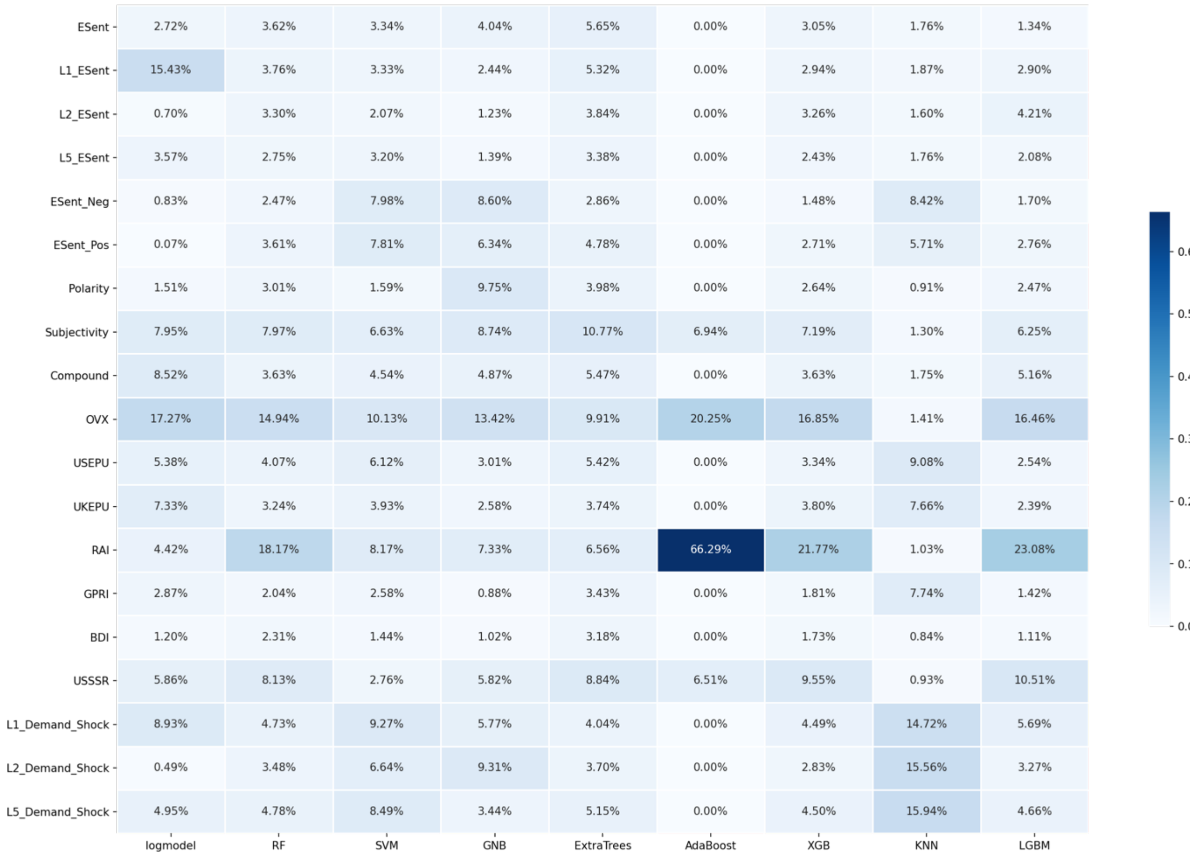
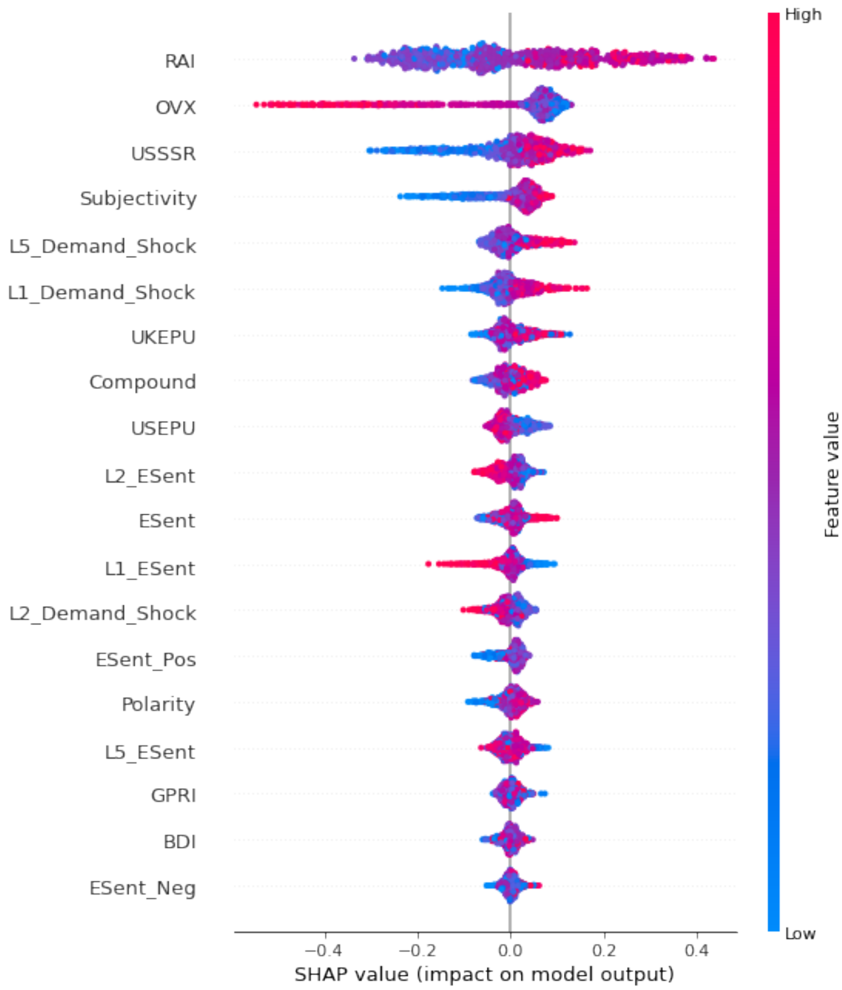


