Investigating Correlations in the Stock Market Using Multi-Source Data and Tensor-Based Models

Word Count: **[2605]** | Submitted on: **23-02-2025**

Björn van Engelenburg
bjornvane@hotmail.com
University of Amsterdam
Amsterdam, The Netherlands

1 INTRODUCTION

Conventional stock market prediction methods typically rely on historical price data to forecast future trends[21]. The reason tensor-based computational frameworks are used is because of the combination of quantitative historical pricing data, news events and user sentiment data [21]. Integrating these heterogeneous information sources can be seen as extremely difficult, since when for example a linear model is used, the assumption is that the data is independent of each other. The correlations and interactions between different features are highly common between these three possibly dependent information sources.

Currently, literature on using tensor-based models in this domain is scarce and mainly used in fields such as neuro-imaging, [7, 8, 14], elucidating great opportunity in the current domain. Other research cases include LSTM-tensor based models or do not take multi-correlated financial instruments into account [9–11, 17]. Other similar research approaches use PCA in their tensor-models or they use multiplex networks to capture dynamic financial interactions in stock markets, focusing on clustering coefficients [3, 19].

As from current literature [21], While a tensor-based computational framework has been created to model the combined effects of various information sources for stock prediction, each stock prediction within that framework is treated as a separate task and learned independently, regardless of the correlations among the stocks[21].

Additionally, more wide-scale frameworks have been used, such as a Bayesian autoregressive tensor model for dynamic analysis of multilayer networks with a focus on shock propagation[14]. Covariance and correlation play important roles in portfolio management of risk as well as portfolio optimization [2, 13].

Consequently, one financial instrument could follow 10 others[13]. This paper will focus on the multi-correlated financial instruments and domain of the financial stock market. We see a gap in the current literature regarding combination of tensor-based models with multi-correlated financial instruments.

To what extent can multi-correlated financial instruments involving news data and tensor ML models enhance stock price prediction? These below questions will be answered to shine more light on and answer the main research question effectively:

- How do tensor-based ML models impact stock price prediction?
- To what extent are multi-correlated financial instruments associated with stock price predictions?
- Can we evaluate the tensor-based model including the multi-correlated instruments?

1.1 Motivation

Finding adequate correlations between financial instruments helps diversify the portfolio as well as aid in in conventional stock market prediction methods [2, 13, 21].

Concluding from literature, the reason tensors could be so telling-giving is because they allow for a 3-D interpretation of associations within the financial domain. In traditional regression (like linear regression), the inputs (predictors) and outputs (responses) are typically vectors or matrices (2D arrays). A 3D tensor could represent a sequence of images (height × width × time) or a dataset where each data point has multiple characteristics organized across multiple dimensions[14].

A common application of tensor regression is Bayesian Dynamic Tensor Regression, where the response (output) and predictors (inputs) are tensors, and the model attempts to capture the dependencies between them. Tensor regression is particularly useful when the data has inherent multi-dimensional structure, and flattening the data into vectors would lead to loss of information about those relationships[14].

Personally, the way certain stocks or financial instruments in this paper are correlated with each other, intrigues us. That is why this paper is focused so heavily on correlations and tensors are included, since they capture some inherent correlations otherwise not visible to the public.

2 RELATED WORK

Through the Data Science / AI department's thesis design set-up, here is where the objective of the research will be stated.

The research gap being filled here is the research on multi-correlated financial instruments using multi-source data with possibly tensor-based models. How will this impact stock market prediction?

2.1 Models

Reaching the objective of answering the research question, also through usage of the sub-research questions: "To what extent can multi-correlated financial instruments involving news data and tensor ML models enhance stock price prediction?" This research paper offers a deep dive into the correlation, covariance and other correlation metrics to better explain the work of tensor models and its effects on high- and multi-dimensional data. Most notably, tensor-regression models are particularly useful for when the data has inherent multi-dimensional structure, and flattening the data into vectors would lead to loss of information about those relationships[9].

Additionally, there is the e-LSTM model used for stock price prediction using a news dataset. Traditional LSTM models struggle to predict stock movements based on media sentiment data due to irregular time intervals, potentially forgetting past influences over long gaps. To address this, an event-based mechanism is introduced to enhance LSTM's ability to process heterogeneous data with uneven temporal distributions [18].

