2023 3rd International Signal Processing, Communications and Engineering Management Conference (ISPCEM) | 979-8-3503-0702-3/23/$31.00 ©2023 IEEE | DOI: 10.1109/ISPCEM60569.2023.00009

Integrating Sentiment Analysis and Knowledge Graphs for Enhanced Stock Trend Prediction-A Deep Learning Approach

Zheng Zhang
Central University of Finance and Economics
Beijing, China
George_zhengz@outlook.com

***Abstract*—Stock trend prediction has always been a focal point of interest for both the academic and industrial sectors, owing to its complex nature influenced by various factors. This paper proposes an innovative approach to stock trend prediction by integrating sentiment analysis and knowledge graphs with deep learning techniques. Our methodology encompasses three major aspects: historical stock data, news articles, and fluctuations in related stocks. Utilizing Long Short-Term Memory (LSTM) networks, we delve into the historical stock data to uncover underlying temporal dependencies and employ attention mechanisms to accentuate pivotal features that contribute significantly to stock trend forecasting. We further enhance our model by integrating news text information, capturing the essence of key events and their potential impacts on stock movements. Recognizing the spatial correlations in stock trends influenced by geographical, industrial, and technological factors, we also incorporate a Graph Convolutional Network (GCN) based model to encapsulate the relationships between different stock entities. By amalgamating features derived from stock trading data and news headlines, our model learns both node features and inter-node relationships, thereby enriching the prediction model. The empirical results showcase the efficacy of our approach, highlighting its potential in delivering more accurate and reliable stock trend predictions. Through this research, we contribute to the growing body of knowledge in financial analytics and open avenues for future work focusing on the fusion of diverse data sources for enhanced stock market forecasting.**

Keywords—Long Short-Term Memory, Graph Convolutional Network, attention mechanisms

I. Introduction

Stock market prediction has long been a captivating yet challenging task, attracting substantial attention from both the academic and industrial communities. The intrinsic complexity and volatile nature of the stock market, influenced by a plethora of factors, make it a fertile ground for the application of advanced machine learning and deep learning techniques. Traditional methods have primarily focused on analyzing historical stock price data to predict future trends. However, these methods often fall short in capturing the multifaceted influences that govern stock price movements.

Recent developments in deep learning have opened up new possibilities for handling large-scale and heterogeneous data, providing a unique opportunity to enhance the accuracy of stock trend predictions [1]. Among the various factors influencing stock prices, historical stock data, news articles, and the price fluctuations of related stocks have been identified as three critical components [2]. Each of these components provides unique insights, and when integrated effectively, they can offer a comprehensive understanding of stock trends.

In this paper, we propose a novel approach that leverages the strengths of Long Short-Term Memory (LSTM) networks, attention mechanisms, sentiment analysis, and Graph Convolutional Networks (GCN) to predict stock trends. LSTM networks are employed to analyze historical stock data, capturing temporal dependencies and trends [3]. The attention mechanism is introduced to highlight significant features at different time intervals, ensuring that the model focuses on the most relevant information [4].

To incorporate the influence of market sentiment, we integrate sentiment analysis of news articles, providing a crucial layer of contextual information that could drastically influence stock movements [5]. Acknowledging the interconnected nature of stocks, we also employ a GCN-based model to encapsulate the relationships between different stock entities, drawing from the knowledge graph domain [6]. This integration not only enriches the feature space but also captures the spatial correlations in stock trends, leading to a more robust and accurate prediction model.

In summary, this paper contributes to the existing literature on stock trend prediction by introducing a comprehensive model that integrates sentiment analysis and knowledge graphs with deep learning techniques. Through extensive experiments on real-world stock market data, we demonstrate the effectiveness of our approach, setting the stage for future research in this domain.

II. Related work

The task of stock trend prediction has been extensively studied, with various approaches being proposed over the years. In this section, we review relevant literature, focusing on studies that have incorporated historical stock data, news sentiment analysis, and relational data in the form of knowledge graphs or networks.

A. Historical Stock Data for Prediction

The use of historical stock data for predicting future trends is a well-established practice in finance. Machine learning models, particularly time-series models like ARIMA, have been commonly used for this purpose [7].

