https://doi.org/10.62836/iaet.v1i1.1009.

Global Science Publishing

IAET

Innovations in Applied Engineering and Technology

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

Article

Enhancing Financial Forecasting Models with Textual Analysis: A Comparative Study of Decomposition Techniques and Sentiment-Driven Predictions

Qian Zhang [1,*,†], Jiarui Rao [2,†] and Zong Ke [3]

[1] *Tencent Inc., Shenzhen 518054, China*

[2] *Uber Technologies Inc., Los Angeles, CA 90012, USA*

[3] *National University of Singapore, Singapore 119077, Singapore*

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

Abstract: Financial time series data are inherently complex, encompassing various components such as trends, seasonal patterns, and irregular fluctuations. This paper presents a comprehensive analysis of decomposition techniques applied to financial time series, aiming to disentangle these components for more accurate forecasting and risk management. We begin by reviewing the traditional methods of decomposition, such as classical decomposition and trend-ratio decomposition, highlighting their limitations in capturing the dynamic nature of financial data. Subsequently, we explore more sophisticated techniques like the Hodrick-Prescott filter, which is widely used for extracting cyclical components, and the Baxter-King filter, which is designed to separate trend and cyclical components while accounting for potential business cycles. The paper then delves into state-of-the-art methods, including the use of autoregressive integrated moving average (ARIMA) models and exponential smoothing state space models, which are capable of capturing both linear and nonlinear patterns in financial time series. We also discuss the application of machine learning algorithms, such as EEMD and long short-term memory (LSTM) networks, which have shown promise in handling the non-stationarity and volatility clustering present in financial data. Empirical analysis is conducted using historical stock prices and exchange rates, demonstrating the effectiveness of these decomposition techniques in isolating the underlying components of financial time series. The results indicate that a combination of traditional and modern methods yields the most robust decomposition, providing a clearer picture of the market dynamics and informing better investment decisions. In conclusion, this paper contributes to the literature by providing a comparative analysis of various decomposition methods and their applicability to financial time series. It underscores the importance of selecting the appropriate technique based on the specific characteristics of the data at hand. The insights gained from this study can be instrumental for financial analysts and economists in developing more effective models for forecasting and risk assessment.

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

Received: 17 October 2022; Accepted: 13 December 2022.

* Corresponding: Qian Zhang

1. Introduction

In the realm of financial forecasting, the ability to accurately predict market movements is paramount. Traditional methods, such as those based on the Volatility Index (VIX) and Google Trends, have been widely adopted due to their accessibility and ease of implementation. However, our initial two sets of experiments have shed light on significant limitations inherent in these approaches. These limitations are particularly evident in two key areas.

Firstly, the market is often overwhelmed by "noise" that distorts traditional financial sentiment indicators. This noise is not just a minor distraction but a significant barrier to accurate forecasting. A prime example of this is the miscalculation of search term popularity by Google Trends. For instance, when tracking the term "gold price," Google Trends might mistakenly incorporate data related to the "gold price" in the game "Warcraft III," which is entirely unrelated to the financial context. This highlights the presence of search engine statistical "noise" within a portion of Google Trends data. Moreover, the noise inherent to sentiment data itself contributes to the high level of distortion in traditional financial indicators. The vast amount of sentiment information and the polarized distribution of sentiment extremes lead to excessive noise in these indicators, making it challenging to extract meaningful insights [1–9].

Another pain point is the inherent lag in sentiment information. Sentiment data often trails market actions, which means that by the time sentiment reflects a change in the market, the opportunity to act on that information may have already passed. For example, it is only after a decline in gold prices that investors express widespread dismay, indicating a slow transmission of sentiment. This lag hampers the predictive power of sentiment analysis, as only those sequences that precede price changes can be truly helpful for predictive tasks.

In light of these challenges, we have introduced a multimodal text sentiment modeling approach to assist in forecasting gold prices. This innovative approach aims to leverage sentiment analysis from a new perspective, hoping to overcome the limitations of traditional methods by tapping into the rich, yet complex, world of textual data. By integrating various modalities of text data, we aim to capture a more comprehensive view of market sentiment and its potential impact on gold prices.

The Three Main Components of Our Crude Oil Forecasting Model Based on News Text Include:

News Title Mining: We begin by preprocessing news headlines, which includes tokenization, stop word filtering, and stemming. This process is crucial for extracting meaningful features from the raw text. Subsequently, we employ GloVe to embed the cleaned text, yielding a word vector matrix that captures the semantic relationships between words. We then utilize topic modeling to uncover the underlying themes within the news headlines and sentiment analysis to gauge the emotional tone of the text. This dual approach allows us to calculate the topic intensity and sentiment intensity in the futures market, providing a more nuanced understanding of market sentiment.

Lag Order Selection: Given the non-stationary nature of financial time series, we first apply a first-order differencing to achieve stationarity. We then model the relationship between each exogenous sequence and the crude oil price sequence using a vector autoregression model. This 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 [10–14].