The work will be evaluated through usefulness of the correlations, particularly the performance of the tensor-based regression model and the correlations to be found within the financial instruments, associations thereof. Additionally, a baseline model will be created to compare against. In terms of time spent, a big part will be coding the tensor-based regression model, finding the multi-correlated financial instruments and depicting its usefulness for stock market prediction and portfolio optimization, comparing it to other works [9]and gathering our own event / news based data as for heterogenuous information fusion.

2.2 Sentiment and NLP

It has become increasingly apparent that weblogs, news or social media data has an influence on predicting stock prices and their movement [6, 17, 21]. We list four main ways to extract sentiment out of all the internet data available[12]:

- **VADER**: A rule-based general sentiment analysis method that uses a combination of qualitative and quantitative methods to empirically validate a list of lexical features. It calculates sentiment scores as *negative, positive, neutral*, and *compound*.
- **TextBlob**: Predefined in NLP, this method provides sentiment scores for polarity and subjectivity. *Polarity* classifies a statement as positive or negative, assigning a score in the range of $[-1, 1]$. *Subjectivity* measures whether the text expresses an opinion, emotion, or factual information.
- **FLAIR**: An NLP framework that facilitates sequence labeling and text classification. Its main function is to provide a unified interface for different types of text analysis and embedding in documents[12].
- **BERT**: Besides these three methods, there is also a deep learning method called BERT [16]. This model undergoes pre-training on a vast collection of general-domain texts using a self-supervised learning approach. It is made by Google and is otherwise known as Bi-Directional Encoder Representations from Transformers. It is essentially designed to understand the context of words more effectively than traditional NLP methods[15].

3 METHODOLOGY

The objective of this study is to analyze the impact of financial news sentiments on stock price movements and integrate this information with historical stock data to improve predictive accuracy. Traditional stock market prediction models primarily rely on historical pricing data, often overlooking the influence of external factors such as financial news and public sentiment. By incorporating multiple sources of information, this study aims to explore how financial news sentiment contributes to market trends and enhances predictive performance. Additionally, correlations between multiple stocks will be included to see the impact of the correlations between stocks on prediction accuracy.

To achieve this, we employ the BERT sentiment scoring method to quantify the sentiment of financial news articles. Bert is chosen as a method because of its ability to extract context from words and because it is grounded in literature [15, 16]. These sentiment scores are then combined with historical stock price data to train a machine learning model. The study utilizes an tensor-based ML model to evaluate how well sentiment data, in isolation and in combination with other sentiment scores, impacts stock price prediction.

The predictive model operates as follows:

Trend Prediction: The model forecasts stock movement on the next trading day using the previous day's closing price and sentiment scores. The next day's price is compared with the previous day to forecast the continuous stock price values. Future Trend Prediction: The model extends its prediction horizon by comparing stock prices over multiple days (e.g., one-week forecasts). The dataset consists of historical stock prices from Yahoo Finance and financial news articles scraped from major financial media sources. The combination of sentiment scores, historical stock data, and the labeling mechanism forms the basis for our analysis of stock price movements across multiple financial sectors.

By integrating multi-source financial data into a tensor-based computational framework, this study aims to assess the extent to which financial news sentiment can enhance stock price prediction. The next sections outline the data collection process, data processing, and model evaluation metrics.

3.1 Datasets

The respective and available datasets are:

(1) Historical Pricing data. This dataset originates from the Yahoo finance package [1].
(2) Event-based / 'news' data. This can be scraped from Reddit or using Gnews, which is a package to acquire news from Google News. It provides an efficient and structured way to access news data without needing direct web scraping.
(3) Richer financial data, for example fundamental data, which can be obtained using financial toolkit website [5].

There are multiple datasets as requested by the data science / AI institute. Then there is also the social / event based information dataset listed above to allow for heterogeneous information fusion. More information to be scraped or gathered if deemed appropriate. Scraping techniques include using Selenium to scrape these events from the Web.

3.2 Data Collection

The historical stock price data used in this study is sourced from Yahoo Finance, a widely used financial data repository frequently utilized in academic research. The dataset contains stock market data for the required time periods and is acquired from the Yahoo Finance package. Each record includes key stock market indicators:

- **Open:** The price at which the stock opened trading on a given day.
- **High:** The highest price reached during that trading session.
- **Low:** The lowest recorded price of the stock on that day.
- **Close:** The final trading price of the stock at market close.

