Text representation for direction prediction of share market - Reproducibility

Ashutosh Anand, Kunal Anand and Anjali Singh

Abstract

This study examines how economic news impacts stock market closing price movements, highlighting the relationship between news events and market trends. It evaluates statistical and deep learning models using text and price data, introducing a parallel CNN model that combines both for higher prediction accuracy. Results show strong correlations between predictions and actual trends, demonstrating the power of leveraging news content to forecast market movements.

1. Introduction

Financial news articles provide insights into events like economic policies, international markets, and commodity trading, which directly influence stock market movements. Articles published during market hours are assumed to impact market trends in real-time. Beyond direct effects, implicit relationships between events in these articles also shape market behavior.

This study explores how financial news articles impact the movement of stock market closing prices. It evaluates the predictive power of news text compared to historical prices and investigates the benefits of combining both. By framing the task as a classification problem, the study focuses on short-term predictions using articles published during market hours.

Unlike prior research that overlooks the context and connections within news text, this work leverages advanced text representation and encoder-based models to capture these nuances. A novel parallel CNN model, combining news articles, historical prices, and technical indicators, is proposed to improve accuracy. Validated using data from 2010–2021 for NIFTY 50, NIFTY Next 50, and NIFTY Bank indices, the model achieves high accuracy and strong trend correlation, demonstrating the value of integrating news and price data for forecasting.

2. Dataset

The dataset combines historical price data for NIFTY 50, NIFTY Next 50, and NIFTY Bank indices with financial news articles from 2010 to 2021, spanning 2,978 trading days. Articles were sourced from the *Economic Times*, focusing only on those published during market hours (9 AM to 4 PM IST), resulting in 215,740 relevant articles post-filtering.

Key technical indicators like ADX, MACD, RSI, and Bollinger Bands were calculated to enrich the dataset. Closing price movement was classified using simple rules: a label of 1 if the price rose or stayed the same compared to the previous day, and 0 otherwise.

This well-curated dataset, blending news insights with historical trends, forms the backbone of our analysis, helping uncover the short-term impact of economic news on market behavior.

3. Evaluation Metrics

1. **Accuracy:**
   Accuracy measures the percentage of correct predictions out of all predictions, calculated as

Introduction to Information Retrieval (IT550), Autumn 2024, DAIICT, India

202318035@daiict.ac.in (A. Anand)
202318035 (A. Anand); 202318057 (K. Anand); 202318050 (A. Singh)

Accuracy = $\frac{n}{N} \times 100$, where n is the number of correct predictions, and N is the total number of predictions. While accuracy gives an overall measure, it may not fully capture the model's ability to discriminate between classes.

2. **ROC-AUC:**
   The Receiver Operating Curve-Area Under Curve (ROC-AUC) assesses a model's ability to distinguish between positive and negative classes. The true positive rate (TPR) and false positive rate (FPR) are calculated at various thresholds, and the area under the ROC curve provides the AUC score. Scores range from 0.5 (random predictions) to 1 (perfect discrimination). Unlike accuracy, ROC-AUC accounts for the model's discriminative power, making it a more robust evaluation metric.
3. **Normalized Cross-Correlation (NCC):**
   NCC evaluates the alignment between predicted and actual time-series data by calculating their correlation when one series is time-shifted relative to the other. A high NCC score at a specific time step indicates strong alignment, and the lag/lead information reveals the sequential relationship. NCC is particularly useful for time-series data as it captures correlations and sequential patterns, which accuracy alone cannot reflect.

4. Results

4.1. Price-Only Approach

Using historical price data and technical indicators, this approach struggled to deliver robust results. Key findings include:

- **Performance:** Transformer Encoder achieved the highest accuracy for NIFTY 50 and NIFTY Bank but had a random-like ROC-AUC of 0.5. Random Forest outperformed other models in ROC-AUC for NIFTY 50 and NIFTY Bank, while SVM excelled for NIFTY Next 50.
- **NCC Scores:** Weak correlations with observed lag/lead patterns showed limited alignment with actual trends.

