Vector Representation of Stocks from the Perspective of Natural Language and Price History
（自然言語と価格に基づく株のベクトル表現）

Dept. of Mathematical Informatics 48186218　杜 鑫
Supervisor　田中 久美子 教授

1 Introduction

News articles influence the dynamics of financial markets. Past works developed various deep learning architectures for news-driven stock price prediction, but have seen few gains. Meanwhile, they did not offer a way to generalize the mutual effects learned from the text and price data for other uses.

This thesis takes a new approach by explicitly describing this mutual effects in terms of a vector, called *stock embedding*. A stock is represented by a vector so that its inner product with a news text vector produces a larger value when the text is more related to the stock.

Two major advantages come from such a vector representation of stocks. First, it concentrates the stock-specific knowledges and enables the sharing of other parameters across stocks. This is an important advantage to alleviate data sparseness and prevent overfitting. Second, stock embeddings are portable. They are easy to put to use in various financial applications other than prediction. We show an example of portfolio optimization, which is the first application of natural language processing (NLP) to the modern portfolio theory to the best of our knowledge. In this example, our method achieved 2.8x times more gains compared to the original method using only price data.

A part of this work has been sent to ACL 2020 [1].

2 Past Works

The idea of stock embeddings in this thesis comes from NLP, where vectors are trained to represent words [6], sentences [5] and even full articles [2]. The geometry of such an embedding system usually contains rich semantic information. For example, two word embeddings with large cosine values are often semantically close to each other. The acquisition of stock embeddings in this thesis is based on this original idea.

The framework for acquiring the stock embeddings in this thesis is directly based on the work in [3, 4]. Compared to these works dedicated to price movement prediction, my thesis concentrates on the generalization and representation of the learned knowledge, and its application to other financial usage.

As one such application, the stock embeddings are evaluated in terms of portfolio optimization. The mean-variance minimization portfolio model proposed by Markowitz [7] was used in my work. To the best of my knowledge, this is the first paper applying NLP techniques to the modern portfolio theory.

3 Text-driven price movement classification

The *text-driven price movement classification* task is used to acquire such stock embeddings. The task is defined as follows:

$$\min_f \text{loss} = -\frac{1}{JT}\sum_{j=1}^{J}\sum_{t=1}^{T}(y_t^j \log \hat{y}_t^j + (1-y_t^j)\log(1-y_t^j)),$$

where

$$\hat{y}_t^j \equiv f(N_{[t-d_1,t+d_2]}), \quad y_t^j \equiv \begin{cases} 1 & \text{if } p_t^j > p_{t-1}^j \\ 0 & \text{otherwise.} \end{cases}$$

f is typically a neural network that processes news texts and outputs the probability of $y_t^j = 1$. When $d_2 < -1$, this is called a *prediction* because no news articles released later than t were used. This task has been shown to be very HARD according to the *efficient market hypothesis*. Experiment results also showed that a neural network can barely learn anything from price and text on this prediction task.

In this thesis, I set $d_1 = 4, d_2 = 0$, so that news articles after a price movement were also incorporated. That is, $\hat{y}_t^j \equiv f(N_{[t-4,t]})$. Note this task is NOT a *prediction*, but referred to by *text-driven price movement classification*. Even though this setting can not be directly used for price prediction, it is still meaningful because the knowledge learned from the data can be extracted and put to other uses.

4 Acquisition of Stock Embeddings

The learning of stock embeddings is based on the following idea. Intuitively, we want the inner product score between a stock embedding and the vector representation of a news text, i.e. $\text{score}_{i,j} = \vec{s}_j \cdot \vec{n}_i$, to produce a high value when the news is related to the vector. A market status vector with respect to stock j on day t, i.e. m_t^j, is then defined as a weighted average of the news vectors on day t:

$$\vec{m}_t^j = \frac{\sum_{\vec{n}_i \in N_t} \exp(\text{score}_{i,j})\vec{n}_i}{\sum_{\vec{n}_i \in N_t} \exp(\text{score}_{i,j})}$$

where the news vectors $\vec{n}_i$ were acquired via BERT [2], a SOTA text encoder. The market status vectors

within $[t-4, t]$ were inputted sequentially, noted by $M^j_{[t-4,t]}$, into a recurrent neural network with gated recurrent units (GRU) followed by a multi-layer perceptron (MLP):

$$\hat{y}^j_t = \text{MLP}(\text{GRU}(M^j_{[t-4,t]})).$$

In the experiment, the stock embeddings were learned on two news article datasets, the WSJ and R&B. By introducing stock embeddings, the classification accuracy of price movements were significantly enhanced, as in Table 1, which suggests that the stock embeddings had learned stock-specific knowledges.

Table 1. Mean classification accuracy (with standard deviations in parentheses) over 10 replications, in percentage (%).