- **Volume:** The total number of shares traded during the day.
- **Adjusted Close:** The closing price after adjusting for corporate actions such as stock splits or dividends.

This structured historical dataset provides a quantitative foundation for stock market analysis and is used in conjunction with financial news sentiment scores to evaluate their impact on stock price movements. The dataset spans multiple companies across various financial sectors to ensure a broad and representative analysis.

3.3 Data Processing

As for processing of the datasets, the Yahoo Finance data and the Gnews datasets are clean. Possibly for the Gnews dataset it could be that some data might be missing. If Reddit is chosen, it could be that it is important to find news associated with the stocks, as the news originating from Reddit might be more unstructured. There are multiple subreddits / webpages containing the news, so it is important to remove duplicate newsposts and make design decisions on what subreddits / webpages to include in the data.

It is important for the event-based / news data to be processed with NLP techniques. Bert will be used here and for this task it is important to check and improve the model so accurate sentiment scores will come out. Lastly, for the richer financial data, the API will have to be acquired to check the validity of the data.

3.4 Methodology

The objective will be to analyse multi-correlated financial instruments and create a tensor-based model, presumably the Bayesian Dynamic Tensor Regression[14]. Assessing valuable correlations and associations thereafter will be important. The correlation formulas can be found below:

Covariance formula:

$$\mathrm{Cov}(X, Y) = \frac{1}{n} \sum_{i=1}^{n} (X_i - \mu_X)(Y_i - \mu_Y)$$

Correlation formula:

$$\mathrm{Corr}(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\sigma_X \sigma_Y}$$

Pearson correlation coefficient:

$$r_{xy} = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n} (X_i - \bar{X})^2} \sqrt{\sum_{i=1}^{n} (Y_i - \bar{Y})^2}}$$

3.5 Modeling

As for a baseline model, this study will be implementing the linear regression model first. This also allows for efficient answering of the research question: *How do tensor-based models impact stock price prediction?* Afterwards, the Bayesian Dynamic Tensor regression will be used [14].

3.6 Evaluation

The main way to evaluate is to compare it to the baseline model. Referring to the first sub-question, this will be answered through comparison with the baseline model and even with models in literature. Relating to the second sub-question, this will be answered through assessing the correlations between the financial instruments and its impact in the prediction. If the modincorporating multi-correlated financial instruments outperform simple models, then these instruments contribute significantly to stock price predictions. As for the third sub-question, can we include the correlations between the stocks in the tensor-based ML model?

This question in turn helps answer the main research question, which is: *To what extent can multi-correlated financial instruments involving news data and tensor ML models enhance stock price prediction?*. The first sub-question tackles the impact of tensor-based ML models. The second sub-question assesses the correlations between the stocks.

The third sub-question assesses the inclusion of the correlations between the stocks into a tensor-based ML model. Answering the main research question, is when we include the news aspect into it.

Metrics that will be used to evaluate are common metrics such as RMSE, MAE, MAPE and R2, as seen in literature [4, 20].

4 RISK ASSESSMENT

The primary challenge in this research lies in the selection and evaluation of the model. Since three different model architectures will be considered—LSTM, e-LSTM, and Bayesian Dynamic Tensor Regression—the choice of the optimal model is crucial for achieving accurate and reliable stock trend predictions. Each of these models has distinct advantages and computational requirements, and their effectiveness will depend on the specific structure and quality of the input data.

Another significant risk concerns the availability and suitability of the datasets. While historical stock price data is accessible via Yahoo Finance, an additional dataset is required to evaluate the model's predictive performance in a real-world scenario. This dataset will incorporate financial news sentiment scores derived from different sentiment analysis techniques (VADER, TextBlob, and Flair). The challenge is that financial news datasets often require extensive preprocessing to align with structured historical stock data. Additionally, integrating and aligning these heterogeneous data sources may introduce inconsistencies that could affect the final predictions.

Furthermore, computational limitations could pose a constraint, particularly when training tensor-based models. Bayesian Dynamic Tensor Regression, for instance, requires extensive matrix operations that may be computationally expensive, especially when handling high-dimensional data with multiple correlated financial instruments. Efficient resource allocation and model optimization strategies will be essential to ensure feasibility within the project's timeline if needed.