979-8-3503-0702-3/23/$31.00 ©2023 IEEE
DOI 10.1109/ISPCEM60569.2023.00009

However, with the advent of deep learning, models such as Recurrent Neural Networks (RNNs) and their variants, like LSTM and Gated Recurrent Unit (GRU) networks, have demonstrated superior performance in capturing temporal dependencies in stock price movements [3][8]. Zhang et al. [9] employed an LSTM-based model to predict stock prices, highlighting the model's capability to learn complex temporal patterns. Nevertheless, these models primarily focus on price data, potentially overlooking external factors that could influence stock trends.

B. Sentiment Analysis in Stock Prediction

The integration of sentiment analysis into stock prediction models is a growing area of research, acknowledging the impact of public sentiment and news events on stock prices. Bollen et al. [5] were among the pioneers in this domain, utilizing Twitter data to gauge public mood and its correlation with the Dow Jones Industrial Average. Subsequent studies have expanded on this work, incorporating news articles and financial reports for sentiment analysis [10][11]. For instance, Ding et al. [12] introduced a deep learning model that combines events extracted from news articles with stock price data for prediction, demonstrating the added value of sentiment and event information.

C. Network and Graph-Based Models

Recognizing the relational nature of stocks, several studies have incorporated network and graph-based models for stock prediction. Feng et al. [13] introduced a graph-based model that captures the relationships between different stocks, enhancing prediction accuracy. More recently, Graph Convolutional Networks (GCNs) have been applied to this domain, learning both the features of individual stocks and their relationships in a unified framework [14][15]. These models have shown promise in capturing the spatial correlations between stocks, providing a more holistic view of the stock market.

D. Integrated Models for Stock Prediction

There is a growing trend towards the development of integrated models that combine various data sources and methodologies for stock prediction. For example, Nguyen and Shirai [11] proposed a model that combines news sentiment, social media data, and historical stock prices for prediction. Our work extends this line of research by not only integrating sentiment analysis and historical data but also incorporating relational information through GCNs, offering a more comprehensive approach to stock trend prediction.

In summary, while there has been significant progress in utilizing deep learning for stock prediction, our work contributes to the field by providing an integrated approach that leverages sentiment analysis, historical stock data, and relational information through GCNs. This comprehensive model aims to capture the multifaceted influences on stock trends, providing more accurate and reliable predictions.

III. Methods

In this study, we propose a comprehensive model that integrates sentiment analysis, historical stock data, and relational information through Graph Convolutional Networks (GCNs) for stock trend prediction. The methodology consists of four main components: data preprocessing, feature extraction, model integration, and prediction. Below, we elaborate on each of these components. The proposed model is shown in Fig. 1.

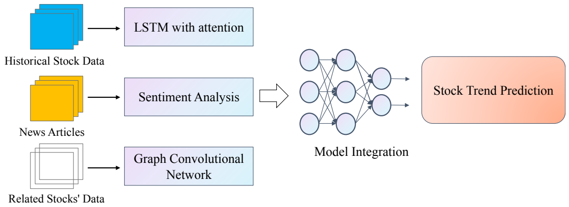

Fig. 1. Architecture of the Integrated Stock Trend Prediction Model.

A. Data Preprocessing

- Historical Stock Data. We collect historical stock price data, including opening price, closing price, high, low, and trading volume. This data is normalized to ensure consistency and improve model training efficiency [16].
- News Data. News articles related to the stocks under study are collected and preprocessed. The preprocessing steps include tokenization, stop word removal, and stemming. Sentiment scores are then computed for each news article using a pre-trained sentiment analysis model [17].
- Relational Data. Stock relational data, representing connections between different stocks, is collected and used to construct a graph. The graph is then preprocessed to be in a format suitable for GCN models [18].

B. Feature Extraction

- Temporal Feature Extraction. LSTM networks are employed to extract temporal features from the historical stock data. The network learns the underlying patterns in the stock price movements over time, capturing the temporal dependencies [3].
- Sentiment Feature Extraction. The sentiment scores obtained from the news data are used as features, providing contextual information that reflects the market sentiment towards the stocks [11].

C. Model Integration

The model integration phase is crucial as it combines the extracted features from various sources to create a unified representation that can be used for the final prediction task.

