공학석사학위논문

웹 검색량 기반 주가 변동 예측을 위한 변화하는 주식 관계 모델링

Modeling Changing Stock Relations using Web Search Volume for Stock Price Movement Prediction

서울대학교 대학원
협동과정 인공지능전공
박재민

웹 검색량 기반 주가 변동 예측을 위한 변화하는 주식 관계 모델링

Modeling Changing Stock Relations using Web Search Volume for Stock Price Movement Prediction

지도교수 강 유

이 논문을 공학석사 학위논문으로 제출함

2022년 12월

서울대학교 대학원

협동과정 인공지능전공

박재민

박재민의 석사 학위논문을 인준함

2022년 12월

위 원 장	권 태 경	(인)
부 위 원 장	강 유	(인)
위 원	송 현 오	(인)

Abstract

Modeling Changing Stock Relations using Web Search Volume for Stock Price Movement Prediction

Jae Min Park
Interdisciplinary Program in Artificial Intelligence
The Graduate School
Seoul National University

Given historical stock prices and web search volumes of selected keywords, how can we accurately predict stock price predictions? Stock price movement prediction is an attractive task for its applicability in real-world investments. Even a slight improvement in performance can lead to enormous profit. However, the task is extremely challenging due to the inherently volatile and random nature of the stock market. To overcome such difficulties, many researchers have tried to utilize relationships between stocks to make predictions. Despite the effort, previous works have failed to incorporate the dynamic characteristic of stock relationships as they heavily relied on predefined concepts to find stock correlations. However, correlations between stocks change over stime and are not dependent on a single criterion.

In this paper, we propose GFS (**G**raph-based **F**ramework using changing relations for **S**tock price movement prediction), a novel framework for stock price movement prediction using web search volumes to capture the changing relations between stocks. GFS combines relationship information from stationary connections based on predefined concepts with variable connects made from the correlations of each stock's web search volumes collected using tickers. In addition, from the fact that stock prices are affected by global trends, we collect web search volumes of 5 keywords that best represent a common denominator of the target stocks. Experimental results on a 1-year dataset of semiconductor stocks listed in the U.S. stock market show that our model achieves higher accuracy than its baselines.

Keywords : Stock Price Movement Prediction, Recurrent Neural Network, Time Series Classification, Google Trends Dataset
Student Number : 2021-21550

Table of Contents

I. Introduction **1**

II. Related Work **7**

2.1 Individual Stock Price Prediction 7

2.2 Correlated Stock Price Prediction 8

III. Proposed Method **9**

3.1 Overview 9

3.2 Attentive Feature Extraction 12

3.3 Utilization of Stationary and Trend Graphs 14

3.4 Keyword-based global trend extraction 15

3.5 Stock Price Movement Prediction 15

IV. Experiment **17**

4.1 Experiment Settings 17

4.2 Classification Performance 18

4.3 Ablation Study 18

V. Conclusion **21**

References **22**

Abstract in Korean **24**

List of Figures

Figure 1. Correlated stock price movements between Pepsi (Ticker: PEP) and Coca-cola (Ticker: COKE) from 2018 to 2021. 3

Figure 2. Stock price movements between NVIDIA (Ticker: NVDA) and Abbott Laboratories (Ticker: ABT) before and after the pandemic outbreak. 4

Figure 3. GFS outperforms its baselines in terms of accuracy; and achieves higher MCC than the average. 5

Figure 4. The overall architecture of GFS. 11

List of Tables

Table 1. Price features representing the daily trend of a stock. 13
Table 2. Google Trends features representing the daily trend of a keyword. 13
Table 3. Summary of our dataset SMC2021. 18
Table 4. The classification performance of GFS and competitors, measured with the accuracy (ACC) and the Matthews Correlation Coefficient (MCC); higher values mean better performances. GFS shows the best performance for all the cases. 19
Table 5. Results of ablation study. Each module of GFS contributes to improving the classification performance of the model. 20

Chapter 1

Introduction

Given historical stock prices and web search volumes of selected keywords, how can we accurately predict stock price movements? Stock price movement prediction is one of the most famous topics in the field of financial AI for its challenges and applicability in real-world investments. Even a slight improvement in performance could lead to enormous profit. Although predicting stock price movement seems simple, the problem is extremely challenging due to the inherently volatile and random nature of the stock market. Therefore, in practice, market analysts often evaluate future movements of stock prices based on correlations with other stocks. Figure 1 demonstrates correlated movements of two stocks, Pepsi and Coca-Cola. The figure shows stock prices from 2018 to 2021. As can be seen, the two prices are highly correlated. Mining such relationships can be a meaningful source of information for predicting future stock prices. For instance, one could guess that Pepsi's stock price will fall if Coca-Cola's price had been falling.

