STOCK MARKET COMMUNITY DETECTOR: SENTIMENT-ENRICHED GRAPHICAL NEURAL NETWORKS FOR FINANCIAL ECOSYSTEM ANALYSIS

Rishabh Ghosh, Eric Liu, Haoxuan Lu, Sheffali Rawat, Anneketh Vij
University of Southern California
{ghoshris,zhongkun,hlu46565,srawat,anneketh}@usc.edu

1 INTRODUCTION

Sentiment analysis has revolutionized the financial industry by offering deep insights into the collective mood of the market. Traditional models, such as FinBERT(Araci, 2019), have been instrumental in this regard, yet they often lack the capability to translate sentiment into concrete financial predictions. This paper proposes an innovative approach that synergizes advanced sentiment analysis with the power of graph neural networks (GNNs)(Jegelka, 2022) to refine stock market forecasts and provide financial professionals with an in-depth, data-centric perspective on market trends derived from social media data. Furthermore, our methodology extends to include community profiling, which discerns the accuracy of users' predictive sentiments, thereby augmenting the predictive power of our model.

In today's dynamic financial environment, precise interpretation and quantification of market sentiment are paramount. Our initiative represents a significant stride in innovation, fulfilling the pressing demand for analytical models that not only comprehend but also quantify the financial ramifications of market sentiment and pinpoint influential community members. Such advancements herald a new era of informed and strategic decision-making within the financial sector.

While models like FinBERT have set the stage for sentiment analysis in finance, they exhibit two main limitations:

- Inability to accurately quantify the financial impact of sentiment, leading to a gap in actionable financial forecasting. Sentiment analysis can detect positive stock sentiment on social media but often fails to predict the precise financial effect, such as the exact percentage increase in stock price. This imprecision leaves a gap in actionable financial forecasting.
- Difficulty in capturing the complex web of relationships and contextual nuances present in financial discussions on social media. Traditional sentiment analysis may identify keywords in financial discussions on social media but struggles to grasp the nuanced opinions and complex relationships, like the varied impacts of new policies on different tech sectors. This leads to an incomplete picture of market sentiment.

To overcome these challenges, we propose a solution that leverages:

- **Advanced Analytical Prowess of GNNs:** Graph Neural Networks (GNNs) excel in analyzing interconnected data. For example, in a network of companies with supplier and consumer relationships, a GNN can detect communities, such as a cluster of companies that frequently interact. This insight is valuable for understanding market trends or financial risk propagation.
- **Synergistic Approach Combining Sentiment Analysis with Structural Data Analysis:** The integration of sentiment analysis with structural data analysis provides a comprehensive market view. For instance, positive investor sentiment towards renewable energy stocks, coupled with structural analysis revealing a close-knit community of such companies, suggests a favorable market trend for these stocks. This combined approach offers insights into the sentiment, context, and relationships within the market.

Our project's contributions are manifold and significant:

- **Reduction in Prediction Time:** Our project has made significant strides in decreasing the time needed to generate predictions. For example, FinBERT previously took minutes, but with our model it now take mere seconds, enabling swift responses to market dynamics.
- **Increased Prediction Precision:** We've significantly improved the accuracy of our stock market predictions. Previously, the correlation between a message's sentiment and the actual stock movement was about 50%. Our models have boosted this accuracy to 60% or higher, supporting more lucrative investment choices.
- **Profit Enhancement:** Our model analyzes financial social media data to assess the accuracy of sources' historical predictions, providing insights into market influence. Unlike FinBERT, which could predict losses, our model consistently forecasts profits, never dipping below zero.
- **Source Reliability Classification:** We have effectively categorized social media sources into reliable and non-reliable groups. This helps in assessing the credibility of the information and strengthens the data's integrity for financial decision-making.

Our empirical evaluations and case studies confirm the effectiveness of our method. It enhances sentiment analysis for finance and provides advanced tools for financial experts to adapt to the dynamic market.

2 Related Work

The integration of sentiment analysis with Graph Neural Networks (GNNs) has been a significant area of focus within the financial sector. This section discusses the spectrum of related work, comparing each study to our proposed model.

