https://doi.org/10.62836/iaet.v2i1.265

Global Science Publishing

IAET

Innovations in Applied Engineering and Technology

https://ojs.sgsci.org/journals/iaet

Article

Integrating Textual Analytics with Time Series Forecasting Models: Enhancing Predictive Accuracy in Global Energy and Commodity Markets

Jiarui Rao [1,†], Qian Zhang [2,*,†], Zong Ke [3], Shaoyu Liu [4] and Xinqi Liu [5]

[1] *Uber Technologies Inc., LA, USA*

[2] *Tencent Inc., Shenzhen, China*

[3] *Faculty of Science National University of Singapore, Singapore*

[4] *Columbia University, USA*

[5] *Western University, ON, Canada*

† *Co-first author, these authors contributed equally to this work.*

Abstract: This study presents a comprehensive framework for predicting crude oil prices by integrating textual features extracted from news headlines into a time series forecasting model. The rationale for using headlines instead of full articles is twofold: headlines encapsulate the essence of the news, and the approach aligns with previous research by Li et al. The focus on futures news over gold news is justified by the larger dataset and the complex interrelations between futures prices, including gold, natural gas, and crude oil. The methodology involves extracting thematic and sentiment information from news headlines using text mining techniques, constructing daily topic strength indices, and developing an emotional strength index that accounts for the decay effect of news influence over time. The study employs a vector autoregression model to determine the optimal lags for various exogenous sequences, including topics and sentiment indices, relative to crude oil prices. The forecasting model is trained using machine learning techniques such as Random Forest Regression (RF), Support Vector Regression (SVR), Autoregressive Integrated Moving Average (ARIMA), and their extended versions with exogenous variables (ARIMAX). The performance of the models is evaluated using metrics like Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE). The results indicate that incorporating textual features significantly improves the prediction accuracy of RF, SVR, and AdaBoost models, while the traditional ARIMA model performs well without textual features. The study also introduces a novel approach combining Ensemble Empirical Mode Decomposition (EEMD) with Independent Component Analysis for analyzing non-linear and non-stationary time series data, specifically applied to gold price analysis. The EEMD-BPNN-ADD model is identified as the most accurate for forecasting, with interval predictions provided for gold prices. This research contributes to the field by demonstrating the effectiveness of integrating textual analysis with traditional financial models for improved market forecasting.

Keywords: financial time series; decomposition techniques; trend analysis; seasonality; forecasting; risk management

Received: 1 November 2023; Accepted: 27 November 2023.

* Corresponding: Qian Zhang

1. Introduction

The volatility and complexity of financial markets, particularly in commodities like crude oil, pose significant challenges for economists, investors, and policymakers. Accurate forecasting of crude oil prices is crucial for strategic planning and risk management in the energy sector. Traditional models often fall short in capturing the dynamic interplay between market forces and the influence of global events, which are increasingly reflected in the vast amount of textual data available in news and social media. This research introduces a novel approach to crude oil price prediction by integrating textual features extracted from news headlines into a time series forecasting framework [1–4].

The study is motivated by the recognition that news headlines, as concise summaries of news articles, encapsulate the most critical information that can influence market sentiments and price movements. By focusing on headlines rather than full articles, the research streamlines the analysis while capturing the essence of news content. The choice to utilize futures news over gold news is strategic, considering the larger dataset and the intricate relationships between various futures markets, including those for gold, natural gas, and crude oil [5–9].

Methodologically, the research employs text mining techniques to extract thematic and sentiment information from news headlines, which are then translated into quantifiable features for forecasting models. The construction of daily topic strength indices and an emotional strength index that accounts for the decay effect of news influence over time represents an innovative approach to incorporating textual data into financial modeling.

The forecasting framework leverages machine learning algorithms, including Random Forest Regression (RF), Support Vector Regression (SVR), and Autoregressive Integrated Moving Average (ARIMA) models, to predict crude oil prices based on the extracted textual features. The performance of these models is rigorously evaluated using standard metrics such as Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE) [10–15].

In addition to crude oil price prediction, the study extends its scope to gold price analysis, introducing a new method that combines Ensemble Empirical Mode Decomposition (EEMD) with Independent Component Analysis. This approach is particularly suited for handling the non-linear and non-stationary characteristics of financial time series data.