Many researchers have tried to incorporate the relationships like Pepsi and Coca-Cola to make stock price movement predictions. They concentrated on connecting stocks based on predefined concepts such as industry or products, concepts that do not change much over time. However, relying on predefined concepts to define relationships face the following challenges: First, stock price movements are not always explainable through predefined concepts. Thus, for a prediction model that utilizes relationship information, relying on a fixed relationship could alter model performance when used on different datasets. Second, fixed relationships cannot capture the changing relationships between stocks. Events such as the COVID-19 pandemic can drastically change the correlations between stocks. Figure 2 shows the changing correlations between the stock prices of NVIDIA and Abbot Laboratories, a pharmaceutical company based in the U.S., before and after the outbreak of the pandemic. The correlation between prices after the pandemic was 0.88, whereas the correlation before the pandemic was only 0.07. Third, relying on predefined concepts cannot target stocks that are in niche sectors.

In this paper, we propose GFS (**G**raph-based **F**ramework using changing relations for **S**tock price movement prediction), a novel framework for stock price movement prediction using web search volumes to capture the changing relations between stocks. The main ideas of GFS are as follows: First, we create context vectors for both historical prices and web search volumes that effectively summarize features of multiple time steps. The created context vectors are used in the subsequent modules of the model. Second, GFS utilizes two graphs, which are the stationary and trend graphs. We use the first graph to represent stationary relationships and the latter for capturing changing relationships. We build trend graphs based on correlations of web search volumes collected from Google Trends as it has been proven that web search volume from Google Trends is an effective factor for stock market predictions. [1] Google trends is a search trends tool that shows frequently a keyword is entered into Google's search engine. Third, GFS learns the global trends by combining global trends context vectors with the feature representations of existing graphs. We capture global trends by extracting 5 keywords from a set of related news articles and creating feature vectors from the web search volumes of the keywords.

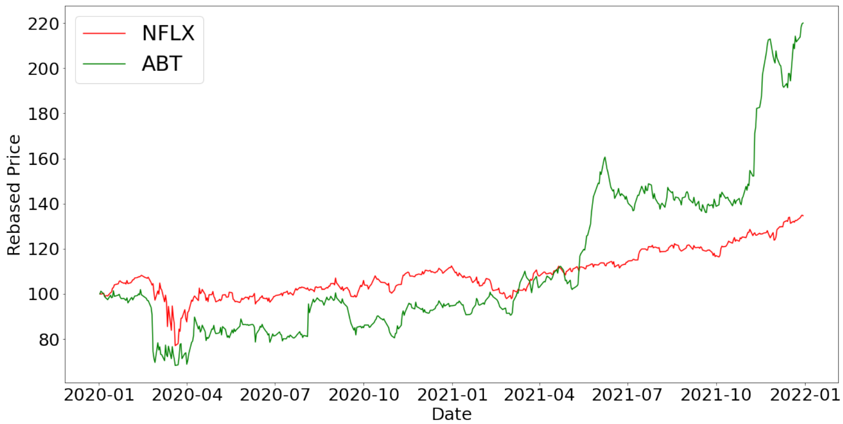
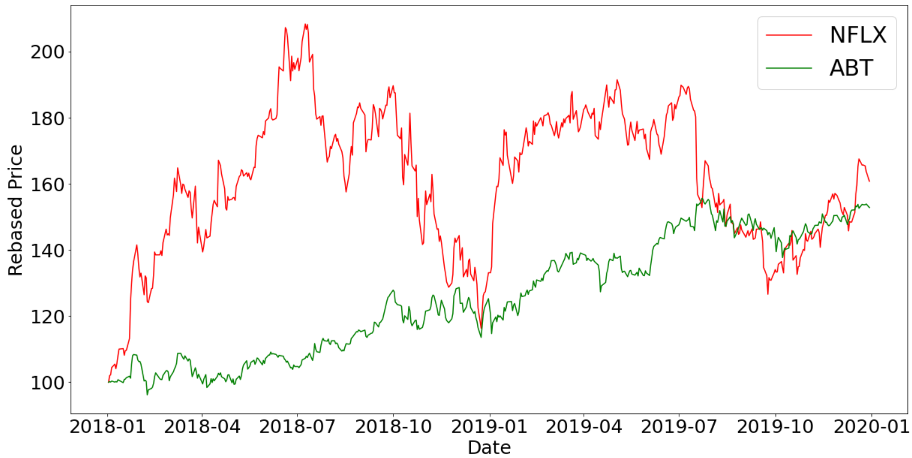
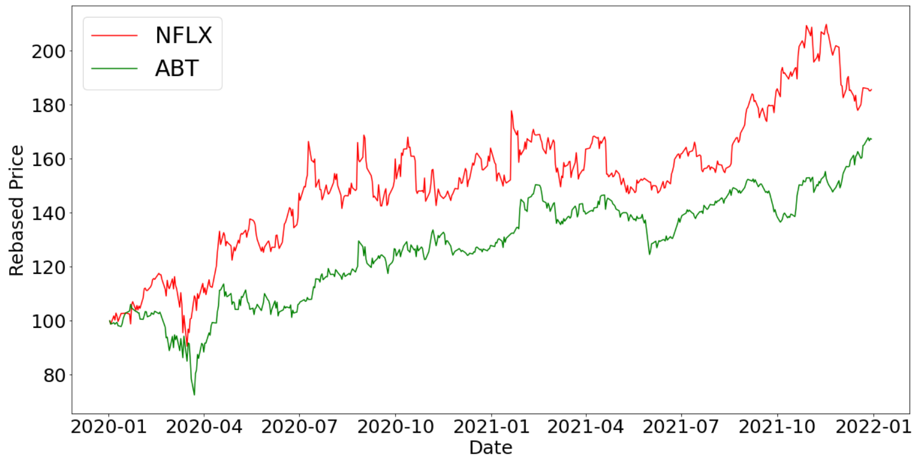