- (Ullah et al., 2022) and (Zhou et al., 2020) have provided a broad overview of GNNs, focusing on their general application and enhancement. Our model specifically leverages GNNs to analyze the interconnected data of financial markets, capturing complex relationships that traditional sentiment analysis models often miss.
- (Zhao et al., 2022b) demonstrated the adaptability of GNNs in sentiment classification. Our model extends this adaptability by classifying sentiment and quantifying its financial impact, addressing the gap in actionable financial forecasting.
- (Xiang et al., 2022) and (Liao et al., 2021) refined sentiment analysis techniques by incorporating linguistic nuances. Our model builds upon this by combining nuanced sentiment analysis with GNNs to provide a more comprehensive view of market sentiment and its implications.
- (Araci, 2019) leveraged pre-trained models like FinBERT to interpret complex sentiments in financial texts. Our model integrates these sentiments with structural data analysis through GNNs, enhancing predictive power concerning stock market trends.
- (Jaggi et al., 2021) explored the predictive power of sentiment analysis on social media platforms. Our model complements this by analyzing sentiment and using GNNs to understand the underlying structural patterns within the data, leading to more informed predictions.
- (Matsunaga et al., 2019) and (Jafari & Haratizadeh, 2022) focused on stock market prediction by combining GNNs with sentiment analysis. Our model offers a more precise quantification of the financial impact of sentiment, filling the gap left by these studies in actionable financial forecasting.
- (Peng et al., 2023) and (Zhao et al., 2022a) utilized company knowledge graphs and integrated stock relations and attention mechanisms to enhance prediction accuracy. Our model incorporates these elements and introduces a novel methodology that quantifies the financial implications of market sentiment, setting new benchmarks in the field.

Our model merges recent advances to address the gaps in sentiment analysis. It combines GNNs' power with thorough sentiment analysis to offer financial experts a detailed market sentiment view, enhancing financial forecasting and decision-making.

3 PROBLEM FORMULATION

3.1 TASK OVERVIEW

Our research aims to enhance financial sentiment analysis by integrating Graph Neural Networks (GNNs) with traditional techniques, focusing on sentiment interpretation, quantification, and community detection to improve stock market trend predictions. The goal is to develop a model that not only interprets market sentiment from social media data but also predicts stock market trends with greater accuracy, aiding financial professionals in decision-making. This comprehensive approach aims to dissect our overarching goals into the following specific objectives:

- **Graph Transformation:** The GNN processes interconnected data, discerning sentiment dependencies and relationships within the financial discourse.
- **Inferred Connections and Sentiment Strength:** The GNN infers latent connections between entities, assigning sentiment strength through edge logits, crucial for understanding sentiment propagation.
- **Node Detection and Impact Evaluation:** The enriched graph identifies influential nodes, evaluating their impact using sentiment consistency and financial metrics, providing actionable insights for decision-makers.
- **Analysis of Inferred Edges:** Our comprehensive approach analyzes inferred edges and logits, offering insights into hidden relationships within the financial network for accurate market trend predictions.

3.2 DATASET DESCRIPTION

Our primary dataset, known as "Financial Tweets" Wallach (2018), consists of social media posts related to the financial sector. Selected for its detailed textual and relational information, it is essential for conducting sentiment analysis on financial subjects. We enhanced the dataset by integrating stock quotes for the mentioned financial products, obtained from Yahoo Finance[1]. This integration offers a holistic perspective on market sentiment as it unfolds. However, due to the unavailability of historical quotes for certain symbols, our initial dataset of over 28,000 records was reduced to just over 23,000. Additionally, the data compilation spanned seven days, in line with the operational constraints of the Twitter API.

3.3 METRICS DEFINITION

We employ the following metrics in our model:

- **Edge Logits:** These represent the strength of sentiment between connected entities in the graph, providing a quantitative sentiment measure.
- **Node Reliability:** This metric assesses the consistency and intensity of sentiment associated with each node, reflecting its influence on the network.
- **Profitability Impact:** It measures the financial impact of entities based on end-of-day (EOD) trade values, linking sentiment to market performance.

These metrics were chosen to offer a comprehensive view of sentiment analysis, combining qualitative and quantitative data for a robust financial market prediction model.