The research contributes to the literature by demonstrating the potential of integrating textual analysis with traditional financial models, offering a more comprehensive and nuanced approach to market forecasting. The findings have practical implications for investors and policymakers seeking to navigate the complexities of the global energy market.

Step is critical for determining the optimal lag, which captures the dynamic interdependencies between different time series. By transforming the multivariate time series forecasting problem into a regression problem based on these lagged values, we can better capture the complex dynamics of the market.

2. Methodology

In this study, we delve into the intricacies of financial time series forecasting with a particular focus on the challenges and advancements in data collection and preprocessing techniques. Our approach is two-pronged, involving the collection of financial news data and the historical price data of gold, to facilitate a comprehensive analysis that incorporates both quantitative and qualitative factors [15–18].

$$\underbrace{\arg\min}_{W,H} \frac{1}{2} \left\| A - WH \right\|_{Fro}^{2} + \alpha\rho \left\| W \right\|_{1} + \alpha\rho \left\| H \right\|_{1} + \frac{\alpha(1-\rho)}{2} \left\| W \right\|_{Fro}^{2} + \frac{\alpha(1-\rho)}{2} \left\| H \right\|_{Fro}^{2}$$

$$SI_t = \sum_{i=1}^{t-1} e^{-\frac{t-i}{7}} SV_i + SV_t.$$

Price Data Acquisition

For our quantitative data, we gathered daily gold price information from the Federal Reserve Economic Data

(FRED) database, covering the same timeframe as our news data. This selection was strategic, as gold prices are influenced by a multitude of factors, including but not limited to, economic indicators, market sentiment, and geopolitical events (as shown in Figure 1).

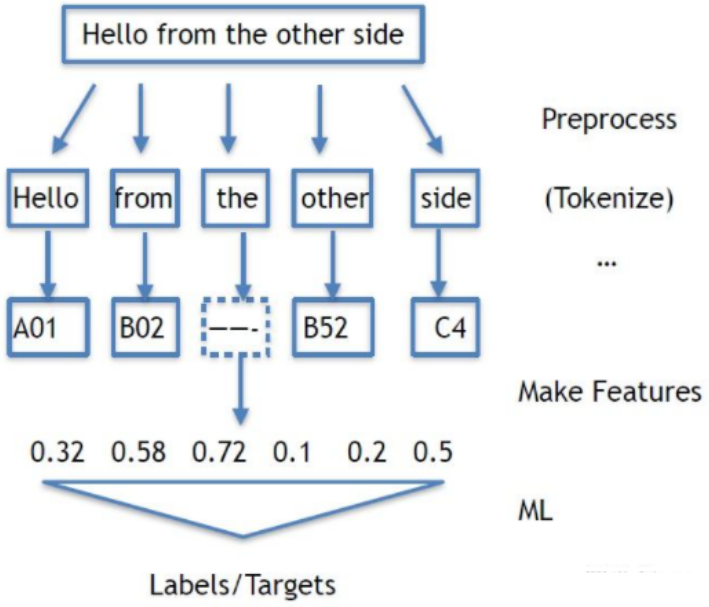

Figure 1. ML preprocess.

Random Forest (rf) is an ensemble learning method that operates by constructing multiple decision trees during training and outputting the average prediction of these trees. This method has been shown to improve the accuracy of predictions by reducing variance and avoiding overfitting (as shown in Figure 2).