Figure 1: Correlated stock price movements between Pepsi (Ticker: PEP) and Coca-cola (Ticker: COKE) from 2018 to 2021.

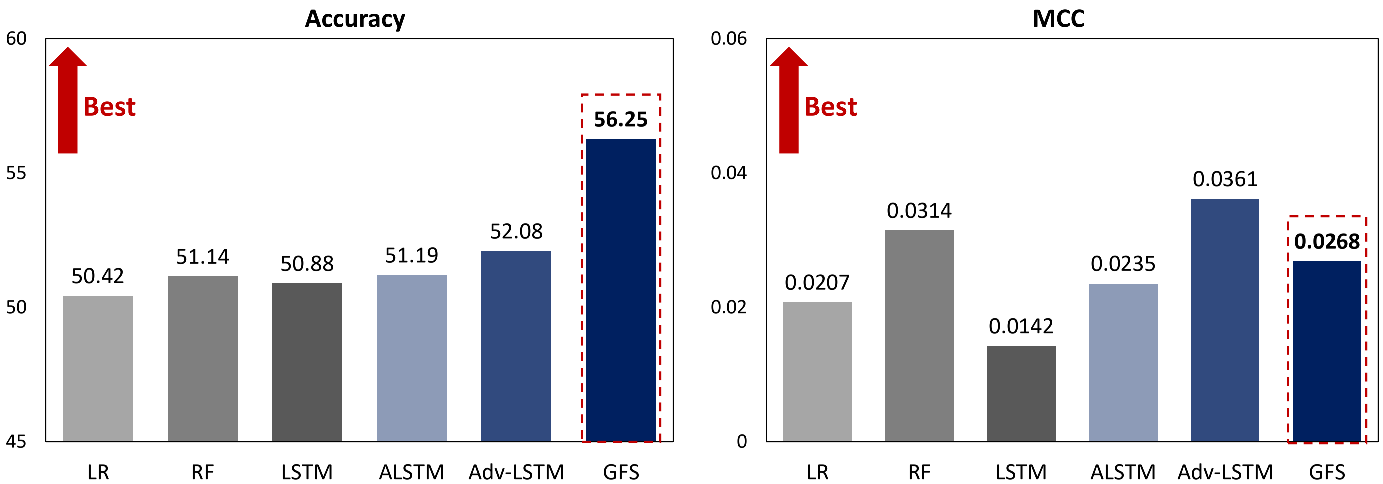


(a) Before COVID-19



(b) After COVID-19

Figure 2: Stock price movements between NVIDIA (Ticker: NVDA) and Abbott Laboratories (Ticker: ABT) before and after the pandemic outbreak.



(a) Accuracy **(b) Matthews Correlation Coefficient**

Figure 3: GFS outperforms its baselines in terms of accuracy; and achieves higher MCC than the average.

The main contributions of this paper are as follows:

- **Model.** We propose GFS, a novel framework for stock price movement prediction using web search volumes to capture the changing relations between stocks.
- **Dataset.** We provide a dataset that contains historical prices, web search volumes, and semiconductor industry-related news articles.
- **Experiments.** Experiments are run on the collected dataset and we show that GFS outperforms its baselines by 4.17%p in accuracy.