4 METHODOLOGY

4.1 DATA PREPROCESSING

We begin data preprocessing by removing URLs from text messages. Next, we add columns for end-of-day (EOD) outcomes of financial symbols to the dataset. These outcomes, while quantitatively valuable, do not reflect sentiment and thus guide our unsupervised analysis.

[1]Accessible via `https://github.com/ranaroussi/yfinance`

4.2 SENTIMENT ANALYSIS WITH FINBERT

Utilizing FinBERT, a BERT fine-tuned SOTA model for financial contexts, we perform sentiment analysis on financial tweets. The model outputs a logit vector, $\mathbf{l} = [l_{\text{pos}}, l_{\text{neg}}, l_{\text{neu}}]^\top$, to probabilistically classify sentiments. We prioritize positive and negative sentiments, as neutral sentiment lacks decision-making value. Consequently, we omit the neutral component, focusing on positive and negative sentiments to clearly interpret market sentiment.

4.3 GRAPH CREATION AND REPRESENTATION

Post-sentiment analysis, we construct a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ to represent sentiment and outcome data. Vertices $\mathcal{V}$ symbolize users and target products, while edges $\mathcal{E}$ link them, reflecting sentiment from FinBERT logits and actual outcomes. This graph structure, with its enriched edge attributes, serves as a tool for detailed sentiment and outcome analysis. Refer Figure 1 for the initial graph that is being created.

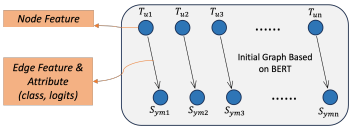

Figure 1: Initial Graph

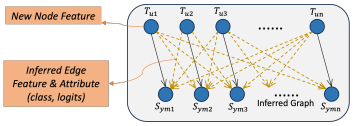

Figure 2: Inferred Graph

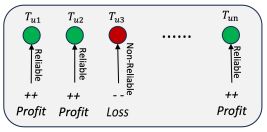

Figure 3: Node Prediction

4.4 GCN NODE ENCODER AND LINEAR LAYER FOR CLASSIFICATION AND LOGITS

We employ a Graph Convolutional Network (GCN) node encoder, something similar to (Kipf & Welling, 2016), for feature learning, tailored to our dataset's specific needs. The GCN utilizes ReLU-activated layers to update node features, following the formula:

$$\mathbf{H}^{(l+1)} = \text{ReLU}\left(\hat{\mathbf{D}}^{-\frac{1}{2}}\hat{\mathbf{A}}\hat{\mathbf{D}}^{-\frac{1}{2}}\mathbf{H}^{(l)}\mathbf{W}^{(l)}\right),$$

where $\hat{\mathbf{A}}$ is the adjacency matrix with self-loops, $\hat{\mathbf{D}}$ is the degree matrix, $\mathbf{H}^{(l)}$ is the hidden layer representation (we have used 16 hidden layers), and $\mathbf{W}^{(l)}$ is the layer's weight matrix. Node features undergo linear transformation to form matrix $\mathbf{Z}$. The matrix's first column, $\mathbf{Z}_{[:,:1]}$, is activated via tanh function for class categorization. The second column, $\mathbf{Z}_{[:,1:2]}$, serves as an additional feature in classification. Subsequent columns, $\mathbf{Z}_{[:,2:]}$, leverage softmax activation to generate sentiment class probabilities. This approach ensures the model outputs definitive class labels alongside sentiment probabilities. We refer to Figure 4 for an overview of our core design.