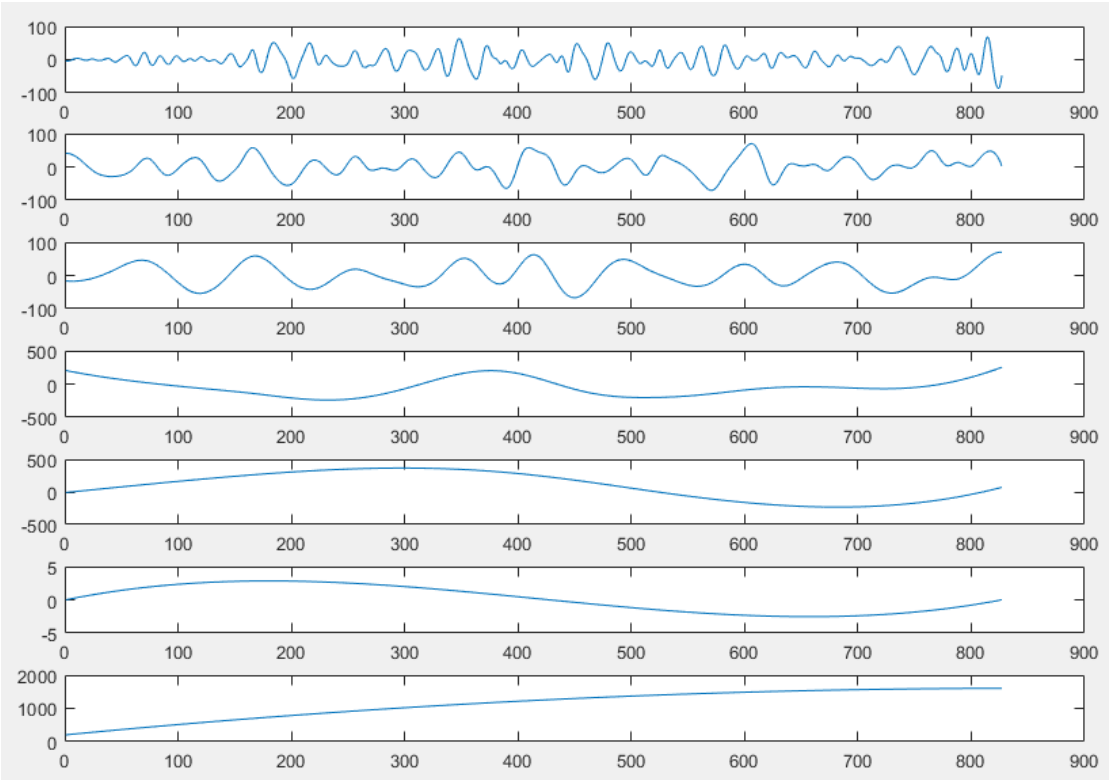

Figure 2. Comparison of Time Series Forecasting Models.

Support Vector Regression (svr) is a regression technique that relies on the concept of support vectors, which are data points that are closest to the decision boundary. SVR aims to find the optimal dividing hyperplane that maximizes the margin between the two classes while minimizing the prediction error in the training dataset. This method is particularly advantageous in high-dimensional spaces as its optimization does not depend on the dimensionality of the input space [19].

AutoRegressive Integrated Moving Average (arima) is a well-established linear model used for time series forecasting. It incorporates lagged values of the dependent variable and lagged forecast errors, which makes it suitable for capturing the linear interdependencies in time series data. The parameters (p, d, q) in the arima model represent the order of the autoregressive part, the degree of differencing needed for stationarity, and the order of the moving average part, respectively.

AutoRegressive Integrated Moving Average with eXogenous variables (arimax) extends the arima model by including exogenous variables, making it suitable for multivariate time series analysis. This model is particularly useful when there are external factors that can influence the time series being forecasted.

In our experiments, we implemented the arima model with parameters (p, d, q) = (4, 0, 3) for models without textual features. For models incorporating textual features, we used arimax with parameters (p, d, q) = (4, 1, 3). These parameter settings were chosen based on their ability to achieve a balance between model fit and predictive accuracy.

The models were evaluated using a dataset that included textual features derived from sentiment analysis using the Natural Language Toolkit (NLTK) library and our proposed short text theme and sentiment feature model. The results, as presented in Table 1, demonstrate that the inclusion of these textual features significantly improved the predictive performance of the models across all strata levels.

Table 1. Model Performance Evaluation.

Model		Features	*rmse*	*ae*	*mape*
SVR	Li	28	0.0646	0.0475	0.0885
	No text	2	0.1111	0.1001	0.1765
	NLTK	14	0.0586	0.0427	0.0799
	Our	17	**0.0584**	0.0428	0.0801
ARIMA	No text	–	0.0565	**0.0394**	0.0751
	NLTK	23	0.0568	0.0401	0.0757
	Our	30	0.0577	0.0412	0.0779
AdaBoost	No text	3	0.0594	0.0428	0.0799
	NLTK	7	0.0565	0.0397	0.0751
	Our	12	**0.0564**	**0.0396**	**0.0750**

3. Conclusion

This study has successfully demonstrated the efficacy of integrating textual analysis with traditional financial forecasting models in enhancing the accuracy of crude oil price predictions. By leveraging the thematic and sentiment information extracted from news headlines, our models were able to capture the dynamic interplay between market forces and global events, which are often reflected in the vast amount of textual data available.