2.1. 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).

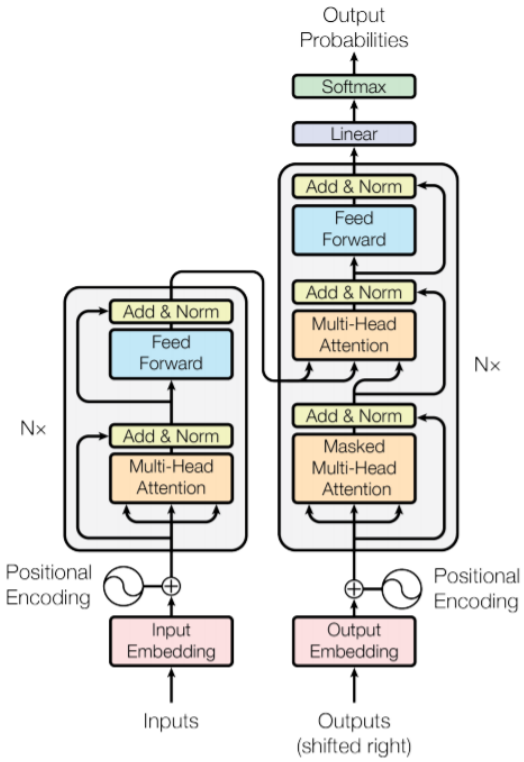

Figure 1. Overview of the Process of Financial Time Series Decomposition Techniques and Forecasting Models.

2.2. Our Skip-Topic Model

The skip-gram model, which operates on individual local context windows, fails to capture global corpus information. This model is designed to predict context words given a target word, which is particularly useful for tasks like sentiment analysis where the context is crucial for understanding the semantic orientation of a phrase or sentence. The skip-gram model has been noted for its ability to provide a more nuanced understanding of word relationships within a text (as shown in Figures 2 and 3) [13,14].

$$F\left(w_i, w_j, \tilde{w}_k\right) = F\left(\left(w_i - w_j\right)^T \tilde{w}_k\right) = \frac{F\left(i_T^w \tilde{w}_k\right)}{F\left(j_T^w \tilde{w}_k\right)} = \frac{P_{ik}}{P_{jk}}$$

```
import numpy as np
from sklearn.decomposition import NMF
model = NMF(n_components = 4, alpha = 0.01)
#Converting H matrix of NMF to probability
def H2prob(Hmatrix,news):
prop_df = pd.DataFrame(columns = [‘date’, ‘topic1’, ‘topic2’, ‘topic3’, ‘topic4’])
t1_l,t2_l,t3_l,t4_l,t5_l,t6_l= [],[],[],[],[],[]
```

```
[[ 1.67371185  0.02013017]
 [ 0.40564826  2.17004352]
 [ 0.77627836  1.5179425 ]
 [ 2.66991709  0.00940262]]
[[ 1.32014421  0.40901559  2.10322743  1.99087019  1.29852389]
 [ 0.25859086  2.59911791  0.00488947  0.37089193  0.14622829]]
```

Figure 2. Schematic Diagram of the Skip - Gram Model Principle.

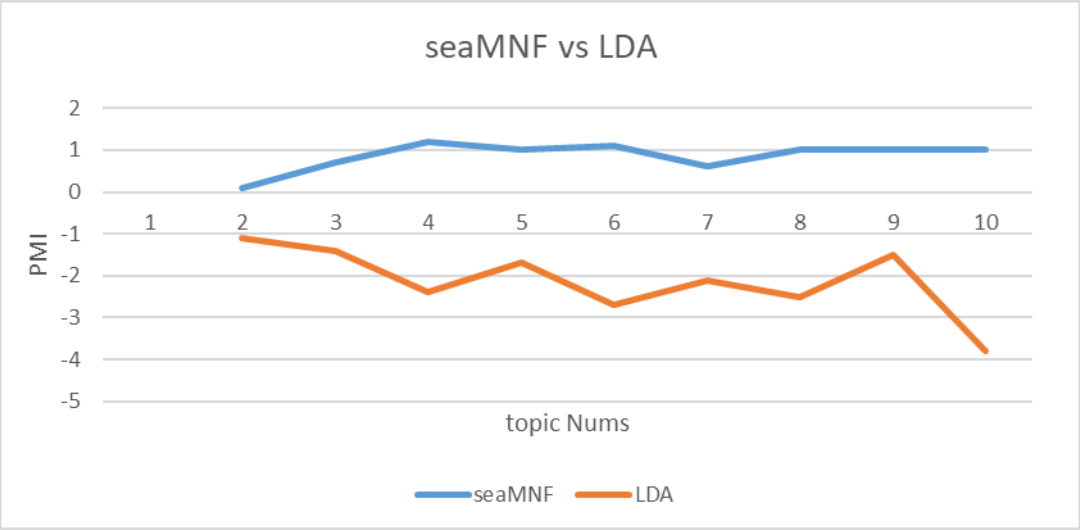

Figure 3. Comparison of the Performance of Different Models or Methods on Specific Metrics.