The rest of the paper is organized as follows: In Section 2, we introduce related works regarding stock price prediction, which are categorized into individual stock price prediction and correlated stock price prediction models. In Section 3, the main ideas and respective details of GFS are introduced. In Section 4, experimental results on the top 16 semiconductor stocks listed in the U.S. stock market are presented. Lastly, we conclude in Section 5.

Chapter 2

Related Work

In this section, related works on stock price movement predictions are introduced. The approaches can be divided into individual stock price prediction models and correlated stock price prediction models.

2.1 Individual Stock Price Prediction

Individual stock price predictions models can be divided into two categories based on the data used: 1) historical price-based prediction models and 2) event-driven stock price prediction models

Historical price-based models aim to find repetitive patterns embedded in the historical prices of stocks. The most widely used methods in this field are Autoregressive (AR) and ARIMA models. However, these two models show low performance under volatile environments, for not being able to effectively reflect the time-dependency of price data. More recent approaches tried to incorporate the time-dependency of data into prediction. These works include utilizing recurrent neural networks (RNN), especially long-short term memory units (LSTM) [2], to address long-time dependency problems. Moreover, [3] attaches an attention module on top of LSTM to create context vectors. [4] modifies the training process of ALSTM to further improve prediction performance.

Limitations in the performance of historical price-based models have triggered the use of other sources of information for the prediction task. Some of the approaches in this domain include considering news [5, 6, 7], SNS [8, 9], and other relevant public company information. We note that none of the approaches above refer to web search volumes collected from Google Trends to make stock price movement predictions.

2.2 Correlated Stock Price Prediction

This approach uses the relationships between multiple stocks to make price movement predictions. The main challenge for this approach is to find quality relationships as model performance is strongly affected by the relationship information. DA-RNN [3] learns correlations between stocks from historical prices but is limited to using closing prices only. [10] proposed a framework that correlates stocks according to patterns apparent in prices. [11] proposed a GNN-based approach that connects stocks based on business relationships to make price movement predictions. [12] also proposed a GNN-based approach that connects stocks based on related shareholder information.

The approaches mentioned above all share the same limitations, which is that these methods all rely on predefined concepts to connect stocks. [13] proposed a novel framework that automatically correlates multiple stocks, but it only utilizes historical prices to find correlations, limiting the type of relationships that can be found.

Our objective in this paper is to incorporate dynamic changes in relations between multiple stocks, overcoming the limitations of the previous works. We note that none of the research above utilizes non-price data to incorporate dynamic changes in stock relations.

Chapter 3

Proposed Method

In this section, we propose GFS (**G**raph-based **F**ramework using changing relations for **S**tock price movement prediction), a novel method for stock price movement using web search volumes to capture dynamic changes in stock relations.

3.1 Overview

The objective of this work is to make accurate stock price movement predictions by capturing the changing relations that can be found in web search volumes. The challenges addressed in this paper are as follows:

1. **Considering time series features.** Time series data used in the model consists of multiple features a day. Therefore, it is important to extract meaningful feature representations while preserving the time-related aspect of each feature. How can we effectively create feature representations from the time series data?

2. **Incorporating changing relations between stocks.** Although relationships such as industry or product types do not change drastically, relations between stocks that affect stock prices change continuously. Therefore, it is important to capture the changing relations between stocks to make accurate price predictions. How can we accurately capture and incorporate the changing relations between stocks?

3. **Capturing global industry trend.** The relationship between stocks is also affected by the global trend as can be seen in Figure 2. How can we capture the global industry trend and incorporate it into the stock price movement predictions?

We address the challenges with the following ideas:

1. **Attentive Feature Extraction.** For each time series, we extract the temporal relationships within the price features of multiple time steps using Attentive LSTM(ALSTM) [3]. ALSTM outputs context vectors that summarize information of multiple time steps and the generated vectors are used as price features in the subsequent modules.

2. **Utilization of stationary and trend graphs** In order to fully exploit the relationship information of the stocks, we adopt Graph Convolutional Network (GCN) to create feature representations. We build two graphs that represent different types of relationships between stocks. The stationary graph represents stationary relationships such as product types, whereas the trend graph represents the changing relationships apparent in the public interest. We select web search volumes as a measure of public interest.

3. **Keyword-based global trend extraction** From a set of related news articles, we extract 5 keywords that best represent the common denominator of the given stocks. Using each keyword as the search term, we collect web search volumes from Google Trends and generate context vectors using the attentive feature extraction method. The context vectors are then integrated with the feature representations of the trend graph; so that we reflect global trends along with stock relations to make accurate predictions.