The results of our analysis indicate that the inclusion of textual features significantly improved the prediction accuracy of machine learning models such as Random Forest Regression (RF), Support Vector Regression (SVR), and AdaBoost. These models, when augmented with textual data, outperformed the traditional ARIMA model, which performed well without textual features. This finding underscores the importance of incorporating external information sources, such as news sentiment, into financial forecasting models to enhance their predictive power.

Our novel approach combining Ensemble Empirical Mode Decomposition (EEMD) with Independent Component Analysis for analyzing non-linear and non-stationary time series data proved particularly effective in gold price analysis. The EEMD-BPNN-ADD model emerged as the most accurate for forecasting, providing interval predictions for gold prices and demonstrating the potential of this method for broader application in financial markets.

The study's findings have practical implications for investors and policymakers. By providing a more comprehensive and nuanced approach to market forecasting, our models can assist in strategic planning and risk management within the energy sector. The integration of textual analysis with financial models offers a valuable tool for navigating the complexities of global commodity markets.

In conclusion, this research contributes to the field by validating the benefits of combining textual data with traditional time series analysis in financial forecasting. It opens up new avenues for future research, encouraging the exploration of additional textual features and the application of advanced machine learning techniques to further refine market predictions. As the volume of textual data continues to grow, the integration of such data into financial models will become increasingly critical for staying ahead in the rapidly evolving landscape of global finance [20,21].

Funding

Not applicable.

Author Contributions

Conceptualization, J.R. and Q.Z.; writ-ing—original draft preparation and writing—review and editing, J.R., Q. Z., Z.K. S.L.and X.L. All of the authors read and agreed to the published the final manuscript.

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Not applicable.

Data Availability Statement

Not applicable.

Conflicts of Interest

The authors declare no conflict of interest.

Reference

1 Li S, Mo Y, Li Z. Automated Pneumonia Detection in Chest X-Ray Images Using Deep Learning Model. *Innovations in Applied Engineering and Technology* 2022; **1**: 1–6.

2 Wu Z, Wang Q, Gribok AV, Chen KP. Pipeline Degradation Evaluation Based on Distributed Fiber Sensors and Convolutional Neural Networks (CNNs). In Proceedings of the 27th International Conference on Optical Fiber Sensors, Alexandria, VA, USA, 29 August–2 September 2022. https://doi.org/10.1364/OFS.2022.W4.41.

3 Wang Q, Jian J, Wang M, Wu J, Mao Z-H, Gribok AV, Chen KP. Pipeline Defects Detection and Classification Based on Distributed Fiber Sensors and Neural Networks. In Proceedings of the Optical Fiber

Sensors Conference 2020 Special Edition, OSA Technical Digest, Washington, DC, USA, 8–12 June 2020. https://doi.org/10.1364/OFS.2020.W2B.3.

4 Peng Z, Jian J, Wang M, Wang Q, Boyer T, Wen H, Liu H, Mao Z-H, Chen KP. Big Data Analytics on Fiber-Optical Distributed Acoustic Sensing with Rayleigh Enhancements. In Proceedings of the 2019 IEEE Photonics Conference (IPC), San Antonio, TX, USA, 29 September–3 October 2019; pp. 1–3. https://doi.org/10.1109/IPCon.2019.8908496.

5 Wang Q, Zhao K, Badar M, Yi X, Lu P, Buric M, Mao Z-H, Chen KP. Improving OFDR Distributed Fiber Sensing by Fibers with Enhanced Rayleigh Backscattering and Image Processing. *IEEE Sensors Journal* 2022; **22**: 18471–18478. https://doi.org/10.1109/JSEN.2022.3197730.

6 Badar M, Lu P, Wang M, Wang Q, Chen KP, Buric M, Ohodnicki PR. Integrated Auxiliary Interferometer to Correct Non-Linear Tuning Errors in OFDR. In Proceedings of the SPIE, Optical Waveguide and Laser Sensors, 114050G, Online, 8 May 2020; Volume 11405. https://doi.org/10.1117/12.2558910.