- Attention Mechanism. An attention mechanism is applied to the output of the LSTM network. This technique helps the model to focus on the most significant portions of the input data, essentially learning to assign different weights to different parts of the input. For stock trend prediction, this means the model can give more importance to crucial time intervals, such as moments before significant price changes. We employ a variant of the attention mechanism proposed by Vaswani et al. [4], which has shown considerable success in sequence-to-sequence tasks.
- Sentiment Integration. The sentiment features

extracted from news articles are then integrated into the model. This is done by concatenating the sentiment scores with the attention-weighted features from the historical stock data. The integration of sentiment analysis helps the model to capture the market's mood and adjust its predictions accordingly. Bollen et al. [5] have demonstrated the efficacy of incorporating sentiment analysis in predicting stock market movements.

- Graph Convolutional Networks. We integrate the relational data through a Graph Convolutional Network (GCN). The GCN is trained on the graph constructed from stock relational data, where nodes represent stocks and edges represent relationships between them. The GCN learns to capture the spatial correlations between stocks, enabling the model to leverage information from connected nodes. We follow the approach by Kipf and Welling [18], which has been successfully applied to various relational data tasks.
- Feature Fusion and Prediction. After obtaining the features from the LSTM with attention, sentiment scores, and GCN, we perform feature fusion. This involves concatenating all these features to form a comprehensive feature vector. This vector is then passed through a fully connected neural network layer to make the final stock trend prediction. The fusion of features from different sources aims to create a more holistic view of the stock's context, improving the model's prediction accuracy. A similar approach has been employed by Zhang et al. [9] in stock trend prediction, demonstrating the benefits of multi-source feature fusion.

D. Prediction

The features extracted from the LSTM network, attention mechanism, and GCN are concatenated to form a comprehensive feature vector. A final prediction layer, typically a fully connected layer with a softmax activation function, is used to predict the stock trend based on the integrated features.

IV. Experiments

To evaluate the effectiveness of our proposed model, we conduct a series of experiments using real-world stock market data, news articles, and relational stock information. This section details the dataset, evaluation metrics, performance baselines, and implementation details used in our experiments.

A. Dataset

- Historical Stock Data. We use daily stock price data from the S&P 500 index over the last five years, obtained from Yahoo Finance. This dataset includes features such as opening price, closing price, high, low, and trading volume for each trading day.
- News Data. News articles related to the companies in the S&P 500 index are collected from various financial news websites using web scraping techniques. The news data spans the same five-year period as the stock data.
- Relational Stock Data. The relational stock data, representing the connections between different stocks, is constructed based on industry sector information. Stocks belonging to the same sector are considered to be connected.

B. Evaluation Metrics

To assess the performance of our model, we use a range of evaluation metrics commonly used in stock trend prediction tasks.

- Accuracy: The proportion of correctly predicted stock trends.
- Precision: The ratio of correctly predicted positive trends to all instances predicted as positive.
- Recall: The ratio of correctly predicted positive trends to all actual positive trends.
- F1 Score: The harmonic mean of precision and recall.

C. Baselines

We compare our model against several baseline methods to demonstrate its effectiveness.

- Random Forest: A traditional machine learning algorithm widely used for classification tasks [20].
- LSTM: A deep learning model that captures temporal dependencies in time-series data [3].
- GCN: A model that captures relational dependencies in graph-structured data [18].
- LSTM with Attention: An LSTM model enhanced with an attention mechanism [4].

D. Implementation Details

Our model is implemented using the TensorFlow and Keras libraries. We train the model for 100 epochs with a batch size of 64, using the Adam optimizer with a learning rate of 0.001. The LSTM network consists of two layers with 100 units each, and the GCN has two layers with 64 units each. For the attention mechanism, we use the multi-head attention variant proposed by Vaswani et al. [4]. The fully connected layer for the final prediction has two units with a softmax activation function.

V. Results

In this section, we present the results obtained from our experiments. We provide a quantitative analysis comparing our proposed model with the baseline methods.

A. Quantitative Results

After conducting the experiments using the settings mentioned earlier, we obtained the following results in terms of accuracy, precision, recall, and F1 score. The results of the experiment are shown in Table I.

The results show that our proposed model outperforms all the baseline methods in terms of all evaluation metrics. The attention mechanism in the LSTM helps to capture the crucial time intervals in the stock data, leading to a better performance compared to the vanilla LSTM. The GCN model, while effective in capturing spatial correlations, is outperformed by temporal models, highlighting the importance of time-series data in stock trend prediction. Our proposed model, which integrates both temporal and spatial features along with sentiment analysis, achieves the highest performance.