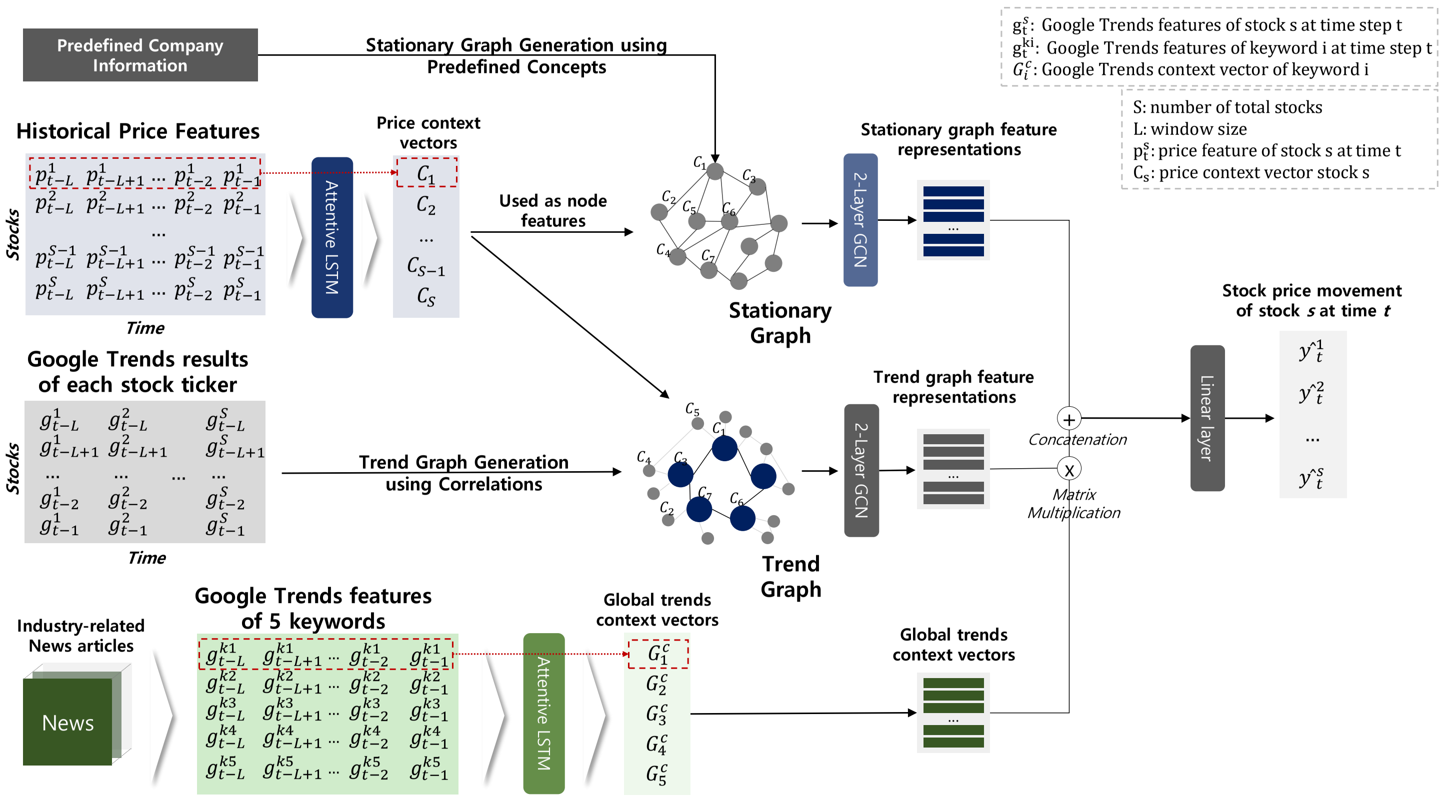


Figure 4: The overall architecture of GFS.

Figure 4 demonstrates the overall architecture of GFS. First, assuming that we have already extracted the 5 keywords and are given corresponding web search volumes, we create context vectors from historical prices and web search volumes using ALSTM. Second, price context vectors are used as features for the two graphs. We feed the two graphs into GCN modules, where node features, edge adjacencies, and edge weights are given as inputs. Third, GCN outputs feature representation of each graph, which are then integrated with the context vectors from the global trends to finally make stock price movement predictions. Other than the keywords extraction process, GFS operates in an end-to-end process.

3.2 Attentive Feature Extraction

The first step of GFS is to create context vectors for each time series data: historical prices and web search volumes. Given time series data with multiple features, the objective is to output context vectors that summarize features of multiple previous time steps.