Figure 9: SHAP values of XGBOOST model



Tables

Table 1: Descriptive statistics

	Mean	Variance	Skewness	Ex.Kurtosis	JB	ERS
ESent	0.546	0.005	-0.734***	2.893***	1521.101***	-10.388***
ESent_Neg	0.072	0.197	2.721***	16.656***	44377.475***	-19.060***
ESent_Pos	0.050	0.129	2.836***	20.541***	65636.536***	-28.853***
Polarity	0.061	0.001	-0.297***	20.224***	59172.855***	-21.435***
Subjectivity	0.326	0.002	0.948***	6.824***	7249.742***	-13.349***
Compound	0.041	0.005	-0.839***	4.558***	3409.207***	-10.814***
OVX	0.000	0.004	1.702***	27.304***	109432.603***	-26.298***
USEPU	0.000	0.228	0.118***	1.989***	580.143***	-29.572***
UKEPU	0.000	0.216	-0.007	2.788***	1123.205***	-3.863***
RAI	0.000	0.007	0.352***	80.975***	947812.757***	-11.183***
GPRI	0.000	0.179	0.006	1.480***	316.460***	-18.656***
BDI	0.000	0.001	0.319***	4.363***	2810.264***	-15.607***
USSSR	0.000	0.001	0.161***	5.694***	4701.887***	-15.584***

