Oil Price Volatility Classification and Prediction with Sentiment Analysis: Machine Learning Approach

Kerem Çiftçi   Eren Ergüzel   Abdullah Rüzgar
Mazhar Mert Ulukaya

January 13, 2025

Abstract

Accurately predicting oil price volatility is crucial for stakeholders in the energy sector—not only for financial stakeholders such as investors, policymakers, and traders but also for sectoral stakeholders, including energy companies and individuals whose businesses are affected by oil prices. This study integrates 15 years of financial news headlines over 22,000 headlines from the Oilprice.com website, sentiment analysis tools such as FinBERT, VADER, and Google Trends data, along with energy metrics from the U.S. Energy Information Administration and market metrics such as VIX and DXY, to predict average monthly oil price volatility and classify it into high and low volatility categories. It employs and compares various machine learning methods, including logistic regression, artificial neural networks (ANN), gradient boosting, elastic net regression, random forests, and LSTM. We observed that the random forest models produced robust results with strong generalization capabilities. However, more linear models, such as logistic regression and elastic net regression, exceeded our expectations. Ultimately, our analysis demonstrated that incorporating sentiment analysis significantly enhanced the predictive power of the models.

1 Introduction

In the energy sector, accurately predicting the volatility of oil prices is paramount for a wide range of stakeholders, including traders, investors, policymakers, and energy companies. Oil price volatility is influenced by a complex interplay of factors such as geopolitical events, economic indicators, and public sentiment. Effective forecasting of these fluctuations facilitates strategic decision-making, enhances risk management, and optimizes trading strategies.

In this project, we aim to both classify and predict the average volatility of oil prices for the upcoming month using an integrated approach that combines sentiment analysis with traditional and advanced machine learning techniques. For the classification task, we categorize monthly average oil price volatility into high and low categories based on historical price fluctuations. Additionally, we treat oil price volatility as a time series problem, leveraging sequential data models and LSTM to capture and incorporate the temporal dynamics inherent in the data.

To achieve these objectives, we conducted web scraping to extract 15 years of financial news headlines from the Oilprice.com website, collecting over 22,000 headlines. We utilize FinBERT, a specialized financial sentiment analysis tool, to analyze news headlines thereby capturing the sentiment that may influence price movements. Complementing this, we incorporate Google Trends data for specific keywords relevant to the energy sector to assess public interest and potential shifts in market sentiment. Furthermore, we integrate comprehensive energy data from the U.S. Energy Information Administration (EIA), including indicators such as crude oil production, inventory levels, and consumption rates, to enhance the predictive power of our models.

Our feature set is composed of historical oil price data, sentiment scores derived from FinBERT-analyzed news headlines, Google Trends indices for targeted keywords, and key energy metrics from EIA. For the classification of volatility, we establish a volatility threshold based on historical monthly price fluctuations to categorize each month's volatility as either high or low. To predict oil price changes for the next month, we employ a suite of machine learning models, including Linear Regression, Logistic Regression, Random Forests, and Neural Networks. Additionally, for the time series prediction of oil price volatility, we apply sequential data models and the DeepAR model to effectively incorporate the temporal features of the data.

To ensure the robustness and reliability of our models, we implement feature selection techniques using Lasso, Elastic Net, and Random Forest regularization methods. These methods help reduce the number of features, thereby identifying the most influential factors driving our predictions and mitigating the risk of overfitting. To validate our results and prevent the inclusion of future information during training, we utilize cross-validation techniques.

We evaluate the performance of our classification models using confusion matrices, Receiver Operating Characteristic (ROC) curves, Area Under the Curve (AUC) scores, and accuracy metrics, providing clear insights into their effectiveness. For the predictive models, we assess performance through appropriate regression metrics and the accuracy of volatility forecasts. By applying both classification and time series prediction methodologies, our project offers a comprehensive analysis of oil price volatility.

In the realm of energy trading and investment, the ability to forecast oil price volatility is critically important as it directly impacts trading strategies, hedging practices, and risk management decisions. By enhancing the accuracy of oil price volatility predictions through advanced sentiment analysis, feature selection, and comprehensive data integration, this project contributes to both the academic understanding of energy market dynamics and provides practical tools for market participants seeking to optimize their trading and investment strategies.

2 Literature Review

The literature on sentiment analysis and its application to financial markets, particularly in the context of volatility prediction, reveals a dynamic and evolving field with significant implications for understanding asset price movements. Sentiment analysis, broadly defined as the computational study of people's opinions, attitudes, and emotions toward individuals, events, or topics (Medhat et al., 2014), has gained prominence across disciplines, including finance, due to advancements in natural language processing (NLP), text mining, and machine learning (ML).

Sentiment analysis as a field has matured significantly since its initial roots in public opinion analysis in the early 20th century (Mäntylä et al., 2018). By 2004, the field experienced exponential growth, driven by the rapid evolution of computational techniques and the proliferation of unstructured data sources like social media, news, and financial reports. While early approaches focused on lexicon- and rule-based methods, modern sentiment analysis leverages ML and deep learning models such as support vector machines (SVM), long short-term memory (LSTM) networks, and transformers. These advancements have enabled sentiment analysis to capture nuanced linguistic features such as sarcasm, subjectivity, and contextual dependencies, addressing the challenges posed by informal and error-prone textual data (Nanli et al., 2012).

In financial markets, sentiment analysis has emerged as a powerful tool for predicting asset prices, volatilities, and market dynamics. The relationship between sentiment and financial performance is well-documented, with studies showing that sentiment-driven models outperform traditional quantitative approaches in various scenarios. For instance, Joseph et al. (2011) demonstrated that online search trends serve as a valid proxy for investor sentiment, while Zhang et al. (2016) highlighted the predictive power of social media sentiment in stock price movements. These studies underscore the value of sentiment as a leading indicator, capable of capturing market participants' psychological biases and emotional responses to information.

In the context of oil markets, sentiment analysis holds particular promise. Oil prices are notoriously volatile, influenced by a complex interplay of macroeconomic factors, geopolitical events, and speculative behaviors. Traditional forecasting models, such as autoregressive moving average (ARMA) and vector autoregression (VAR), often struggle to incorporate the unstructured qualitative data that influence price movements. Sentiment analysis addresses this gap by quantifying textual data, such as news and social media posts, into actionable metrics that reflect market optimism or pessimism. Studies like those by Chiong et al. (2018) and Ren et al. (2018) have integrated sentiment analysis into ML frameworks, demonstrating improved forecasting accuracy for financial assets.

Most existing studies focus on structured data, such as futures prices and macroeconomic indicators, to predict oil prices. However, the inclusion of sentiment metrics, derived from news headlines, geopolitical developments, and even weather patterns, provides a more comprehensive view of market dynamics. For example, Heston and Sinha (2017) revealed that while daily news sentiment influences short-term price movements, longer-term trends require a broader sentiment perspective. This finding aligns with broader financial market research, where positive sentiment is often associated with increased risk-taking and market optimism (Forgas, 1995; Baker and Wurgler, 2006).

Qadan and Nama (2020) investigate the impact of investor sentiment on oil prices and volatility using data from 1986 to 2016. They find that sentiment significantly predicts oil price movements and contributes to volatility.

3 Methods

We classified the volatility level above 32 as months of high volatility, and below 32 as months of low volatility.