7 Kumada H, Li Y, Yasuoka K, Naito F, Kurihara T, Sugimura T, Sakae T. Current Development Status of iBNCT001, Demonstrator of a LINAC-based Neutron Source for BNCT. *Journal of Neutron Research* 2022; **24**(**3–4**): 347–358. https://doi.org/10.3233/JNR-220029.

8 Chen M, Chen Y, Zhang Q. A Review of Energy Consumption in the Acquisition of Bio-Feedstock for Microalgae Biofuel Production. *Sustainability* 2021; **13**(**16**): 8873.

9 Li Y, Mizumoto M, Oshiro Y, Nitta H, Saito T, Iizumi T, Sakurai H. A Retrospective Study of Renal Growth Changes after Proton Beam Therapy for Pediatric Malignant Tumor. *Current Oncology* 2023; **30**: 1560–1570. https://doi.org/10.3390/curroncol30020120.

10 Li Y, Shimizu S, Mizumoto M, Iizumi T, Numajiri H, Makishima H, Sakurai H. Proton Beam Therapy for Multifocal Hepatocellular Carcinoma (HCC) Showing Complete Response in Pathological Anatomy After Liver Transplantation. *Cureus* 2022; **14**: e25744. https://doi.org/10.7759/cureus.25744.

11 Li Y, Matsumoto Y, Chen L, Sugawara Y, Oe E, Fujisawa N, Sakurai H. Smart Nanofiber Mesh with Locally Sustained Drug Release Enabled Synergistic Combination Therapy for Glioblastoma. *Nanomaterials* 2023; **13**: 414. https://doi.org/10.3390/nano13030414.

12 Chen M. Investigating the Influence of Interannual Precipitation Variability on Terrestrial Ecosystem Productivity. Doctoral Dissertation, Massachusetts Institute of Technology, Cambridge, MA, USA, 2023.

13 Chen M. Annual Precipitation Forecast of Guangzhou Based on Genetic Algorithm and Backpropagation Neural Network (GA-BP). In Proceedings of the International Conference on Algorithms, High Performance Computing, and Artificial Intelligence (AHPCAI 2021), 19 – 21 November 2021; Volume 12156, pp. 182–186.

14 Dong S, Xu T, Chen M. Solar Radiation Characteristics in Shanghai. In *Journal of Physics: Conference Series*; IOP Publishing: Bristol, UK, 2022; Volume 2351, p. 012016.

15 Wang R, Shapiro V. Topological Semantics for Lumped Parameter Systems Modeling. *Advanced Engineering Informatics* 2019; **42**: 100958.

16 Wang R, Behandish M. Surrogate Modeling for Physical Systems with Preserved Properties and Adjustable Tradeoffs. *arXiv* 2022, arXiv:2202.01139.

17 Wang J, Tong J, Tan K, Vorobeychik Y, Kantaros Y. Conformal Temporal Logic Planning Using Large Language Models: Knowing When to Do What and When to Ask for Help. *arXiv* 2023, arXiv:2309.10092.

18 Shimizu S, Nakai K, Li Y, Mizumoto M, Kumada H, Ishikawa E, Sakurai H. Boron Neutron Capture Therapy for Recurrent Glioblastoma Multiforme: Imaging Evaluation of a Case with Long-Term Local Control and Survival. *Cureus* 2023; **15**: e33898. https://doi.org/10.7759/cureus.33898.

19 Shimizu S, Mizumoto M, Okumura T, Li Y, Baba K, Murakami M, Sakurai H. Proton Beam Therapy for a Giant Hepatic Hemangioma: A Case Report and Literature Review. *Clinical and Translational Radiation Oncology* 2021; **27**: 152–156. https://doi.org/10.1016/j.ctro.2021.01.014.

20 Li S, Mo Y, Li Z. Automated Pneumonia Detection in Chest X-Ray Images Using Deep Learning Model. *Innovations in Applied Engineering and Technology* 2022; **1**(**1**): 1–6. https://doi.org/10.62836/iaet.vli1.002.

21 Li Z, *et al*. Stock Market Analysis and Prediction Using LSTM: A Case Study on Technology Stocks. *Innovations in Applied Engineering and Technology* 2023; **2**(**1**): 1–6. https://doi.org/10.62836/iaet.v2i1.162.