Table 2: Hyperparameters

Model	Parameter Name & Description	Final Parameter
LR	Penalty: Regularization type; Solver: Algorithm to use for optimization	'liblinear'
RF	Bootstrap: Whether bootstrap samples are used when building trees; Max Depth: Maximum depth of the tree; Max Features: Number of features to consider; Min Samples Leaf: Minimum number of samples required to be at a leaf node; Min Samples Split: Minimum number of samples required to split an internal node; N Estimators: Number of trees in the forest	{'bootstrap': True, 'max_depth': 90, 'max_features': 3, 'min_samples_leaf': 4, 'min_samples_split': 10, 'n_estimators': 300}
SVM	C: Regularization parameter; Gamma: Kernel coefficient; Kernel: Specifies the kernel type to be used	{'C': 100, 'gamma': 0.1, 'kernel': 'rbf'}
Naïve	Var Smoothing: A regularization parameter to avoid division by zero	{'var_smoothing': 3.5111917342151275e-06}
ExtraTrees	Bootstrap: Whether bootstrap samples are used when building trees; Max Depth: Maximum depth of the tree; Max Features: Number of features to consider; Min Samples Leaf: Minimum number of samples required to be at a leaf node; Min Samples Split: Minimum number of samples required to split an internal node; N Estimators: Number of trees in the forest	{'bootstrap': True, 'max_depth': 90, 'max_features': 3, 'min_samples_leaf': 3, 'min_samples_split': 10, 'n_estimators': 300}
AdaBoost	N Estimators: The maximum number of estimators to use; Learning Rate: Weight applied to the weak learners; Algorithm: Specifies the boosting algorithm	{'algorithm': 'SAMME', 'learning_rate': 1.02, 'n_estimators': 20}
XGBoost	Max Depth: Maximum depth of the tree; Learning Rate: Step size used in gradient descent; Subsample: Proportion of the dataset used for fitting the model	{'learning_rate': 0.01, 'max_depth': 7, 'subsample': 0.5}
KNN	N Neighbors: Number of neighbors to use for classification; Weights: Weight function used in prediction; Metric: Distance metric used for the data	{'metric': 'manhattan', 'n_neighbors': 9, 'weights': 'distance'}
LGBM	Num Leaves: Maximum number of leaves in one tree; Min Child Samples: Minimum number of samples in a leaf node; Min Child Weight: Minimum sum of instance weight in a leaf node; Subsample: Fraction of data used for training; Colsample Bytree: Fraction of features used for each tree; Reg Alpha: L1 regularization term; Reg Lambda: L2 regularization term	{'colsample_bytree': 0.952164731370897, 'min_child_samples': 111, 'min_child_weight': 0.01, 'num_leaves': 38, 'reg_alpha': 0, 'reg_lambda': 0.1, 'subsample': 0.3029313662262354}

Table 3: Forecasting results

Models	Accuracy	Precision	Recall	Log-Loss	F1-score	Jaccard	ROC_AUC
LR	0.542	0.588	0.327	16.516	0.420	0.266	0.545
RF	0.592	0.62	0.509	14.715	0.559	0.388	0.593
SVM	0.500	0.509	0.437	18.039	0.470	0.307	0.501
Naïve	0.510	0.513	0.733	17.658	0.603	0.432	0.506
ExtraTrees	0.547	0.601	0.321	16.343	0.419	0.265	0.550
AdaBoost	0.544	0.582	0.363	16.446	0.447	0.288	0.547
XGBoost	**0.602**	**0.622**	**0.556**	**14.334**	**0.587**	**0.415**	**0.603**
KNN	0.499	0.507	0.467	18.074	0.486	0.321	0.499
LGBM	0.578	0.611	0.469	15.200	0.530	0.361	0.580

Table 4: Robustness test based on train test split

	60:40			80:20		
	Accuracy	F1-score	ROC_AUC	Accuracy	F1-score	ROC_AUC
LR	0.545	0.502	0.542	0.545	0.492	0.543
RF	0.579	0.556	0.578	0.537	0.498	0.536
SVM	0.521	0.493	0.519	0.524	0.494	0.523
Naïve	0.514	0.615	0.528	0.519	0.596	0.523
ExtraTrees	0.546	0.509	0.544	0.536	0.472	0.534
AdaBoost	0.556	0.537	0.555	0.556	0.533	0.555
XGBoost	**0.586**	**0.563**	**0.585**	**0.563**	**0.512**	**0.562**
KNN	0.468	0.465	0.469	0.481	0.448	0.480
LGBM	0.555	0.524	0.554	0.552	0.512	0.550

Table 5: Robustness test based on supply shock prediction

	Accuracy	Precision	Recall	Log-Loss	F1-score	Jaccard	ROC_AUC
LR	0.594	0.605	0.597	14.646	0.601	0.429	0.594
RF	0.604	0.611	0.627	14.265	0.619	0.448	0.604
SVM	0.543	0.555	0.542	16.481	0.548	0.378	0.543
Naïve	0.573	0.572	0.664	15.373	0.615	0.444	0.571
ExtraTrees	0.591	0.595	0.629	14.75	0.611	0.44	0.59
AdaBoost	0.595	0.597	0.644	14.611	0.619	0.448	0.593
XGBoost	**0.620**	**0.624**	**0.647**	**13.711**	**0.635**	**0.466**	**0.619**
KNN	0.510	0.522	0.512	17.658	0.517	0.349	0.510
LGBM	0.609	0.615	0.632	14.092	0.623	0.453	0.608