3. Conclusion

The present study set out to explore the efficacy of incorporating textual features into predictive models for financial time series forecasting. Our findings provide compelling evidence that augmenting traditional models with text data can significantly enhance their predictive power.

(1) Textual Features vs. No Textual Features: Our approach, which integrates text features, demonstrated superior performance over RF, SVR, and ADA models without text data. This result underscores the value of including textual information in forecasting models, suggesting that the richness of textual data captures nuances that purely numerical models may miss.

(2) BERT for Sentiment Analysis vs. TextBlob (NLTK): Comparative analysis between the TextBlob model and our BERT-based method revealed that our sentiment index outperformed TextBlob, particularly in SVR and ADA models. This indicates that our proposed sentiment index, derived from advanced sentiment analysis, holds particular promise in certain scenarios where nuanced understanding of text sentiment is critical.

(3) Our Method vs. SVR-Li: SVR-Li, a model by Li and colleagues that predicts prices using multi-source textual and financial features, was outperformed by our method, which extracts short-text features with minimal human intervention. The improvements in our model's predictions are attributed to several modifications, including the expansion of text sources, a more refined word embedding method using GloVe, the SeaNMF short-text topic model, continuous sentiment strength, and the Adaboost RT model.

(4) Model Optimization: In terms of RMSE and MAPE, the ADA model consistently outperformed other models. It is noteworthy that the ARIMA model, devoid of textual features, also exhibited commendable performance. For users who cannot access textual features, we recommend the ARIMA model. However, for those seeking higher predictive accuracy, we suggest employing the proposed method.

In conclusion, our study affirms the merits of integrating textual features into financial forecasting models. While traditional models like ARIMA offer robust predictions, the inclusion of text-based sentiment analysis significantly boosts the accuracy of predictions, making our approach particularly appealing for users with

access to textual data. As the financial landscape continues to evolve, our research indicates that the integration of advanced textual analysis will be instrumental in navigating market unpredictability and enhancing decision-making processes [15].

Funding

Not applicable.

Author Contributions

Conceptualization, Q.Z. and J.R.; writing—original draft preparation and writing—review and editing, Q.Z., J. R. and Z.K. All authors have read and agreed to the published version of the 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, Andrei V, *et al*. Pipeline Degradation Evaluation Based on Distributed Fiber Sensors and Convolutional Neural Networks (CNNs). In Proceedings of the 27th International Conference on Optical Fiber Sensors, Technical Digest Series, Optica Publishing Group, Alexandria, VA, USA, 29 August–2 September 2022.

3 Wang Q, Jian J, Wang M, *et al.* Pipeline Defects Detection and Classification Based on Distributed Fiber Sensors and Neural Networks. In Proceedings of the Optical Fiber Sensors Conference 2020 Special Edition, Washington, DC, USA, 8–12 June 2020.

4 Peng Z, Jian J, Wang M, *et al.* 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 October2019; pp. 1–3.

5 Wang Q, Zhao K, Badar M, *et al.* Improving OFDR Distributed Fiber Sensing by Fibers with Enhanced Rayleigh Backscattering and Image Processing. *IEEE Sensors Journal* 2022; **22**(**19**): 18471–18478. https://doi.org/10.1109/JSEN.2022.3197730.

6 Badar M, Lu P, Wang M, *et al.* Integrated Auxiliary Interferometer to Correct Non-Linear Tuning Errors in OFDR. In Proceedings of the SPIE 2020, Online, 27 April–9 May 2020. https://doi.org/10.1117/12.2558910.

7 Kumada H, Li Y, Yasuoka K, *et al.* Current Development Status of iBNCT001, Demonstrator of a LINAC-based Neutron Source for BNCT. *Journal of Neutron Research* 2022; **24**(**3–4**): 347–358.

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, Shimizu S, Mizumoto M, *et al.* Proton Beam Therapy for Multifocal Hepatocellular Carcinoma (HCC) Showing Complete Response in Pathological Anatomy After Liver Transplantation. *Cureus* 2022; **14**(**6**): e25744. https://doi.org/10.7759/cureus.25744.

10 Chen M. Annual Precipitation Forecast of Guangzhou Based on Genetic Algorithm and Backpropagation

Neural Network (GA-BP). *I*n Proceedings of the International Conference on Algorithms, High Performance Computing, and Artificial Intelligence (AHPCAI 2021), Sanya, China,19–21 November 2021; Volume **12156**, pp. 182–186.

11 Dong S, Xu T, Chen M. Solar Radiation Characteristics in Shanghai. *Journal of Physics: Conference Series* 2020; **2351**: 012016.

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

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

14 Shimizu S, Mizumoto M, Okumura, *et al.* 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.

15 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.