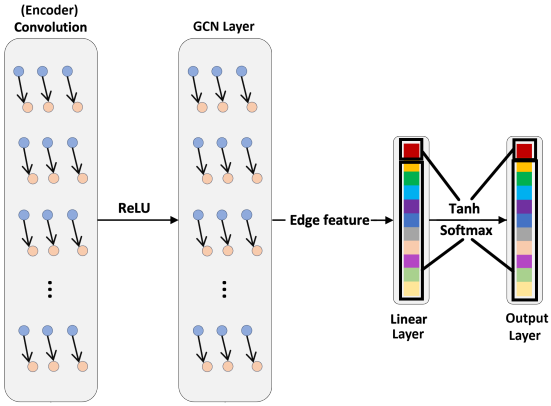

Figure 4: GCN Encoder and Linear Layer

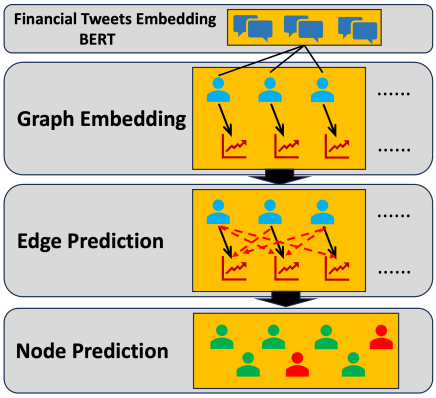

Figure 5: High level view of our model

4.5 EDGE PREDICTION AND NODE DETECTION

We perform edge prediction by deducing new edges and their sentiment logits to improve the graph's connectivity; see Figure 2. Node detection is conducted by applying thresholds to classify and determine influential nodes based on sentiment and market impact; see Figure 3. Given our dataset comprises only equities, we use mean values for thresholding. The model is refined through a training loop that minimizes the mean squared error (MSE) loss, improving the accuracy of edge feature predictions and the graph's depiction of the financial network. We Refer to Figure 5 for a very high overview of our model.

5 RESULTS AND DISCUSSION

5.1 ACCURACY OF CORRECTLY PREDICTED MESSAGES

In this unsupervised learning context, while we don't have a traditional accuracy metric, we do measure the accuracy of logits against the true end-of-day outcomes. FinBERT correctly predicted 11,799 out of 23,171 total messages. Our model improved upon this, correctly predicting 11,891 messages, marking a 0.78% increase over FinBERT's performance.

5.2 TOTAL PROFIT ALONG WITH MODEL VARIATIONS AND BIAS EXPERIMENTS

Our model surpasses FinBERT in generating profits across 5393 nodes, as outlined in Table 1. It adeptly navigates intricate financial networks to devise superior trading tactics. Tailored for diverse scenarios, the model's flexibility provides a range of trading options, exemplified in Table 2.

Node Classifier Type	**Our Model Reliable Node %**	**FinBERT Reliable Node %**	**Our Model Correctly Predicted Reliable Edges %**	**FinBERT Correctly Predicted Reliable Edges %**	**Our Model Net Profit %**	**FinBERT Net Profit %**
1	47.84%	50.69%	47.99%	51.67%	0.0996	-0.0218
2	89.49%	51.06%	40.84%	46.10%	0.1880	-0.0572
3	37.33%	50.87%	47.21%	51.06%	0.0999	-0.0214
4	37.09%	50.83%	47.12%	51.17%	0.1007	-0.0222
5	37.34%	50.88%	47.23%	51.05%	0.0992	-0.0208
6	39.64%	50.84%	46.48%	51.17%	0.1070	-0.0170
7	37.33%	50.87%	47.21%	51.06%	0.0999	-0.0214

Table 1: Updated Comparative Analysis of Our Model Vs FinBERT Across Different Node Classifier Types

Node Classifier Type	**Description**	**Edge Threshold**	**Outcome Threshold**
1	Based on edge type prediction	Mean	-
2	Based on true outcome prediction	-	Mean
3	Balanced bias	Mean	Mean
4	More bias to outcome prediction	Mean + 1%	Mean
5	More bias to edge prediction	Mean	Mean + 1%
6	Less bias to edge prediction	Mean - 1%	Mean
7	Less bias to outcome prediction	Mean	Mean - 1%

Table 2: Mapping of Node Classifier Types and it's Metrics

The Table 1, utilizing the entire dataset, reveals that Type 2 classifiers tend to overfit. Net profits are derived from trading based on the model's predictions, where positive logits suggest buying, and negative logits indicate selling. Although FinBERT appears to have more dependable nodes and predicts more connections than our model, it fails to measure them, resulting in a net loss. Addressing this issue was our initial goal.

5.3 Enhanced Product Message Prediction and Trading Profitability

Our model excels beyond FinBERT by forecasting and quantifying product-related user messages. Prior to the market opening on the final day, the model was trained, enabling it to swiftly predict new edge connections and generate approximately 1 million logits for these edges in mere seconds. With the onset of trading, these pre-computed logits facilitated accurate sentiment analysis for strategic trading. The model's superior performance was evident, securing a profit margin of 0.411%, significantly outstripping FinBERT's 0.025%.