We used a total of 4 different machine-learning algorithms to perform oil price volatility classification, and 4 different models to predict oil price volatility. These algorithms are as follows, respectively, and with their reasons,
For classification:

- **Artificial Neural Network:**Our neural network is designed with an input layer of 16 neurons, corresponding to the sixteen features utilized in our analysis. We incorporated 2 hidden layers containing 64 neurons and 32 neurons respectively. They connect to a single neuron in the output layer. The output neuron leverages a sigmoid activation function, producing values between 0 and 1. Values nearing 0 are interpreted as low oil price volatility, while those approaching 1 indicate high volatility. We applied the ReLU activation function in the hidden layer, in order to capture the nonlinear relations within the data. This architecture was specifically chosen for oil price classification to accurately model complex patterns and enhance the network's ability to distinguish between varying volatility levels. Additionally, we used earlystopping as a regularization method to reduce the risk of overfitting without gaining any additional improvement on the validation fold.

- **Random Forest Classifier:** We employed the Random Forest Classifier algorithm to identify the most effective decision tree variations for classifying oil price volatility. Its ensemble approach enhances classification accuracy and robustness by reducing overfitting and efficiently handling a large number of features. This makes Random Forest well-suited for distinguishing between high and low oil price volatility.

- **Gradient Boosting**: We selected Gradient Boosting for classifying oil price volatility because it effectively models complex and non-linear relationships by combining multiple decision trees. Its ability to reduce overfitting through built-in regularization ensures accurate and reliable predictions, even when handling a large number of features.

- **Logistic Regression:** We utilized the Logistic Regression model to classify volatility levels as high or low. The inherent characteristic of Logistic Regression, which produces output probabilities between 0 and 1, makes it ideally suited for our binary classification task.

For prediction:

- **Artificial Neural Network:**We implemented an artificial neural network consisting of one input layer with 16 neurons for sixteen features, and two hidden layers which contain 64 and 32 neurons respectively, and employed the Rectified Linear Unit (ReLU) activation function to capture non-linear relationships within the data effectively. The network architecture includes a single neuron in the output layer without an activation function, allowing it to generate continuous numerical values that correspond to the deterministic nature of volatility prediction. Additionally, we used earlystopping as a regularization method to reduce the risk of overfitting without gaining any additional improvement on the validation fold.

- **Random Forest Regression:** We utilized the Random Forest Regression algorithm to identify the most optimal decision tree configurations for predicting oil price volatility. This ensemble method enhances the model's robustness and accuracy by reducing the risk of overfitting and facilitating feature dimensionality reduction. Similar to the Random Forest Classifier, this approach leverages multiple decision trees to improve predictive performance and ensure reliable volatility estimates.

- **Gradient Boosting:**We selected Gradient Boosting for predicting oil price volatility due to its capability to model complex and non-linear relationships by leveraging an ensemble of decision trees. Its built-in regularization mechanisms help mitigate overfitting, ensuring accurate and reliable predictions, even when working with datasets containing a large number of features.

- **Elastic Net Regression:** To capture potential linear relationships in the data, we employed the Elastic Net regression model. Elastic Net combines the properties of Lasso and Ridge regression, enabling feature selection by shrinking less important feature coefficients toward zero while also maintaining a balance to avoid excessive shrinkage of correlated features. This approach mitigates the risk of overfitting and enhances the efficiency of the model by focusing on the most significant predictors of oil price volatility while accounting for multicollinearity among the features.

4 Data

Our target variable is the daily 30-day volatility, sourced from Bloomberg Terminal. Bloomberg calculates volatility by taking the logarithmic returns of consecutive days and determining the standard deviation over the preceding 30 days. However, to capture within-month variability, we compute the monthly average of daily volatility we sourced from Bloomberg Terminal.

We utilize 60 diverse predictors, broadly categorized as follows: web scraping (7), Google Trends (17), EIA monthly petroleum and other liquids production data (30), geopolitical risk (GPR), federal funds rate, US Consumer Price Index (CPI), and market variables (3). The dataset spans the period from June 2011 to September 2024.

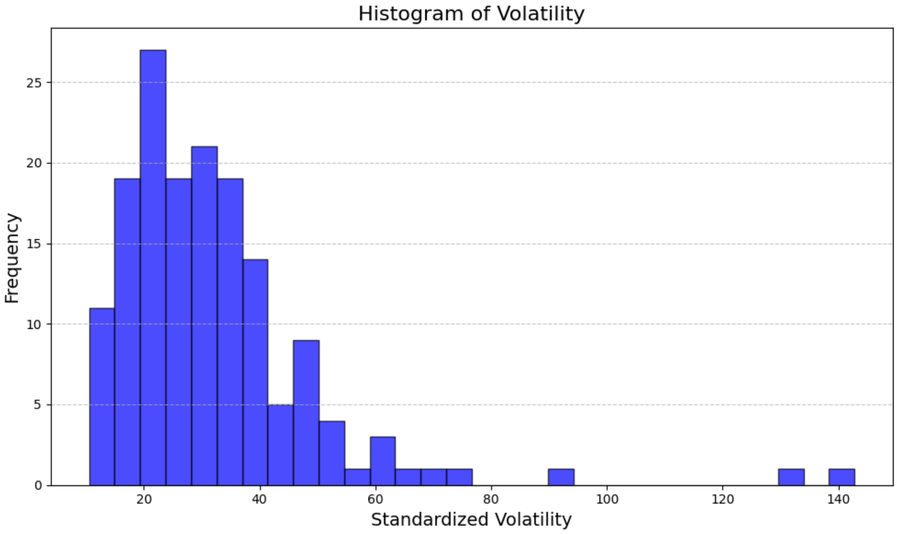


Figure 1: Histogram of Standardized Volatility

We filter these features down to the top 20 in the first step using the random forest feature importance method, specifically focusing on the classification of monthly average 30-day volatility into high or low categories. Following this, a correlation matrix of these top 20 features is presented to explore their interrelations and potential redundancy.