LSTM. LSTM has been proven to be effective in making feature representations of time series data with multiple features. Given a feature vector z_{st}, where s and t denote stock and time indices respectively, LSTM updates two state vectors h_{t-1}, c_{t-1} of the previous time step and creates new state vectors h_t, c_t. The state vector h_T from the last step becomes the final representation of the given time series. However, this results in the limitation that information from the previous time steps $[1, 2, ..., t-1]$ are forgotten.

Attentive LSTM. In order to extract more accurate feature representations, we adopt Attentive LSTM (ALSTM). Instead of using the last hidden state as an output, ALSTM utilizes the attention mechanism to compute a weighted sum of all the hidden

Table 1: Price features representing the daily trend of a stock.

Features	**Calculation**
$c\_open,\ c\_high,\ c\_low$	$e.g.,\ c\_open = open_t/close_t - 1$
n_close	$e.g.,\ n\_close = close_t/close_{t-1} - 1$
$5\_day,\ 10\_day,\ 15\_day$ $20\_day,\ 25\_day,\ 30\_day$	$e.g.,\ 5\_day = \frac{\sum_{i=0}^{4} close_{t-i}/5}{close_t} - 1$

Table 2: Google Trends features representing the daily trend of a keyword.

Features	**Calculation**
1_day	$e.g.,\ 1\_day = volume_t/volume_t - 1$
$5\_day,\ 10\_day,\ 15\_day$ $20\_day,\ 25\_day,\ 30\_day$	$e.g.,\ 5\_day = \frac{\sum_{i=0}^{4} 1\_day_{t-i}}{5}$

states $[h_1, h_2, \dots h_T]$ from LSTM using attention scores as weights. Each attention weight measures the importance of time step i regarding the current time step T. The result of the weighted sum produces a context vector C_s. The equations for each component are as follows:

$$a_i = \frac{exp(exp(h_j^T h_T))}{\sum_{j=1}^{T} exp(h_j^T h_T)}, C_s = \sum_{i=0}^{T} a_i h_i \tag{3.1}$$

By calculating the weighted sum of every hidden state vector, ALSTM can incorporate information from previous time steps, overcoming the limitation of LSTM. [14]

3.3 Utilization of Stationary and Trend Graphs

The second step is to create graphs that connect related stocks. The stationary graph is given, whereas the trend graph is constructed at every time step. Both graphs consist of node features, edge index, and edge weights. We feed each graph into a 2-layer GCN module that outputs a feature representation of the input graph features as GCN allows each node to share information from not only their direct neighbors but also other nodes that are connected to the neighbors. The details of each graph are as follows:

Stationary Graph. We build the stationary graph using predefined concepts. The types of predefined concepts used in the model are manufacturer type, product type, and country of origin, where satisfying each criterion gives two stocks an edge weight of 0.4, 0.4, and 0.2 respectively. Other than the node features, the graph remains constant throughout the whole training process.

Trend Graph. To incorporate changing relations concerning public interest, we build a trend graph at every time step using correlations of web search volume change ratios. Instead of simply connecting stocks based on correlations, we use a novel graph generation method that emphasizes key stocks that have high aggregated correlations. First, we calculate correlations between every pair. Second, the top 5 stocks with the highest total correlations are selected as central nodes. Third, each stock is connected to one of the central nodes, one the stock has the highest correlation with. Lastly, stocks with correlations higher than 0.8 are connected, regardless of the node's centrality. This method allows stocks that have low correlation with other stocks to be a part of the network so that the model can incorporate information about more stocks and even stocks without direct connection.

3.4 Keyword-based global trend extraction

From the given news articles, we extract 5 keywords that best represent the common property of the target stocks. In this paper, we gathered information on the stocks in the semiconductor industry. Therefore, the 5 keywords are terms that best represent the semiconductor industry. Using the keywords as search terms, we collect web search volumes of each keyword. The collected web search volumes are created into context vectors in the attentive feature extraction process. We perform matrix multiplication of the context vectors and the feature representations of the trend graph to finally output a trend representation that also reflects changing global trends. The details of the integration process are further explained in the next subsection.

3.5 Stock Price Movement Prediction

The last step is to make predictions using the feature representations of the two graphs as inputs.

Reflecting Global Trend. To reflect global trends on the feature representation of the trend graph, we transpose the feature matrix that consists of global trends context vectors and perform multiplication with the existing trend graph representation. This results in a new trend graph representation of shape S x 5, where S denotes the number of total stocks and 5 is the number of keywords. Then, we concatenate the refined trend feature representations with the existing feature representation of the stationary graph and feed them into the final prediction layers.

Final Prediction. We apply a single linear layer to the concatenated graph feature representations to produce the final prediction as the following equation:

$$\hat{y} = \sigma(HW + b)$$

The logistic sigmoid function σ is applied to interpret each element $\hat{y}_d^s$ as the price movement probability for stock s. probability higher than 0.5 indicates that a stock price is expected to rise, otherwise fall. Each element $\hat{y}_d$ is the output of GFS for stock price movement prediction at time step d.