To mitigate these risks, a stepwise approach will be followed, beginning with the training and evaluation of simpler models (such as LSTM and e-LSTM) before progressing to more complex tensor-based models. Before starting with these models, a baseline model will be used even. Data preprocessing will be conducted in multiple stages to ensure alignment between financial news sentiment scores and stock price movements. Additionally, computational resources will be monitored throughout the training process, with

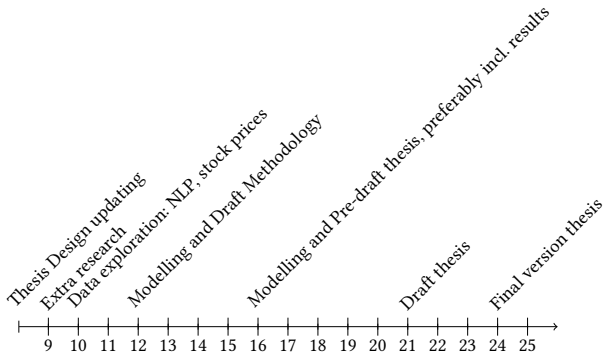
optimizations implemented where necessary. If the selected dataset proves insufficient, the alternative data sources to enhance model performance and evaluation will be explored.

In terms of the data, the only risks are with financial toolkit for the fundamental data, as it is not tested yet. The Yahoo Finance package as well as the Gnews package seem pretty feasible. They are well-established and have been around for long, especially Yahoo Finance. As a back-up, there is the Reddit API which is acquired and enables social media news evaluation.

As for the modelling risks, there is a chance that the bayesian dynamic tensor regression cannot be implemented correctly. In that case, the models e-LSTM or LSTM will be touched upon.

5 PROJECT PLAN

I am outlining my project timeline to provide a clear and realistic estimate of the time required for each phase of the research. This timeline reflects my structured approach to managing tasks efficiently and ensures that the work remains on track. It also allows the UvA supervisor to see the progress and verify that I am meeting key milestones as planned. To illustrate this, I will present my schedule using a visual representation in the form of a timeline.



REFERENCES

[1] Ran Aroussi. 2023. yfinance: Download market data from Yahoo! Finance's API. https://pypi.org/project/yfinance/ Accessed: 2024-09-29.

[2] Arindam Bandyopadhyay. 2022. 141Correlation Theorem and Portfolio Management Techniques. In *Basic Statistics for Risk Management in Banks and Financial Institutions*. Oxford University Press. https://doi.org/10.1093/oso/9780192849014.003.0006 arXiv:https://academic.oup.com/book/0/chapter/361609472/chapter-pdf/48709924/oso-9780192849014-chapter-6.pdf

[3] Paolo Bartesaghi, Gian Paolo Clemente, and Rosanna Grassi. 2022. A tensor-based unified approach for clustering coefficients in financial multiplex networks. *Information Sciences* 601 (2022), 268–286. https://doi.org/10.1016/j.ins.2022.02.084

[4] Narayana Darapaneni, Anwesh Reddy Paduri, Himank Sharma, Milind Manjrekar, Nutan Hindlekar, Pranali Bhagat, Usha Aiyer, and Yogesh Agarwal. 2022. Stock Price Prediction using Sentiment Analysis and Deep Learning for Indian Markets. arXiv:2204.05783 [q-fin.ST] https://arxiv.org/abs/2204.05783

[5] Financial Modeling Prep. 2025. Pricing Plans. https://site.financialmodelingprep.com/pricing-plans?couponCode=jeroen Accessed: 2025-02-23.

[6] Eric Gilbert and Karrie Karahalios. 2010. Widespread Worry and the Stock Market. *Proceedings of the International AAAI Conference on Web and Social Media* 4, 1 (May 2010), 58–65. https://doi.org/10.1609/icwsm.v4i1.14023

[7] R. Guerrero, R. Wolz, A.W. Rao, and D. Rueckert. 2014. Manifold population modeling as a neuro-imaging biomarker: Application to ADNI and ADNI-GO. *NeuroImage* 94 (2014), 275–286. https://doi.org/10.1016/j.neuroimage.2014.03.036

[8] Lexin Li Hua Zhou and Hongtu Zhu. 2013. Tensor Regression with Applications in Neuroimaging Data Analysis. *J. Amer. Statist. Assoc.* 108, 502 (2013), 540–552. https://doi.org/10.1080/01621459.2013.776499 arXiv:https://doi.org/10.1080/01621459.2013.776499 PMID: 24791032.