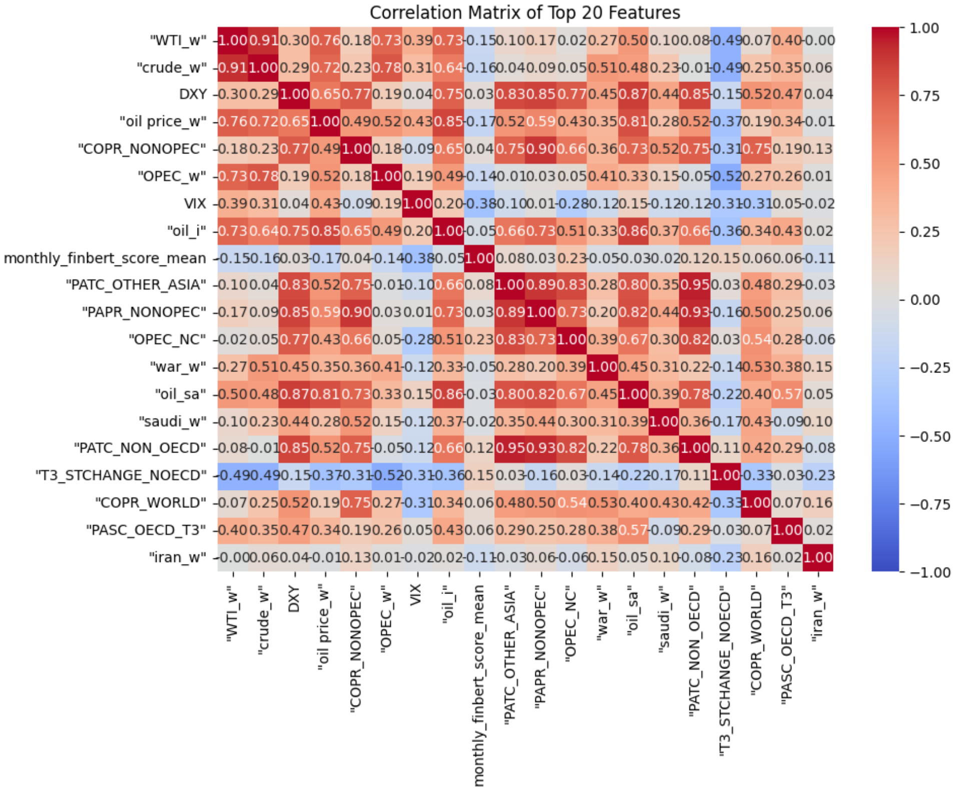


Figure 2: Correlation Matrix of Top 20 Features

Web Scraping

We extracted 15 years of financial news headlines from the oilprice.com website, totaling over 22,000 entries. Sentiments were analyzed using FinBERT (Yang et al., 2020) and VADER (Hutto Gilbert, 2014). The monthly_finbert_score_mean was derived by the summation of the average positive and negative probabilities across headlines. This variable captured meaningful insights, while headline count proved irrelevant (Feng et al., 2021; Jain Gupta, 2020).

FinBERT and VADER Algorithms

- **FinBERT**: FinBERT, a sentiment analysis model based on BERT, is tailored to financial text, such as news articles and earnings reports. It classifies sentiments as positive, negative, or neutral by interpreting domain-specific language effectively (Yang et al., 2020).

- **VADER**:VADER is a lexicon-based tool designed for social media and informal text. It computes a Compound Score (-1 to +1) to capture overall sentiment using a predefined lexicon of sentiment-weighted words (Hutto Gilbert, 2014). However this variable was filtered out during the feature selection.

Google Trends

Geopolitical tensions in Israel, Iran, and Saudi Arabia are critical for oil price volatility due to their impact on global supply. Literature highlights how conflicts and alliances in these regions influence crude oil prices (Kilian Murphy, 2014; Zhang et al., 2020). Google Trends data, such as search rates for terms like "WTI," "oil price," and "Iran," reflects market sentiment and signals significant events. Prior studies demonstrate the effectiveness of Google Trends in volatility prediction, acting as a proxy for investor attention (Da et al., 2011; Preis et al., 2013).

Feature	**Description**	**Correlation with Target (%)**
`WTI_w`	Global search rate for the term "WTI."	70%
`crude_w`	Global search rate for the term "crude."	67%
`oil price_w`	Global search rate for the term "oil price."	64%
`OPEC_w`	Global search rate for the term "OPEC."	48%
`oil_i`	Search rate for the term "oil" in Israel.	30%
`war_w`	Global search rate for the term "war."	24%
`oil_sa`	Search rate for the term "oil" in Saudi Arabia.	38%
`saudi_w`	Global search rate for the term "Saudi."	13%
`iran_w`	Global search rate for the term "Iran."	1%

Table 1: Features and Their Correlation with the Target Variable

EIA Data

We used publicly available EIA monthly petroleum and other liquids production data, introducing 30 different measures covering production, consumption, and inventory withdrawal data across various regions and products. For example, studies such as Kilian and Hicks (2013) demonstrate that OPEC and non-OPEC production levels directly affect oil price volatility by influencing global supply stability. Similarly, non-OECD consumption patterns are critical for understanding shifts in demand from emerging markets. These notions were highlighted by Fattouh and Economou (2018). These metrics provide crucial context for analyzing price dynamics.

- `COPR_NONOPEC`: Non-OPEC Crude Oil Production
- `PATC_OTHER_ASIA`: Petroleum consumption in countries excluding China and Eurasia
- `PAPR_NONOPEC`: Non-OPEC petroleum production
- `OPEC_NC`: OPEC other liquids production
- `PATC_NON_OECD`: Total non-OECD petroleum consumption
- `T3_STCHANGE_NOECD`: Non-OECD Total Crude Oil and Other Liquids Inventory Net Withdrawals
- `COPR_WORLD`: World Crude Oil Production
- `PASC_OECD_T3`: OECD inventory excluding the USA

VIX

The VIX, a proxy for market volatility, reflects expectations of near-term stock market volatility based on SP 500 index option prices (Ang et al., 2006). Similarly, the OVX, or "oil VIX," captures expected crude oil price volatility, providing critical insights for market participants. These indices are widely used to gauge market sentiment and risk.

DXY

The DXY (U.S. Dollar Index) measures the value of the U.S. dollar relative to a basket of major foreign currencies. Its fluctuations are closely linked to global oil prices due to oil's pricing in USD. Golub (1983) highlights the correlation between exchange rate movements and oil prices, emphasizing how dollar strength or weakness impacts global oil market dynamics.

GPR

The Geopolitical Risk (GPR) Index measures global geopolitical tensions and uncertainties. Developed by Caldara and Iacoviello (2022), it has been applied in various studies predicting oil price movements and volatility. For instance, Chishti et al. (2020) tested its impact on oil markets, though it was excluded during our feature selection process due to lower importance.

Feature Selection

We later filtered out correlated variables down to 16 to avoid multicollinearity.