The model aims to minimize Binary Cross Entropy loss, which is computed as follows:

$$Loss(X, y) = \text{-}(y \log(p) + (1\text{-}y) \log(1\text{-}p)), \tag{3.2}$$

where X is an input tensor and y is the label, which implies actual stock movements. P is the computed price movement probability.

Chapter 4

Experiment

In this section, the performance of GFS is evaluated experimentally. This paper aims to answer the following questions:

- **Q1. Classification Performance**: Can the model make accurate stock price movement predictions?
- **Q2. Ablation Study**: Does each idea contribute to improving the model performance?
- **Q3. Hyperparameter Senstivity**: Is GFS robust to the window size and dimension of hidden units?

4.1 Experiment Settings

Dataset. As no publicly available datasets exist for this problem, we build a dataset consisting of historical price data, web search volumes for each stock ticker, and semiconductor industry-related news articles collected from major news media for the time period of 01/2021 to 12/2021. We label each instance according to the change rate of closing prices. The change rate is calculated as $close_t/close_{t-1} - 1$.

Parameter settings. We optimize GFS using the Adam optimizer with a batch size of 16. We also select the learning rate, and lag window size in the range of [0.0001, 0.01] and [5, 10, 15, 20, 40] respectively. Our selected best model uses a window size of 20, with a learning rate of 0.001.

Table 3: Summary of our dataset SMC2021.

Dataset	# Stocks	# News	# Trends	Train	Valid	Test
SMC2021	16	2,741	21	01/2021-	06/2021-	09/2021-12/2021

Baselines The performance of GFS is compared with the following baselines:

- LR: A logistic regression model using historical stock price features.
- RF: A random forest classifier using historical stock price features.
- LSTM: LSTM model based on historical stock price features. [2]
- ALSTM: LSTM using attention mechanism. [3]
- Adv-ALSTM: ALSTM with adversarial training. [4]

Evaluation metrics. We use two metrics to evaluate the performance: Accuracy (ACC) and Matthews Correlation Coefficient (MCC) [15] which considers the imbalance of classes.

4.2 Classification Performance

We compare the classification performance of GFS and its baselines. The results are summarized in Table 4 and Figure 3. GFS achieves up to 4.17%p higher accuracy than its best competitor, and 0.0014 higher MCC than the average.

4.3 Ablation Study

We compare the performance of GFS and its variants where each module is removed from the original model in Table 5. Results from the ablation study shows that

Table 4: The classification performance of GFS and competitors, measured with the accuracy (ACC) and the Matthews Correlation Coefficient (MCC); higher values mean better performances. GFS shows the best performance for all the cases.

	SMC2021	
Method	ACC	MCC
LR	50.42	0.0207
RF	51.14	0.0314
LSTM	50.88	0.0142
ALSTM	51.19	0.0235
Adv-ALSTM	52.08	0.0360
GFS (proposed)	**56.25**	**0.0268**

each module contributes to improving the model performance.

- GFS-SG: GFS without the stationary graph (SG)
- GFS-TG: GFS without the trend graph (TG)
- GFS-GT: GFS without the global trends (GT)

We can notice that each module improves the prediction accuracy, and GFS with both graphs records the highest performance.

In addition, we note that TG is more important than SG, with it capturing dynamic changes in stock relationships. SG operates as a supporting module on top of TG, which improves the stability of stock relationships by incorporating definitive stock-to-stock connections. Also, reflecting the global trends (GT) aids other modules to make more accurate predictions.

Table 5: Results of ablation study. Each module of GFS contributes to improving the classification performance of the model.

	SMC2021	
Method	ACC	MCC
ALSTM	52.08	0.0360
GFS-SG	47.24	0.0209
GFS-TG	51.32	-0.0481
GFS-GT	53.00	-0.0207
GFS (proposed)	**56.25**	**0.0268**

Chapter 5

Conclusion

We propose GFS, an accurate stock price movement prediction model that utilizes stationary relationships and changing trend relationships. We create graph embeddings of the two graphs and combine the result from the trend graph with the created global industry trend context vectors. We feed the final embeddings to the prediction layers to make stock movement predictions. Experimental results show that GFS achieves higher accuracy than all of its baselines by incorporating changing relations between the stocks. Future works include extending the model to predict stock movements using only web search volume data or other sources of information to build temporal graphs. Also, the keyword extraction process could be performed dynamically.