[9] Qing Li, Yuanzhu Chen, Li Ling Jiang, Ping Li, and Hsinchun Chen. 2016. A Tensor-Based Information Framework for Predicting the Stock Market. *ACM Trans. Inf. Syst.* 34, 2, Article 11 (Feb. 2016), 30 pages. https://doi.org/10.1145/2838731

[10] Qing Li, LiLing Jiang, Ping Li, and Hsinchun Chen. 2015. Tensor-Based Learning for Predicting Stock Movements. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*. 1784–1790. https://aaai.org/papers/9452-tensor-based-learning-for-predicting-stock-movements/

[11] Qing Li, Jinghua Tan, Jun Wang, and Hsinchun Chen. 2021. A Multimodal Event-Driven LSTM Model for Stock Prediction Using Online News. *IEEE Transactions on Knowledge and Data Engineering* 33, 10 (2021), 3323–3337. https://doi.org/10.1109/TKDE.2020.2968894

[12] Junaid Maqbool, Preeti Aggarwal, Ravreet Kaur, Ajay Mittal, and Ishfaq Ganaie. 2023. Stock Prediction by Integrating Sentiment Scores of Financial News and MLP-Regressor: A Machine Learning Approach. *Procedia Computer Science* 218 (01 2023), 1067–1078. https://doi.org/10.1016/j.procs.2023.01.086

[13] H. Markowitz. 1959. *Portfolio Selection: Efficient Diversification of Investments*. Wiley. https://books.google.nl/books?id=UEIUAQAAMAAJ

[14] Matteo Iacopini Monica Billio, Roberto Casarin and Sylvia Kaufmann. 2023. Bayesian Dynamic Tensor Regression. *Journal of Business & Economic Statistics* 41, 2 (2023), 429–439. https://doi.org/10.1080/07350015.2022.2032721 arXiv:https://doi.org/10.1080/07350015.2022.2032721

[15] Priyank Sonkiya, Vikas Bajpai, and Anukriti Bansal. 2021. Stock price prediction using BERT and GAN. arXiv:2107.09055 [q-fin.ST] https://arxiv.org/abs/2107.09055

[16] Matheus Gomes Sousa, Kenzo Sakiyama, Lucas de Souza Rodrigues, Pedro Henrique Moraes, Eraldo Rezende Fernandes, and Edson Takashi Matsubara. 2019. BERT for Stock Market Sentiment Analysis. In *2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI)*. 1597–1601. https://doi.org/10.1109/ICTAI.2019.00231

[17] Jinghua Tan, Jun Wang, Denisa Rinprasertmeechai, Rong Xing, and Qing Li. 2019. A Tensor-based eLSTM Model to Predict Stock Price Using Financial News. In *Proceedings of the 52nd Hawaii International Conference on System Sciences*. 10. http://hdl.handle.net/10125/59606

[18] Jinghua Tan, Jun Wang, Denisa Rinprasertmeechai, Rong Xing, and Qing Li. 2019. A Tensor-based eLSTM Model to Predict Stock Price Using Financial News. https://doi.org/10.24251/HICSS.2019.201

[19] Jun Wang, Yexun Hu, Tai-Xiang Jiang, Jinghua Tan, and Qing Li. 2023. Essential tensor learning for multimodal information-driven stock movement prediction. *Knowledge-Based Systems* 262 (2023), 110262. https://doi.org/10.1016/j.knosys.2023.110262

[20] Anuradha Yenkikar and C. Narendra Babu. 2023. Comparison of Machine Learning Algorithm for Stock Price Prediction Using Sentiment Analysis. In *2023 International Conference on Emerging Smart Computing and Informatics (ESCI)*. IEEE. https://doi.org/10.1109/ESCI56872.2023.10099875

[21] Xi Zhang, Yunjia Zhang, Senzhang Wang, Yuntao Yao, Binxing Fang, and Philip S. Yu. 2018. Improving stock market prediction via heterogeneous information fusion. *Knowledge-Based Systems* 143 (2018), 236–247. https://doi.org/10.1016/j.knosys.2017.12F.025