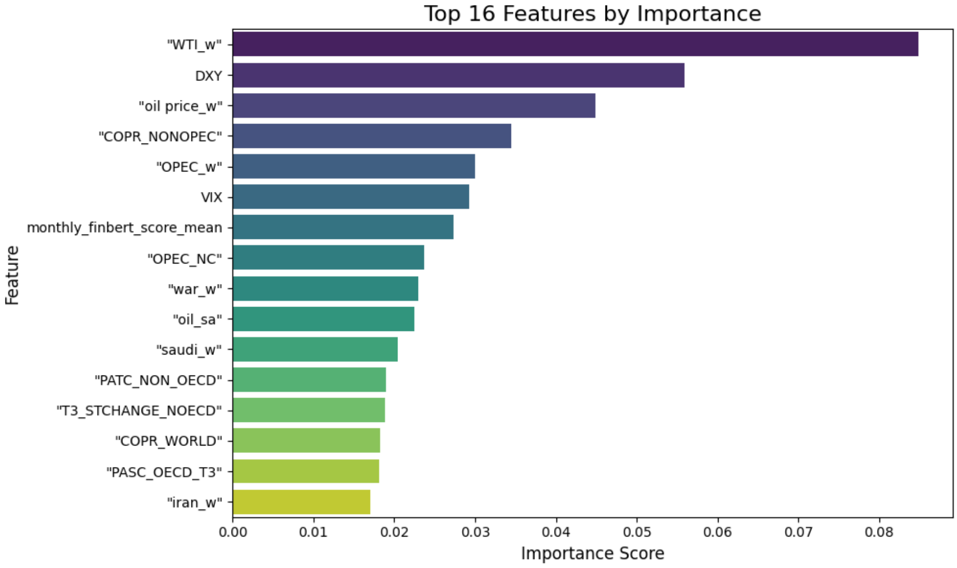


Figure 3: Top 16 Features by Importance

Feature	mean	std	min	25%	50%	75%	max	skewness	kurtosis
”WTI_w”	8.081761	8.910930	2.00	4.000000	7.00000	9.000000	100.000000	7.422553	72.514406
DXY	93.265723	8.822127	73.90	86.430000	95.46000	98.970000	112.120000	-0.408885	-0.721206
”oil price_w”	21.484277	11.065315	7.00	14.500000	20.00000	27.000000	83.000000	1.762797	6.913598
”COPR_NONOPEC”	47.337736	1.983479	43.09	46.315000	47.50000	48.905000	51.360000	-0.307704	-0.645457
”OPEC_w”	2.622642	1.931351	1.00	2.000000	2.00000	3.000000	18.000000	4.177928	26.376842
VIX	18.301258	6.869176	9.51	13.580000	16.31000	20.555000	53.540000	1.893456	5.003212
monthly_finbert_score_mean	0.305826	0.078899	0.00	0.282599	0.31074	0.350252	0.479862	-1.867226	5.537902
”OPEC_NC”	5.148113	0.165752	4.66	5.010000	5.19000	5.280000	5.490000	-0.622892	0.011505
”war_w”	27.823899	5.874785	19.00	23.000000	27.00000	31.000000	51.000000	0.961973	1.299037
”oil_sa”	53.767296	17.682930	21.00	36.500000	58.00000	66.000000	97.000000	-0.281702	-0.618921
”saudi_w”	16.081761	3.006256	12.00	14.000000	15.00000	17.000000	33.000000	1.594967	5.458205
”PATC_NON_OECD”	50.702453	3.909386	42.39	47.945000	50.90000	53.395000	58.120000	-0.154328	-0.735171
”T3_STCHANGE_NOECD”	0.069623	1.626513	-10.36	-0.470000	0.38000	0.950000	3.260000	-2.451764	11.511205
”COPR_WORLD”	75.164025	2.767633	65.64	73.035000	76.18000	77.165000	79.900000	-1.005236	0.777534
”PASC_OECD_T3”	2823.672956	162.059979	2545.00	2679.000000	2813.00000	2929.500000	3211.000000	0.422664	-0.627080
”iran_w”	8.553459	6.615716	5.00	6.000000	7.00000	9.000000	80.000000	8.475266	87.629495

Table 2: Statistical Table of the Features

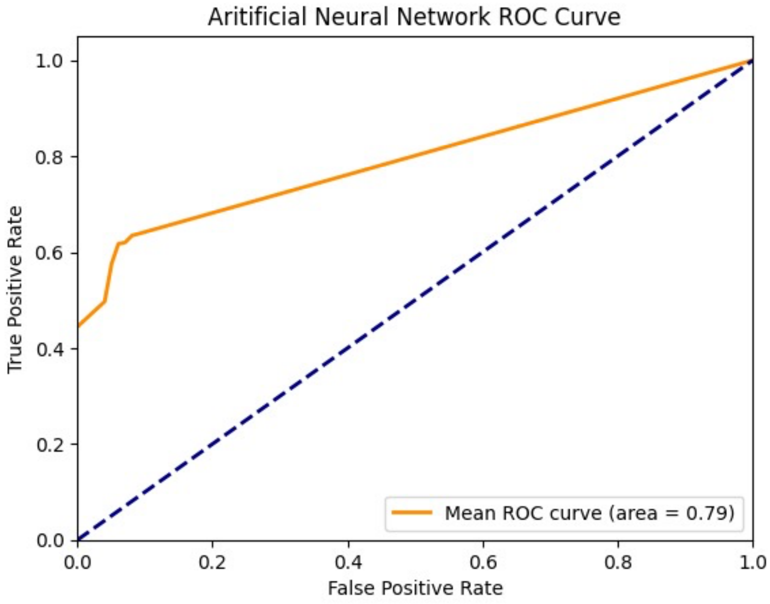
5 Empirical Results

5.1 Classification Models

Neural Network

Metric	Class -1	Class 1	Macro Avg	Weighted Avg
Precision	0.72	0.97	0.84	0.82
Recall	0.99	0.45	0.72	0.77
F1-Score	0.83	0.61	0.72	0.74
Support	94	65	159	159
Accuracy			0.77	
Mean Cross-Entropy Loss			3.5143	
Confusion Matrix			$\begin{bmatrix} 93 & 1 \\ 36 & 29 \end{bmatrix}$	

Table 3: Performance Metrics for Fold 6

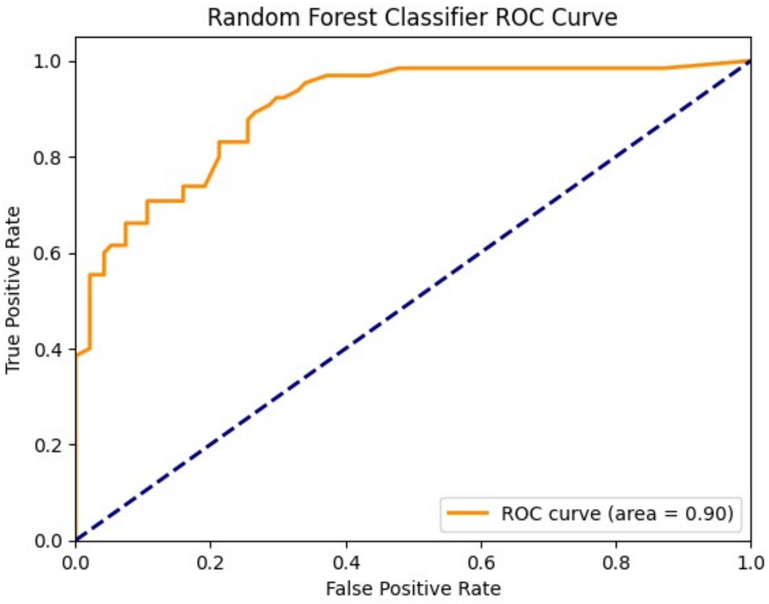


The neural network model displayed a bias toward one class, frequently predicting the lower variance class. Its complexity and non-linear nature may have led to overfitting, particularly given the data's potentially linear relationships.

Random Forest Classifier

Metric	Class -1	Class 1	Macro Avg	Weighted Avg
Precision	0.81	0.78	0.79	0.80
Recall	0.86	0.71	0.78	0.80
F1-Score	0.84	0.74	0.79	0.80
Support	94	65	159	159
Accuracy		0.80		
Mean Cross-Entropy Loss		0.5947		
Confusion Matrix		$\begin{bmatrix} 81 & 13 \\ 19 & 46 \end{bmatrix}$		

Table 4: Performance Metrics for Random Forest Classifier

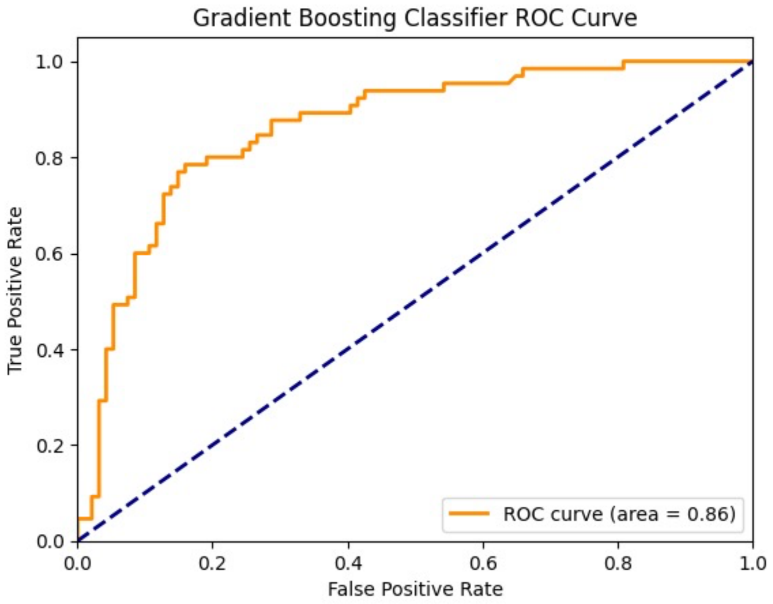


The random forest classifier model achieved the highest ROC curve, reflecting strong overall performance. However, its accuracy was slightly lower compared to other models, likely due to its low variance and high generalization characteristics.