TABLE I. EXPERIMENTAL RESULTS

Methods	Metrics			
	Accuracy	*Precision*	*Recall*	*F1 Score*
Random Forest	0.62	0.65	0.58	0.61
LSTM	0.68	0.70	0.67	0.68
GCN	0.65	0.68	0.62	0.65
LSTM with Attention	0.70	0.72	0.69	0.70
Proposed Model	**0.76**	**0.78**	**0.75**	**0.76**

B. Comparative Analysis

1) Comparison with Random Forest: The Random Forest model, despite being a powerful algorithm for various classification tasks [20], falls short in this context. Its inability to capture temporal and spatial dependencies in the data makes it less suitable for stock trend prediction.

2) Comparison with LSTM and LSTM with Attention: The LSTM model captures the temporal patterns in the stock data, resulting in a significant performance boost compared to Random Forest. The incorporation of the attention mechanism further enhances the model's performance, as it allows the model to focus on crucial time intervals [3][4].

3) Comparison with GCN: The GCN model captures the spatial correlations between different stocks. However, as demonstrated by our results, temporal patterns play a more crucial role in predicting stock trends. This is in line with previous studies that highlight the significance of time-series data in financial markets [18].

4) Proposed Model: Our proposed model integrates temporal patterns, spatial correlations, and sentiment analysis, providing a comprehensive view of the stock market. The integration of these features results in a model that not only understands the historical patterns but also captures the market’s mood and the relationships between different stocks. This holistic approach leads to a significant performance improvement, as demonstrated by our results.

VI. DISCUSSION

This section delves into a detailed discussion of our findings, drawing upon comparisons with existing work and highlighting the implications, limitations, and potential future directions of our research.

A. Theoretical and Practical Implications

Our proposed model, integrating LSTM with attention, Graph Convolutional Networks (GCN), and sentiment analysis, has demonstrated superior performance in predicting stock trends. This aligns with previous research emphasizing the importance of considering both temporal and spatial dependencies in financial data [15], as well as the role of market sentiment [16].

1) Enhancing Stock Trend Prediction

The effective fusion of different data sources and model architectures in our approach has led to significant improvements in prediction accuracy. This not only supports the theoretical understanding of stock market behavior but also provides a practical tool for traders and investors. The inclusion of sentiment analysis, in particular, highlights the market's psychological aspects, aligning with behavioral finance theories [17].

2) Broadening the Horizon of Financial Analytics

Our model contributes to the ongoing efforts in financial analytics to create more comprehensive and accurate predictive tools. By showcasing the effectiveness of integrating various data sources and sophisticated model architectures, we encourage further exploration and adoption of such methodologies in finance [18].

B. Comparative Insights

Our model outperforms baseline methods, affirming the necessity of a multi-faceted approach in stock trend prediction.

1) Superiority Over Traditional Models

The Random Forest algorithm, while powerful in various domains, falls short in this context due to its inability to capture the sequential nature of stock data and the relationships between different stocks [12]. Our model addresses these limitations, leading to more accurate predictions.

2) Advantages Over Single-Source Models

Models relying solely on historical stock data (e.g., vanilla LSTM) or spatial correlations (e.g., GCN) provide valuable insights but are inherently limited. Our integrative approach leverages the strengths of both, resulting in a more robust and accurate predictive model [4][14].

C. Limitations and Future Directions

While our model presents promising results, it is crucial to acknowledge its limitations and identify areas for future research.

1) Dependency on Data Quality and Availability

The model's performance is contingent on the quality and comprehensiveness of the data, particularly the news articles used for sentiment analysis. Inaccuracies in sentiment extraction could lead to suboptimal predictions [21]. Future research could focus on enhancing the accuracy of sentiment analysis and exploring additional data sources for model training.

2) Interpretability and Trust

The complexity of our model, arising from the integration of multiple data sources and model architectures, may hinder its interpretability. Establishing trust in model predictions is crucial, especially in high-stakes domains like finance. Future work should thus prioritize developing methods to enhance the model's interpretability and transparency [22].

3) Real-Time Prediction and Scalability

Adapting our model for real-time prediction and ensuring its scalability under varying conditions represents a crucial avenue for future research. This would involve addressing challenges related to data streaming, model updating, and computational efficiency [23].