References

[1] T. Preis, H. S. Moat, and H. E. Stanley, "Quantifying trading behavior in financial markets using google trends," *Scientific reports*, vol. 3, no. 1, pp. 1–6, 2013.

[2] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[3] Y. Qin, D. Song, H. Chen, W. Cheng, G. Jiang, and G. Cottrell, "A dual-stage attention-based recurrent neural network for time series prediction," *arXiv preprint arXiv:1704.02971*, 2017.

[4] F. Feng, H. Chen, X. He, J. Ding, M. Sun, and T.-S. Chua, "Enhancing stock movement prediction with adversarial training," *arXiv preprint arXiv:1810.09936*, 2018.

[5] S. Mohan, S. Mullapudi, S. Sammeta, P. Vijayvergia, and D. C. Anastasiu, "Stock price prediction using news sentiment analysis," in *2019 IEEE Fifth International Conference on Big Data Computing Service and Applications (BigDataService)*, pp. 205–208, IEEE, 2019.

[6] X. Ding, Y. Zhang, T. Liu, and J. Duan, "Deep learning for event-driven stock prediction," in *Twenty-fourth international joint conference on artificial intelligence*, 2015.

[7] Q. Li, L. Jiang, P. Li, and H. Chen, "Tensor-based learning for predicting stock movements," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 29, 2015.

[8] H. Wu, W. Zhang, W. Shen, and J. Wang, "Hybrid deep sequential modeling for social text-driven stock prediction," in *Proceedings of the 27th ACM international conference on information and knowledge management*, pp. 1627–1630, 2018.

[9] Y. Xu and S. B. Cohen, "Stock movement prediction from tweets and historical prices," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1970–1979, 2018.

[10] C. Li, D. Song, and D. Tao, "Multi-task recurrent neural networks and higher-order markov random fields for stock price movement prediction: Multi-task rnn and higer-order mrfs for stock price classification," in *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 1141–1151, 2019.

[11] D. Matsunaga, T. Suzumura, and T. Takahashi, "Exploring graph neural networks for stock market predictions with rolling window analysis," *arXiv preprint arXiv:1909.10660*, 2019.

[12] Y. Chen, Z. Wei, and X. Huang, "Incorporating corporation relationship via graph convolutional neural networks for stock price prediction," in *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pp. 1655–1658, 2018.

[13] J. Yoo, Y. Soun, Y.-c. Park, and U. Kang, "Accurate multivariate stock movement prediction via data-axis transformer with multi-level contexts," in *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pp. 2037–2045, 2021.

[14] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[15] B. W. Matthews, "Comparison of the predicted and observed secondary structure of t4 phage lysozyme," *Biochimica et Biophysica Acta (BBA)-Protein Structure*, vol. 405, no. 2, pp. 442–451, 1975.

요 약

과거 주가와 관련 키워드 웹 검색량이 주어졌을 때 주가의 변동을 어떻게 정확하게 예측할 수 있을까? 주가 예측은 많은 각광을 받고 있으며 약간의 성능 개선으로도 실제 투자에서 많은 이익을 얻을 수 있기에 매우 매력적인 주제이다. 주가의 움직임을 예측한다는 것은 비록 간단해보이지만 주가의 본질적인 변동성으로 인해 매우 어렵다. 이를 극복하기 위한 방안으로 많은 방법들이 주식 간 상관관계 정보를 활용하기 위해 시도해 왔다. 그러나 이전 연구들은 사전에 정의된 정보를 기반으로 고정된 관계만을 사용하거나 과거 가격만을 사용하여 계속해서 변화하는 주식들간의 관계를 예측에 활용하는데 실패하였다.

본 논문에서는 주식 관계의 동적 변화를 사용해 주가의 변동을 예측하는 방법인 GFS (**G**raph-based **F**ramework using changing relations for **S**tock price prediction)를 제안한다. GFS는 사전에 정의된 정보를 활용한 그래프와 함께 웹 검색량으로부터 주식들간의 상관관계를 계산하여 매번 새로운 그래프를 생성하여 사용한다. 또한, GFS는 뉴스로부터 글로벌 산업 트렌드를 나타내는 키워드를 추출하여 얻은 웹 검색량의 특성을 효과적으로 추출하여 글로벌 산업 트렌드 벡터를 생성한다. 두 그래프와 글로벌 산업 트렌드 벡터는 모두 GFS가 정확한 주가 변동을 예측 하는 것에 상당 부분 기여하며, 실험 결과를 통해 GFS가 주가 변동 예측 분야에서 최고 수준의 정확도를 제공함을 확인할 수 있다.

주요어 : 주가 예측, 순환 신경망, 시계열 분류, 구글 트렌드 데이터

학번 : 2021-21550