Gradient Boosting

Metric	Class -1	Class 1	Macro Avg	Weighted Avg
Precision	0.83	0.81	0.82	0.82
Recall	0.88	0.74	0.81	0.82
F1-Score	0.86	0.77	0.81	0.82
Support	94	65	159	159
Accuracy			0.81	
Mean Cross-Entropy Loss			0.6894	
Confusion Matrix			$\begin{bmatrix} 81 & 13 \\ 17 & 48 \end{bmatrix}$	

Table 5: Performance Metrics for Gradient Boosting Classifier

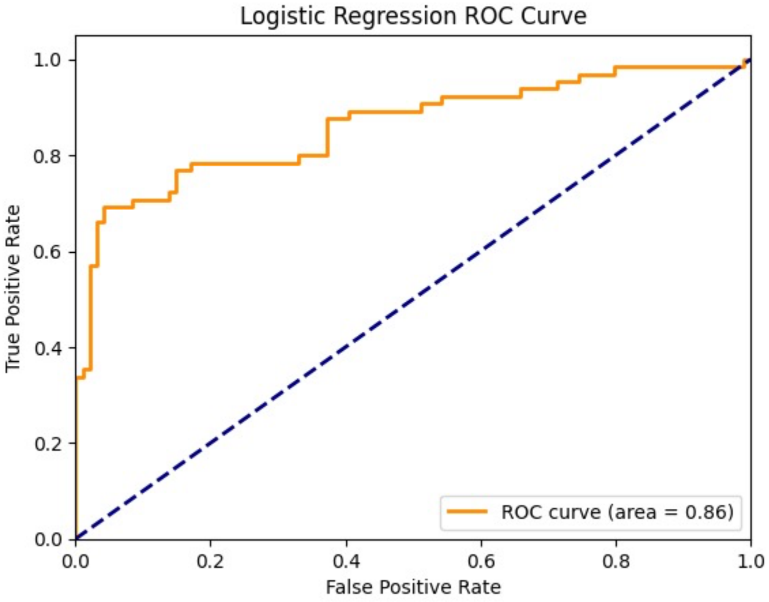


Gradient boosting demonstrated higher variance, which contributed to greater accuracy. Nevertheless, it exhibited a lower AUC and higher cross-entropy loss, likely due to the model's inherently high variance, which emphasizes capturing complex patterns.

Logistic Regression

Metric	Class -1	Class 1	Macro Avg	Weighted Avg
Precision	0.82	0.77	0.80	0.80
Recall	0.85	0.74	0.79	0.81
F1-Score	0.84	0.76	0.80	0.80
Support	94	65	159	159
Accuracy	0.81			
Mean Cross-Entropy Loss	0.4886			
Confusion Matrix	$\begin{bmatrix} 80 & 14 \\ 17 & 48 \end{bmatrix}$			

Table 6: Performance Metrics for Logistic Regression Classifier



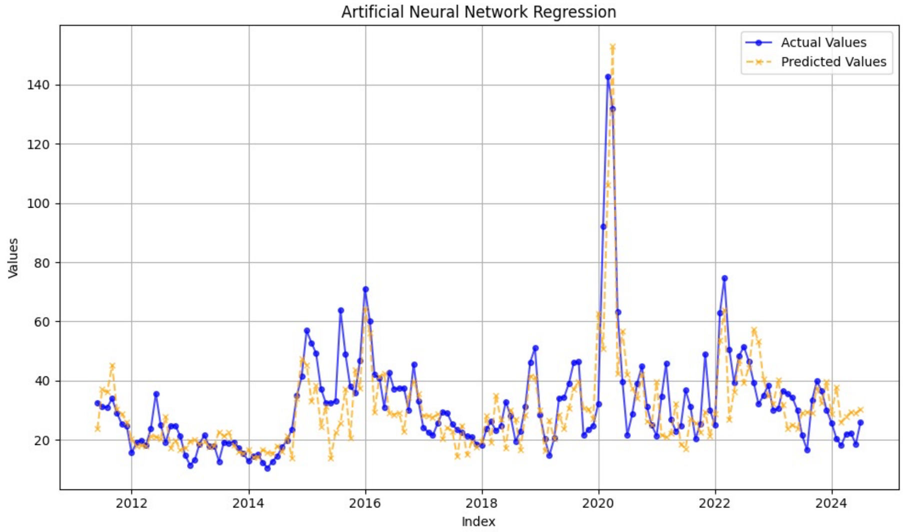
Despite its simplicity, logistic regression performed remarkably well, surpassing initial expectations. Its strong performance highlights the suitability of linear models for the dataset's underlying structure.

Classification models have performed remarkably well. This is because the data is classified into low and high volatility based on a threshold level of 32. In this classification method, high volatility months are considered "black swans," meaning they are rarer than low volatility months. In our case, there are 94 low volatility months and 65 high volatility months. These high volatility months cause the data to be skewed: when volatility is high, it is extremely high, but most of the time, it remains low.

5.2 Prediction Models

Given the data's skewed nature toward the higher volatility class, the success of classification models was anticipated. To deepen our understanding, we extended our analysis using predictive models to evaluate how effectively our data captures and explains volatility. This approach allowed us to assess not only the classification accuracy but also the explanatory power and generalization capabilities of the features in predicting volatility trends.

Neural Network

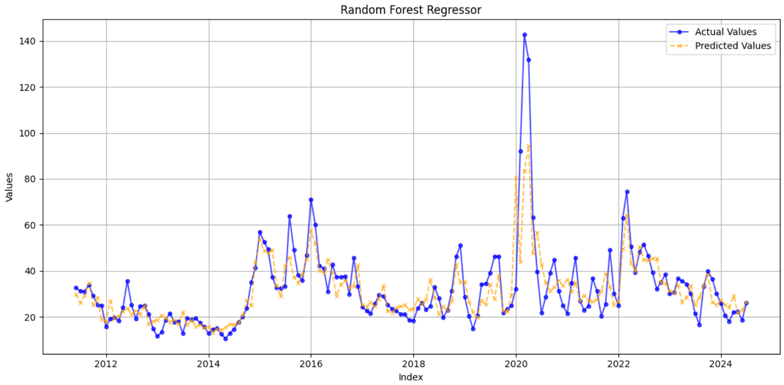


Metric	Value
Mean MSE	114.3496
Mean MAE	7.8588
Mean R^2	0.5720

Table 7: Performance Metrics for ANN

Due to the complexity of neural networks, predictions are highly variable, which may have contributed to its relatively lower performance in cross-validation. Despite employing early stopping and simplifying the architecture by reducing layers and neurons, overfitting remains a potential issue. Additionally, the non-linearities introduced by the NN may have hindered its ability to generalize effectively.