Method for m^j_t	WSJ	R&B
previous work in [3]	52.1(2.51)	52.2(0.41)
previous work in [4]	54.3(0.99)	57.2(1.08)
proposed	**60.1**(0.91)	**68.8**(1.67)

5 Application of Stock Embeddings to Portfolio Optimization

A portfolio is essentially a set of investment proportions $w = [w_1, \ldots, w_J]^T$ assigned to J stocks. A good portfolio usually has a high gain and low uncertainty (risk). Given the prices of stock j at time t and $t-1$, the return of the stock can be computed by $r^t_j = p^t_j / p^{t-1}_j$. The portfolio's return is then $r_t = \sum_{j=1}^J w_j r^t_j$, with the variance

$$\text{Var}(r_t) = \sum_{j=1}^{J} \sum_{j'=1}^{J} w_j w_{j'} \text{cov}(r^t_j, r^t_{j'}).$$

Let $w = [w_1, \ldots, w_J]^T$ and $\Sigma = (\hat{\text{cov}}(r^t_j, r^t_{j'}))_{J \times J}$ where $\hat{\text{cov}}$ is the estimated covariance, then we can construct a portfolio w with expected portfolio return E and minimized variance (risk) by solving the problem below:

$$\begin{aligned}
\min_{w} \quad & \text{Var}(r_t) = w^T \Sigma w && \text{(2a)} \\
\text{subject to} \quad & w^T \hat{r} = E, && \text{(2b)} \\
& w^T \mathbf{1} = 1, && \text{(2c)} \\
& 0 \le w_j \le 1 \quad j = 1, ..., J, && \text{(2d)}
\end{aligned}$$

where $\mathbf{1} = \big(1, 1, \ldots, 1\big)^T$ is a J-dimension column vector consisting of 1's. Note that higher E usually means higher risk born by the investor.

Since stock embeddings reflects the characteristics of stocks in response to news articles or the financial events behind, their cosine values can be used to measure the correlation among stocks. For each pair of stocks, the cosine of their stock embeddings is calculated, which forms a cosine matrix Σ^{cos}:

$$\Sigma^{cos}_{j,j'} := \cos(s_j, s_{j'})$$

By replacing the Σ by Σ^{cos}, we can get another set of investment proportions, which forms a different portfolio. Investment simulations were conducted on the two datasets, in which portfolios are generated yearly, and the averaged annual gains are evaluated. Figure 1 shows the averaged real annual return of the portfolios from different methods in the R&B dataset. Among them, the S&P 500 Index is a popular portfolio containing 505 stocks. The *covariance* method corresponds to Σ, the *Weighted BERT* corresponds to a pure text-driven method based on BERT [2], and the *stock embedding* corresponds to the Σ^{cos} which is proposed in this thesis. It is observed that the stock embedding-based method achieved a 2.8x more gains compared to the covariance method.

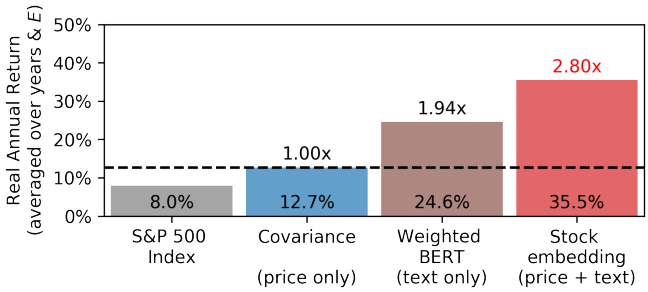

Fig. 1. Averaged real annual gains from different methods on R&B (2006-2013).

6 Conclusion

This thesis proposed a method to extract stock embeddings from both price history and news article texts. This is done by use of the task *text-driven price movement classification*. The higher accuracies achieved by stock embeddings demonstrated that the stock embeddings have learned stock-specific knowledges from the text and the price. Moreover, a further experiment of applying the stock embeddings to portfolio optimization strongly suggests the potentials of the stock embeddings in financial applications, including portfolio optimization.

Bibliography

[1] Xin Du and Kumiko Tanaka-Ishii. Stock embeddings acquired from news articles and price history, and an application to portfolio optimization. Submitted to ACL 2020.

[2] Devlin et al. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, pages 4171–4186, June 2019.

[3] Ding et al. Deep learning for event-driven stock prediction. In *IJCAI*, pages 2327–2333, 2015.

[4] Hu et al. Listening to chaotic whispers: A deep learning framework for news-oriented stock trend prediction. In *ACL*, pages 261–269, 2018.

[5] Lin et al. A structured self-attentive sentence embedding. In *ICLR*, 2017.

[6] Mikolov et al. Distributed representations of words and phrases and their compositionality. In *NeurIPS*, pages 3111–3119, 2013.

[7] Harry Markowitz. Portfolio selection. *The journal of finance*, 7(1):77–91, 1952.