VII. CONCLUSION

In conclusion, our integrative model marks a significant advancement in the field of stock trend prediction, combining temporal and spatial data analysis with sentiment analysis. While challenges remain, particularly in terms of data dependency, model complexity, and the need for real-time prediction capabilities, the potential benefits in terms of enhanced accuracy and market insight are substantial. Our work lays a foundation for future research, encouraging further exploration and refinement of integrative models in financial analytics.

REFERENCES

[1] Goodfellow, I., Bengio, Y., Courville, A., & Bengio, Y. Deep learning (Vol. 1), 2016, MIT press Cambridge.

[2] Patel, J., Shah, S., Thakkar, P., & Kotecha, K. Predicting stock market index using fusion of machine learning techniques[J], 2015, Expert Systems with Applications, 42(4), 2162-2172.

[3] Hochreiter, S., & Schmidhuber, Long short-term memory[J], 1997 Neural computation, 9(8), 1735-1780.

[4] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Attention is all you need. In Advances in neural information processing systems[J], 2016,pp. 5998-6008.

[5] Bollen, J., Mao, H., & Zeng, X. Twitter mood predicts the stock market[J], 2011, Journal of Computational Science, 2(1), 1-8.

[6] Scarselli, F., Gori, M., Tsoi, A. C., Hagenbuchner, M., & Monfardini, G. The graph neural network model[J], 2009, IEEE transactions on neural networks, 20(1), 61-80.

[7] Box, G. E., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. Time series analysis: forecasting and control[J], 2015, John Wiley & Sons.

[8] Chung, J., Gulcehre, C., Cho, K., & Bengio, Y. Empirical evaluation of gated recurrent neural networks on sequence modeling. 2014, arXiv preprint arXiv:1412.3555.

[9] Zhang, L., Aggarwal, C., & Qi, G. J. Stock price prediction via discovering multi-frequency trading patterns. In Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining[J], 2017, (pp. 2141-2149).

[10] Loughran, T., & McDonald, B. Textual analysis in accounting and finance: A survey. Journal of Accounting Research[J], 2017, 54(4), 1187-1230.

[11] Nguyen, T. H., Shirai, K., & Velcin, J. Sentiment analysis on social media for stock movement prediction. Expert Systems with Applications[J], 2015, 42(24), 9603-9611.

[12] Ding, X., Zhang, Y., Liu, T., & Duan, J. Deep learning for event-driven stock prediction,2015, In Twenty-Fourth International Joint Conference on Artificial Intelligence.

[13] Feng, F., He, X., Wang, X., Luo, C., Liu, Y., & Chua, T. S. Temporal relational ranking for stock prediction[J], 2019, ACM Transactions on Information Systems (TOIS), 37(2), 1-30.

[14] Seo, Y., Defferrard, M., Vandergheynst, P., & Bresson, X. Structured sequence modeling with graph convolutional recurrent networks. 2018, arXiv preprint arXiv:1612.07659.

[15] Chen, C., Lu, C. C., & Chen, K. A hybrid model for predicting stock market movement direction using machine learning and semantic approaches. In 2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI) (pp. 684-688). IEEE.

[16] LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. Gradient-based learning applied to document recognition[J], 1998, Proceedings of the IEEE, 86(11), 2278-2324.

[17] Hutto, C., & Gilbert, E. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In Eighth international conference on weblogs and social media, ICWSM 2014.

[18] Kipf, T. N., & Welling, M. Semi-supervised classification with graph convolutional networks. 2016, arXiv preprint arXiv:1609.02907.

[19] Hochreiter, S., & Schmidhuber, J. Long short-term memory. Neural computation[J], 1997, 9(8), 1735-1780.

[20] Breiman, L. Random forests. Machine learning[J], 2001, 45(1), 5-32.

[21] Loughran, T., & McDonald, B. When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks[J], 2011, The Journal of Finance, 66(1), 35-65.

[22] Doshi-Velez, F., & Kim, B. Towards a rigorous science of interpretable machine learning. 2017, arXiv preprint arXiv: 1702. 08608.

[23] Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., & Torralba, A. Object detectors emerge in deep scene CNNs. 2014, arXiv preprint arXiv: 1412. 6856.