Random Forest Regression

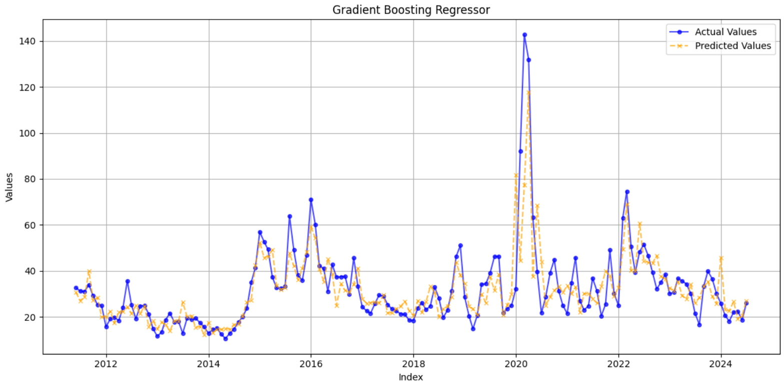


Metric	Value
Mean MSE	106.8964
Mean MAE	6.3011
Mean R^2 Score	0.6549

Table 8: Performance Metrics for Random Forest Regressor (RFR)

The random forest regressor exhibited low variability and strong generalization, resulting in robust performance. Its ability to generalize well makes it particularly suitable for larger prediction windows.

Gradient Boosting Regression

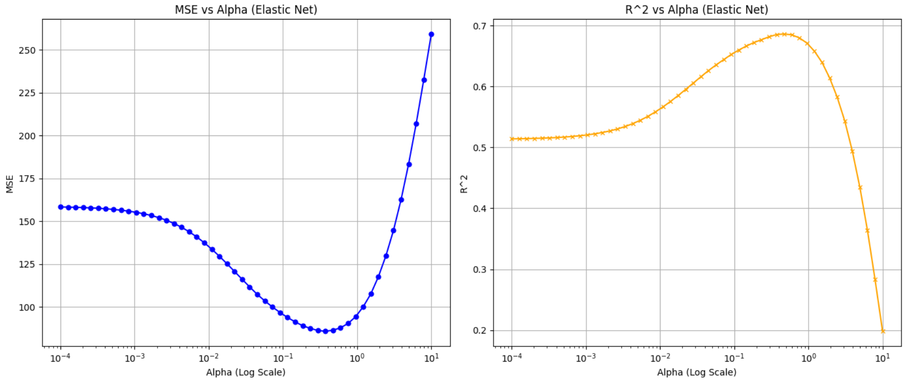


Metric	Value
Mean MSE	110.2532
Mean MAE	6.3783
Mean R^2 Score	0.6588

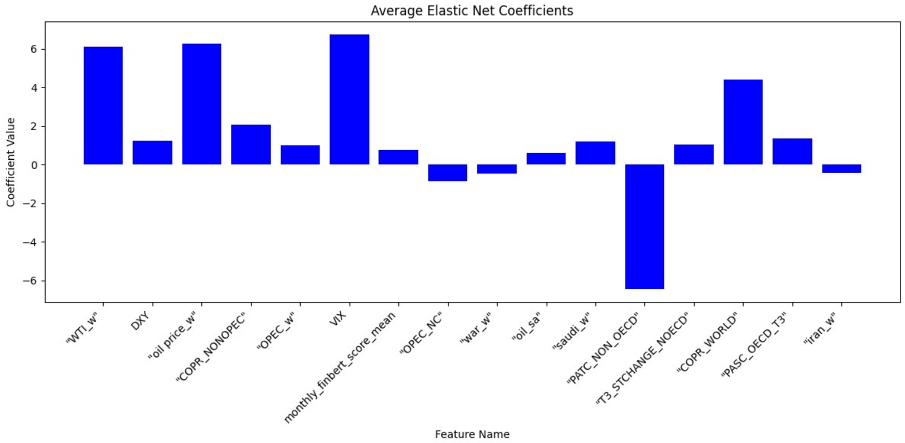
Table 9: Performance Metrics for Gradient Boosting Regressor (GBR)

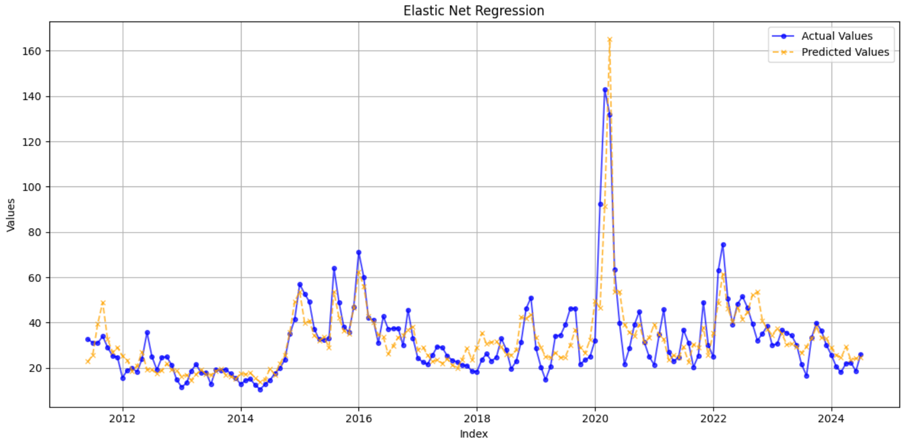
Gradient boosting demonstrated high variability, inherent to its boosting principles, but still performed effectively. While it is better at capturing outliers, its performance may be less stable over larger prediction horizons compared to the random forest regressor.

Elastic Net Regression



We employed the Elastic Net model, which requires the optimization of two hyperparameters: the L1 ratio(L1/L1+L2) and the alpha parameter. Through extensive experimentation, we determined that the optimal alpha value for the model is 0.37, as it minimizes the Mean Squared Error (MSE) while simultaneously maximizing the R^2 coefficient of determination. The corresponding weights of the model coefficients, determined based on this optimal alpha parameter, are presented below.





Metric	Value
Mean MSE	86.3683
Mean MAE	6.3558
Mean R^2 Score	0.6817

Table 10: Performance Metrics for Elastic Net (ElNet)

The elastic net model was optimized by tuning the L1 ratio to identify the best hyperparameters. The absence of zero-converging features suggests a strong linear relationship between the selected features (chosen via random forest feature selection) and the target variable. This result highlights the relevance and predictive power of the chosen features.

5.3 Comparison of prediction performance with and without sentiment data

In this section, to better investigate the relationships within the dataset, we increase the data frequency to a daily level and employ time series models such as GARCH and LSTM. These models enable us to capture sequential patterns and dependencies, facilitating a deeper exploration of the temporal structure of the data. By leveraging these approaches, we aim to identify meaningful inferences about the dynamics of volatility and its underlying drivers. Initially using only lagged volatility as a feature. We then incorporate the sentiment feature and observe the resulting change in MSE. In a sense, we try to create a "ceteris paribus" effect, isolating the influence of sentiment on volatility prediction

As our focus is on volatility prediction, we examined the popular Generalized Autoregressive Conditional Heteroskedasticity (GARCH) model, an extension of the ARCH model. GARCH posits that variance depends on both past squared residuals and past variances, capturing the phenomenon of volatility clustering. This means that periods of high volatility are typically followed by high volatility, and periods of low volatility are followed by low volatility. While GARCH models do not directly model the causes of volatility, they effectively capture the clustering effects that result from market reactions to news and behavioral feedback loops.

An intuitive explanation for this can be illustrated through a real-world example. Consider a company where a piece of news becomes public, suggesting an increase in future cash flows. This news may lead to the company's stock being perceived as underpriced, prompting traders to rush to buy the stock. In this scenario, various market participants react: traders seek to profit, existing shareholders may see an opportunity to sell, and arbitrageurs take positions in futures, spot, or options markets, potentially disrupting prices. In commodity and FX markets, companies and traders may adjust hedging strategies in response to the news, which can amplify price movements and contribute to volatility clustering. Behavioral phenomena such as herding and sentiment-driven trading also often exacerbate overreactions, leading to temporary mispricing and heightened volatility.