Table 6: Robustness test based on risk shock prediction

	Accuracy	Precision	Recall	Log-Loss	F1-score	Jaccard	ROC_AUC
LR	0.795	0.889	0.587	7.375	0.707	0.547	0.767
RF	0.923	0.916	0.9	2.77	0.908	0.831	0.92
SVM	0.547	0.46	0.441	16.343	0.45	0.29	0.532
Naïve	0.682	0.73	0.388	11.461	0.507	0.339	0.642
ExtraTrees	0.794	0.924	0.557	7.41	0.695	0.533	0.762
AdaBoost	0.923	0.918	0.897	2.77	0.908	0.831	0.92
XGBoost	**0.926**	**0.925**	**0.897**	**2.666**	**0.911**	**0.836**	**0.922**
KNN	0.534	0.436	0.368	16.793	0.399	0.249	0.511
LGBM	0.924	0.92	0.897	2.735	0.909	0.833	0.92

Appendix A

Table A1: Variable specification

Variable	Description	Data Source
Demand_Shock	Measures the influence of changes in global economic activity on oil prices. Binary variable created based on the positive and negative values of Demand_Shock, where the value is 0 for a negative shock and 1 otherwise.	Authors calculation form Datastream
Supply_Shock	Captures production-related disturbances affecting the oil market. Binary variable created based on the positive and negative values of Supply_Shock, where the value is 0 for a negative shock and 1 otherwise.	Authors calculation form Datastream
Risk_Shock	Reflects the impact of stock market volatility and market expectations on oil prices. Binary variable created based on the positive and negative values of Risk_Shock, where the value is 0 for a negative shock and 1 otherwise.	Authors calculation form Datastream
ESent	Overall energy sentiment index constructed using Twitter data	Authors calculation form Twitter data
ESent_Neg	Changes in number of negative tweets	Authors calculation form Twitter data
ESent_Pos	Changes in number of positive tweets	Authors calculation form Twitter data
L1_ESent	Lagged energy sentiment index by one period.	Authors calculation form Twitter data
L2_ESent	Lagged energy sentiment index by two periods.	Authors calculation form Twitter data
L5_ESent	Lagged energy sentiment index by five periods.	Authors calculation form Twitter data
Polarity	Measures the positivity or negativity of sentiment in energy-related tweets.	Authors calculation form Twitter data
Subjectivity	Assesses the degree of personal opinion or factual information in energy-related tweets.	Authors calculation form Twitter data
Compound	Aggregated sentiment score combining polarity and subjectivity.	Authors calculation form Twitter data
OVX	Oil Volatility Index indicating the market's expectation of future oil price volatility.	Datastream
USEPU	U.S. Economic Policy Uncertainty Index	Baker et al. (2016)
UKEPU	U.K. Economic Policy Uncertainty Index	Baker et al. (2016)

RAI	Global Risk Aversion Index	Bekaert et al. (2022)
GPRI	Geopolitical Risk Index	Caldara and Iacoviello (2022)
BDI	Baltic Dry Index measuring global shipping rates for dry bulk commodities.	Datastream
USSSR	U.S. shadow sort Rates indicating U.S. monetary policy stance	Krippner, L. (2013)
L1_Demand_Shock	Lagged Demand_Shock by one period.	Authors calculation form Datastream
L1_Supply_Shock	Lagged Supply_Shock by one period.	Authors calculation form Datastream
L1_Risk_Shock	Lagged Risk_Shock by one period.	Authors calculation form Datastream
L2_Demand_Shock	Lagged Demand_Shock by two periods.	Authors calculation form Datastream
L2_Supply_Shock	Lagged Supply_Shock by two periods.	Authors calculation form Datastream
L2_Risk_Shock	Lagged Risk_Shock by two periods.	Authors calculation form Datastream
L5_Demand_Shock	Lagged Demand_Shock by five periods.	Authors calculation form Datastream
L5_Supply_Shock	Lagged Supply_Shock by five periods.	Authors calculation form Datastream
L5_Risk_Shock	Lagged Risk_Shock by five periods.	Authors calculation form Datastream