Conclusion: The price-only approach failed to uncover meaningful patterns, highlighting its limitations in capturing market dynamics.

4.2. Text-Only Approach

Leveraging news text with various classifiers yielded better results:

- **Performance:** SVM with TF-IDF achieved high accuracy and ROC-AUC among traditional methods. Encoder-based models, particularly FinBERT, improved accuracy and captured nuanced relationships.
- **NCC Observations:** Initial weak correlations improved over time, with encoder models achieving the best NCC scores and minimal lag.

Conclusion: Text-based features provided significant improvements, with FinBERT demonstrating the ability to extract complex relationships from financial news.

4.3. Combined Text and Price Approach

Integrating news and price data through a Parallel CNN model delivered the best results:

- **Performance:** The Parallel CNN achieved the highest accuracy and ROC-AUC across all indices, leveraging FinBERT for superior text representation.
- **Key Findings:** For NIFTY 50, 71.6% accuracy and an NCC score of 0.4261 were achieved. Similar strong results were observed for NIFTY Next 50 and NIFTY Bank.

4.4. Comparison of Best Approaches with the Research Paper

The study compared the three approaches (price-only, text-only, and combined):

Table 1
Average Accuracy of NIFTY Indices Prediction (Research Paper)

Approach	NIFTY 50	NIFTY NEXT 50	NIFTY BANK)
Baseline	0.6267	0.6064	0.6111
Text Only	0.7828 (+36.79%)	0.7463 (+28.93%)	0.7542 (+32.47%)
Combining Text and Price	0.8179 (+38.44%)	0.7683 (+29.54%)	0.7741 (+35.4%)

Table 2
Average Accuracy of NIFTY Indices Prediction (Our Results)

Approach	NIFTY 50	NIFTY NEXT 50	NIFTY BANK
Price only	0.5122	0.5022	0.4901
Text only	0.6842 (+25.13%)	0.6412 (+21.67%)	0.6307 (+22.29%)
Combining Text and Price	0.712 (+28.06%)	0.6502 (+22.76%)	0.6613 (+25.88%)

5. Key Challenges and Learnings

5.1. Key Challenges

One major challenge was accurately capturing the nuanced and unstructured context of financial news headlines. Choosing effective text representation methods like TF-IDF, BERT, or FinBERT required significant experimentation to uncover implicit relationships. Integrating this textual data with historical prices and technical indicators was equally complex, as it required precise alignment between news and market movements.

Another hurdle was handling the constantly shifting relationship between news and market indices. Models had to be robust to these changes, especially during out-of-sample testing. The parallel CNN architecture needed to capture both local and sequential patterns in text and price data efficiently. Implementing the expanding window strategy further added complexity, requiring retraining with new data while ensuring generalization. Balancing metrics like accuracy, ROC-AUC, and NCC demanded careful fine-tuning to optimize performance across models.

5.2. Key Learnings

Key insights emerged from these challenges. FinBERT excelled at capturing the context and subtle relationships in financial news, outperforming traditional methods. Combining text and price data provided a clearer picture of market behavior, with the parallel CNN model effectively leveraging both for accurate predictions.

The expanding window training strategy proved crucial in adapting to shifting market dynamics, while CNNs captured meaningful patterns in both text and price sequences. Metrics like ROC-AUC and NCC offered a deeper understanding of prediction reliability beyond accuracy.

This approach not only improved stock price predictions but also showed potential for applications like news recommendation, sentiment analysis, and fake news detection, demonstrating the power of multi-modal modeling.

6. References

1. S. Gangopadhyay and P. Majumder. 2023. Text representation for direction prediction of share market. *Expert Systems with Applications* 211 (2023), 118472. DOI:https://-doi.org/10.1016/j.eswa.2022.118472.
2. D. Araci. 2019. FinBERT: Financial sentiment analysis with pre-trained language models. *arXiv preprint arXiv:1908.10063.*
3. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805.*
4. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems* (NeurIPS), 5998–6008.