As a result, periods of heightened activity often lead to further volatility. Such feedback loops illustrate why volatility often persists and why the GARCH model is well-suited to capturing this volatility clustering phenomenon.

GARCH Model with Sentiment Features

We fit a GARCH(1,1) model with an exogenous variable to predict volatility using the variance equation.

$$\sigma_t^2 = \omega + \alpha_1 \epsilon_{t-1}^2 + \beta_1 \sigma_{t-1}^2 + \gamma X_t$$

In this model, volatility is modeled via the return variance equation. We incorporated an exogenous variable, a sentiment feature called "len_news," which represents the number of news articles about oil on a given day. Despite the coefficient for this feature being low but positive, it was statistically significant, as indicated by a very low p-value. This aligns with the idea that sentiment data can reflect market attention, which indirectly impacts volatility.

Parameter	**Estimate**	**Std. Error**	**t-Value**	**p-Value**
mu	0.000539	0.000015	35.9070	0e+00
omega	0.000014	0.000003	4.9204	1e-06
alpha1	0.160694	0.011786	13.6343	0e+00
beta1	0.802931	0.007396	108.5700	0e+00
vxreg1	0.000001	0.000000	1699.8167	0e+00

Table 11: GARCH Model Parameters

We then extended the analysis by testing an LSTM model, which is well-suited for sequential time-series data. Initially, we constructed a baseline model using only lagged volatility to predict future volatility. Subsequently, we added the sentiment feature ("len_news") and observed an improvement in predictive performance.

Sentiment data alone (e.g., Google Trends data or financial news headline sentiment analysis) is a relatively weaker predictor compared to lagged volatility or other strong features. However, adding sentiment data to the model enhanced its predictive ability, as expected.

Model	**Features**	**RMSE**
Random Forest	"len news" and "daily volatility lag 1"	2.19
	Only "daily volatility lag 1"	2.2
LSTM	len news" and "daily volatility lag 1"	6
	Only "daily volatility lag 1"	8

Table 12: Comparison of RMSE Across Models and Features

We also compared predicted vs. actual volatility using cross-validation. To avoid false assumptions or being misled by RMSE from split validation approach, we performed cross-validation tailored for time-series data, preserving the temporal order. We refrain from using traditional methods like LOOCV or K-fold to ensure the time-series structure was not violated. The inclusion of sentiment improves the RMSE and MSE, indicating that it captures additional variance. However, the improvement is marginal, suggesting that the "len_news" feature is likely secondary to core predictors like lagged volatility.

Model	**Features**	**Mean MSE**
LSTM	"daily_volatility_lag_1", "len_news"	241
	"daily_volatility_lag_1"	272

Table 13: Cross Validation Results

Our results suggest that sentiment data (e.g., Google Trends, financial news headline analysis) is an important addition to volatility modeling and prediction. However, it is not the dominant predictor. While some news significantly influence volatility, many news articles provide redundant or already-known information, limiting their impact.

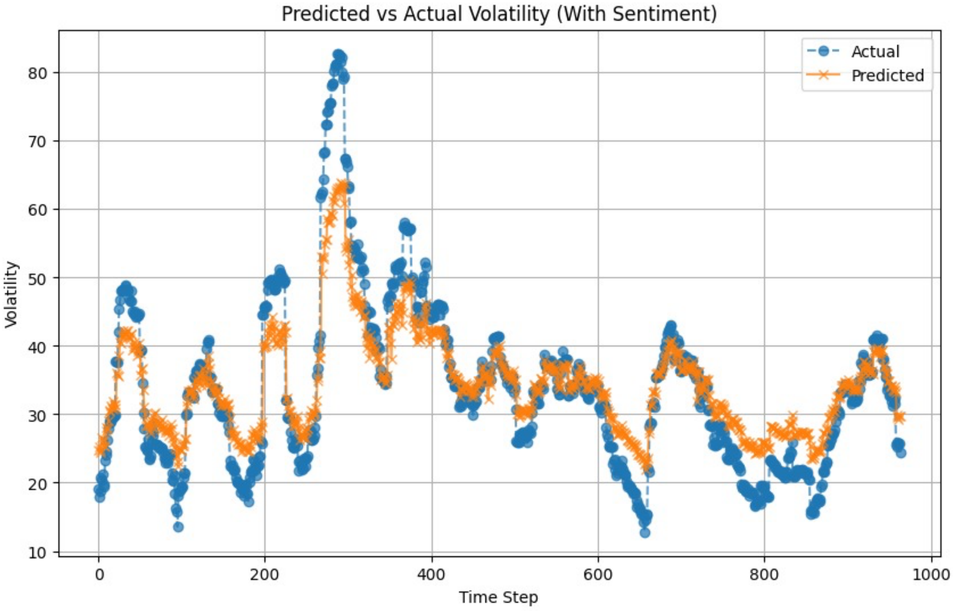


Figure 4: Predicted vs Actual Volatility (With Sentiment)

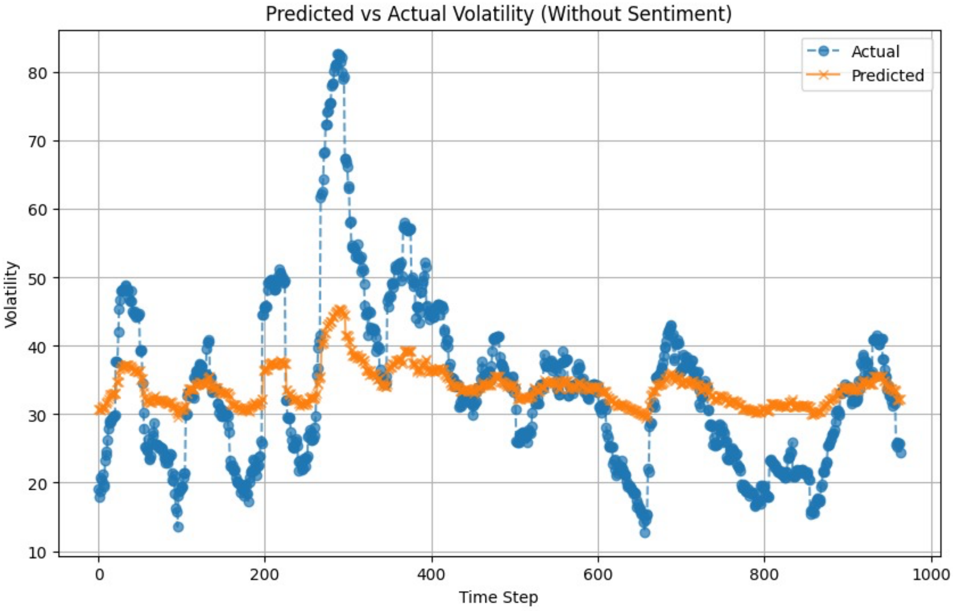


Figure 5: Predicted vs Actual Volatility (Without Sentiment)

6 Comments

We did not anticipate achieving such high accuracy scores and performance metrics, including the AUC score and a low loss function. However, despite further investigations, we were unable to identify any issues in the in-sample testing. This outcome may stem from the limited number of data points available in our dataset. Additionally, the dataset is not normally distributed, as evidenced by the histogram of our data (Figure 1), which could introduce some disturbance in our models.