5.4 Efficient Computational Model for Financial Analysis

Our approach outperforms FinBERT by reducing computation time through sentiment analysis and graph encoding integration. While FinBERT takes roughly 2 hours to process data, our model completes training in a mere 30 seconds, crucial for real-time decision-making in the financial market. It leverages FinBERT for immediate message processing and updates daily logits post-market for the subsequent trading day, ensuring no delays during vital trading times.

5.5 t-SNE Visualization of Edge Feature Distribution

Figure 6 presents a t-SNE visualization that delineates the distribution of edge features within our dataset. A label of 1 indicates correct prediction of the message, whereas a label of -1 denotes an incorrect prediction. The leftmost graph in Figure 6 depicts the state post-FinBERT analysis and pre-training, the center graph illustrates the scenario after training our model with labels according to FinBERT's predictions, and the rightmost graph displays the results post-training our model with labels based on our own predictive outcomes.

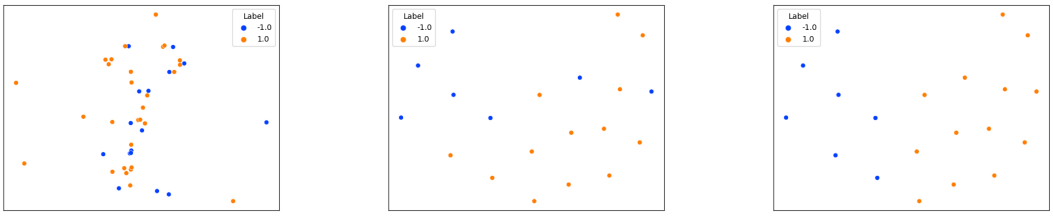

Figure 6: t-SNE distribution of edge features before (left), after training with FinBERT labels (center), and after training with our model's predictions (right).

6 Conclusion and Future Work

Our study[2] presents a novel approach by integrating FinBERT with a Graph Convolutional Network (GCN) to enhance financial data analysis. This combination allows for a more comprehensive understanding of market sentiments, leading to better-informed trading decisions. The model's ability to quantify profit and loss provides traders with a distinct advantage over traditional BERT models.

The main limitation we encountered was the dataset's limited scope, confined to only 7 days of Twitter data. This constraint hinders the depth and breadth of our analysis. To improve our project, expanding the dataset to include a wider range of financial products and a longer time frame would be beneficial. Additionally, incorporating data from various sources could provide a more robust and diverse dataset for analysis.

Future improvements include scaling the model to handle diverse financial datasets and real-time data streams, which would bolster its practicality for dynamic market conditions and contribute to more nuanced trading strategies. Further research will also explore the integration of more sophisticated graph-based techniques, which could lead to the development of more holistic trading strategies. These improvements are geared towards establishing the model as a versatile and comprehensive decision-making tool in the ever-evolving financial market landscape.

[2]The code and all associated results are available at: `https://github.com/RishabhGhosh/StockMarketCommunityDetector/`

REFERENCES

Dogu Araci. Finbert: Financial sentiment analysis with pre-trained language models, 2019. URL https://doi.org/10.48550/arXiv.1908.10063.

Alireza Jafari and Saman Haratizadeh. Gcnet: Graph-based prediction of stock price movement using graph convolutional network. *Engineering Applications of Artificial Intelligence*, 116:105452, 2022. ISSN 0952-1976. doi: https://doi.org/10.1016/j.engappai.2022.105452. URL https://www.sciencedirect.com/science/article/pii/S0952197622004420.

Mukul Jaggi, Priyanka Mandal, Shreya Narang, Usman Naseem, and Matloob Khushi. Text mining of stocktwits data for predicting stock prices. *Applied System Innovation*, 4(1):13, February 2021. ISSN 2571-5577. doi: 10.3390/asi4010013. URL http://dx.doi.org/10.3390/asi4010013.

Stefanie Jegelka. Theory of graph neural networks: Representation and learning, 2022. URL https://arxiv.org/pdf/2204.07697.pdf.