Moreover, we were not expecting linear and relatively less complex models, such as logistic regression and Elastic Net regression, to perform so effectively. This observation suggests that starting with highly complex and non-linear algorithms may not always be the best approach, as these models may fail to capture the linear relationships within the data and are prone to overfitting.

The reliable performance of our predictive models indicates that our classification models are also dependable. As a result, they complement each other effectively.

The fact that our subset of the 16 most important features produced promising results demonstrates the robustness of our dataset across various machine learning models.

Furthermore, our final analysis highlighted that sentiment analysis is systematically significant, even within a time-series context. Our models demonstrated that sentiment data enhances the predictive power of sequential analyses as well. This finding underscores the potential for further exploration into the time-series analysis of sentiment data and its relationship with oil price volatility.

7 Conclusion

In summary, this study explored the dual application of classification and prediction models to forecast volatility and assess the explanatory power of sentiment data. The skewed nature of the dataset toward higher volatility made classification models particularly effective, with the random forest achieving the highest ROC curve, reflecting robust generalization capabilities. Gradient boosting excelled in capturing complex patterns but faced stability challenges, while logistic regression surpassed expectations, proving the suitability of linear models for this dataset. Neural networks, despite their complexity, struggled with overfitting and class bias, underscoring their limitations for datasets with potentially linear relationships.

The transition to predictive models highlighted the dataset's ability to model volatility. Random forest demonstrated strong generalization and reliability, while elastic net regression outperformed it by better leveraging the selected features, indicating a linear relationship in the data. Gradient boosting excelled in capturing outliers but showed less stability over longer windows, and neural networks struggled with generalization despite adjustments.

While random forest remains a robust choice for both prediction and classification, elastic net regression's superior performance can be attributed to its regularization properties, which align well with the already selected features.

Our findings also emphasize that sentiment data, while valuable in enhancing volatility modeling, plays a complementary rather than dominant role. The redundancy and repetitive nature of many sentiment features limit their standalone impact, as markets often preemptively price in their effects.

Future research could focus on integrating time-series models, such as LSTM or RNN, to capture market momentum more effectively. Further, augmenting the dataset with additional features and employing advanced feature engineering techniques could enhance the models' predictive accuracy. Testing these models in real-world scenarios or simulated environments will provide a deeper understanding of their practical applications, paving the way for more robust volatility prediction frameworks.

References

[1] Ang, A., Hodrick, R. J., Xing, Y., Zhang, X. (2006). The cross-section of volatility and expected returns. *The Journal of Finance, 61*(1), 259-299.

[2] Baker, M., Wurgler, J. (2006). Investor sentiment and the cross-section of stock returns. *The Journal of Finance, 61*(4), 1645-1680.

[3] Baumeister, C., Kilian, L. (2016). Forty years of oil price fluctuations: Why the price of oil may still surprise us. *The Journal of Economic Perspectives, 30*(1), 139-160.

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

[5] Chishti, M., Mehmood, S., Butt, H. A. (2020). Geopolitical risk and its impact on the oil market: Evidence from GPR index. *Resources Policy, 66*(101613), 1-10.

[6] Chiong, R., Fan, Z., Hu, Z., Adam, M. T., Lutz, B., Neumann, D. (2018). A sentiment analysis-based machine learning approach for financial market prediction via news disclosures. In *Proceedings of the Genetic and Evolutionary Computation Conference Companion* (pp. 278-279). New York, NY: ACM.

[7] Da, Z., Engelberg, J., Gao, P. (2011). In search of attention. *The Journal of Finance, 66*(5), 1461-1499.

[8] Feng, S., Xu, Y., Li, S. (2021). Sentiment analysis for volatility prediction in financial markets using BERT models. *Journal of Financial Data Science, 3*(2), 45-62.

[9] Fattouh, B., Economou, A. (2018). OPEC at the crossroads. *Oxford Institute for Energy Studies.*

[10] Forgas, J. P. (1995). Mood and judgment: The affect infusion model (AIM). *Psychological Bulletin, 117*(1), 39-66.

[11] Geman, H., Ohana, S. (2009). Forward curves, scarcity, and price volatility in oil and natural gas markets. *Energy Economics, 31*(4), 576-585.

[12] Hamilton, J. D. (2009). Causes and consequences of the oil shock of 2007-08. *Brookings Papers on Economic Activity, 40*(1), 215-261.

[13] Heston, S. L., Sinha, N. R. (2017). News vs. sentiment: Predicting stock returns from news stories. *Financial Analysts Journal, 73*, 67-83.

[14] Hutto, C. J., Gilbert, E. (2014). VADER: A parsimonious rule-based model for sentiment analysis of social media text. *Proceedings of the International AAAI Conference on Web and Social Media, 8*(1), 216-225.

[15] Jain, P., Gupta, A. (2020). Using social media sentiment analysis to predict stock market volatility. *International Journal of Forecasting, 36*(3), 816-835.

[16] Joseph, K., Wintoki, M. B., Zhang, Z. (2011). Forecasting abnormal stock returns and trading volume using investor sentiment: Evidence from online search. *International Journal of Forecasting, 27*, 1116-1127.

[17] Kilian, L., Hicks, B. (2013). Did unexpectedly strong economic growth cause the oil price shock of 2003–2008? *Journal of Forecasting, 32*(5), 385-394.

[18] Kilian, L., Murphy, D. P. (2014). The role of inventories and speculative trading in the global market for crude oil. *Journal of Applied Econometrics, 29*(3), 454-478.

[19] Lakatos, R., Bogacsvics, G., Hajdu, A. (2022). Predicting the direction of the oil price trend using sentiment analysis. In *Proceedings of the 2022 IEEE 2nd Conference on Information Technology and Data Science (CITDS)* (pp. 16-18). IEEE.

[20] Mäntylä, M. V., Graziotin, D., Kuutila, M. (2018). The evolution of sentiment analysis—A review of research topics, venues, and top cited papers. *Computer Science Review, 27*, 16-32.

[21] Nanli, Z., Ping, Z., Weiguo, L., Meng, C. (2012). Sentiment analysis: A literature review. In *Proceedings of the 2012 International Symposium on Management of Technology (ISMOT)* (pp. 572-576). Hangzhou, China.

[22] Preis, T., Moat, H. S., Stanley, H. E. (2013). Quantifying trading behavior in financial markets using Google Trends. *Scientific Reports, 3*(1684), 1-6.

[23] Qadan, M., Nama, H. (2018). Investor sentiment and the price of oil. *Energy Economics, 69*, 42-58.

[24] Ren, R., Wu, D. D., Liu, T. (2018). Forecasting stock market movement direction using sentiment analysis and support vector machine. *IEEE Systems Journal, 13*, 760-770.

[25] Sadik, Z., Date, P., Mitra, G. (2019). Forecasting crude oil futures prices using global macroeconomic news sentiment. *IMA Journal of Management Mathematics, June.*

[26] Yang, Y., Uy, M., Huang, Y. (2020). FinBERT: A pre-trained financial language representation model for financial text mining. *arXiv preprint arXiv:2006.08094*.

[27] Zhang, L., Liang, Z., Xiao, K., Liu, Q. (2016). Forecasting price shocks with social attention and sentiment analysis. In *Advances in Social Networks Analysis and Mining (ASONAM)* (pp. 559-566). IEEE/ACM.

[28] Zhang, D., Broadstock, D. C., Cao, H. (2020). Impact of geopolitical risks on crude oil returns and volatility. *Energy Economics, 86*(104619), 1-10.