Thomas N. Kipf and Max Welling. Variational graph auto-encoders, 2016. URL https://arxiv.org/pdf/1611.07308.pdf.

Wenxiong Liao, Bi Zeng, Jianqi Liu, Pengfei Wei, Xiaochun Cheng, and Weiwen Zhang. Multi-level graph neural network for text sentiment analysis. *Computers Electrical Engineering*, 92:107096, 2021. ISSN 0045-7906. doi: https://doi.org/10.1016/j.compeleceng.2021.107096. URL https://www.sciencedirect.com/science/article/pii/S0045790621001051.

Daiki Matsunaga, Toyotaro Suzumura, and Toshihiro Takahashi. Exploring graph neural networks for stock market predictions with rolling window analysis, 2019. URL https://doi.org/10.48550/arXiv.1909.10660.

Hao Peng, Ke Dong, and Jie Yang. Stock price movement prediction based on relation type guided graph convolutional network. *Engineering Applications of Artificial Intelligence*, 126:106948, 2023. ISSN 0952-1976. doi: https://doi.org/10.1016/j.engappai.2023.106948. URL https://www.sciencedirect.com/science/article/pii/S0952197623011326.

Ihsan Ullah, Mario Manzo, Mitul Shah, and Michael G. Madden. Graph convolutional networks: analysis, improvements and results. *Applied Intelligence*, 52(1):13, July 2022. ISSN 1573-7497. doi: 10.1007/s10489-021-02973-4. URL https://doi.org/10.1007/s10489-021-02973-4.

David Wallach. Financial tweets. https://www.kaggle.com/datasets/davidwallach/financial-tweets/data, 2018.

Chunli Xiang, Junchi Zhang, Fei Li, Hao Fei, and Donghong Ji. A semantic and syntactic enhanced neural model for financial sentiment analysis. *Information Processing Management*, 59(4):102943, 2022. ISSN 0306-4573. doi: https://doi.org/10.1016/j.ipm.2022.102943. URL https://www.sciencedirect.com/science/article/pii/S0306457322000656.

C. Zhao, X. Liu, J. Zhou, Y. Cen, and X. Yao. Gcn-based stock relations analysis for stock market prediction. *PeerJ Computer Science*, 8:e1057, 2022a. doi: 10.7717/peerj-cs.1057. URL https://doi.org/10.7717/peerj-cs.1057.

Meng Zhao, Jing Yang, Jianpei Zhang, and Shenglong Wang. Aggregated graph convolutional networks for aspect-based sentiment classification. *Information Sciences*, 600:73–93, 2022b. ISSN 0020-0255. doi: https://doi.org/10.1016/j.ins.2022.03.082. URL https://www.sciencedirect.com/science/article/pii/S0020025522003103.

Jie Zhou, Ganqu Cui, Shengding Hu, Zhengyan Zhang, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li, and Maosong Sun. Graph neural networks: A review of methods and applications. *AI Open*, 1:57–81, 2020. ISSN 2666-6510. doi: https://doi.org/10.1016/j.aiopen.2021.01.001. URL https://www.sciencedirect.com/science/article/pii/S2666651021000012.

A APPENDIX

Task Description	**Member (Last Name)**
Develop and fine-tune the FinBERT model to state-of-the-art performance levels, and expand the dataset through augmentation techniques.	Liu, Rawat
Create and implement a script to retrieve financial quotes from Yahoo Finance, and enhance the dataset with additional data augmentation.	Lu, Vij
Design and implement a Graph Convolutional Network (GCN) for accurate edge prediction and node detection within the financial market graph.	Ghosh
Conduct a thorough analysis of the model results, including statistical evaluation and interpretation of the financial market predictions.	Ghosh, Liu, Rawat, Vij
Investigate alternative graph-based methodologies for quantifying and analyzing financial market structures and relationships.	Lu
Prepare and present a poster summarizing the research findings, methodologies, and implications for the financial market analysis.	Liu, Lu, Vij
Compose the final research paper, detailing the project's objectives, methodologies, results, and conclusions.	Ghosh, Rawat

Table 3: Distribution of Tasks Among Team Members