Investment Behaviors Can Tell What Inside: Exploring Stock Intrinsic Properties for Stock Trend Prediction

Chi Chen
Tsinghua University
Beijing, China
chenchi14@mails.tsinghua.edu.cn

Li Zhao
Microsoft Research
Beijing, China
lizo@microsoft.com

Jiang Bian
Microsoft Research
Beijing, China
Jiang.Bian@microsoft.com

Chunxiao Xing
Tsinghua University
Beijing, China
xingcx@tsinghua.edu.cn

Tie-Yan Liu
Microsoft Research
Beijing, China
Tie-Yan.Liu@microsoft.com

ABSTRACT

Stock trend prediction, aiming at predicting future price trend of stocks, plays a key role in seeking maximized profit from the stock investment. Recent years have witnessed increasing efforts in applying machine learning techniques, especially deep learning, to pursue more promising stock prediction. While deep learning has given rise to significant improvement, human investors still retain the leading position due to their understanding on stock intrinsic properties, which can imply invaluable principles for stock prediction. In this paper, we propose to extract and explore stock intrinsic properties to enhance stock trend prediction. Fortunately, we discover that the repositories of investment behaviors within mutual fund portfolio data form up a gold mine to extract latent representations of stock properties, since such collective investment behaviors can reflect the professional fund managers' common beliefs on stock intrinsic properties. Powered by extracted stock properties, we further propose to model the dynamic market state and trend using stock representations so as to generate the dynamic correlation between the stock and the market, and then we aggregate such correlation with dynamic stock indicators to achieve more accurate stock prediction. Extensive experiments on real-world stock market data demonstrate the effectiveness of stock properties extracted from collective investment behaviors in the task of stock prediction.

CCS CONCEPTS

• **Information systems** → **Data mining**; • **Applied computing** → **Economics**.

KEYWORDS

Stock Prediction; Mutual Fund Portfolio Data; Matrix Factorization

KDD '19, August 4–8, 2019, Anchorage, AK, USA

ACM ISBN 978-1-4503-6201-6/19/08…$15.00
https://doi.org/10.1145/3292500.3330663

ACM Reference Format:
Chi Chen, Li Zhao, Jiang Bian, Chunxiao Xing, and Tie-Yan Liu. 2019. Investment Behaviors Can Tell What Inside: Exploring Stock Intrinsic Properties for Stock Trend Prediction. In *The 25th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '19), August 4–8, 2019, Anchorage, AK, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3292500.3330663

1 INTRODUCTION

Among a myriad of investment channels, the stock market has been continually considered as a big profitable potential. Generating a superior rate of return consistently over a further time horizon, nevertheless, requires a masterful understanding of the market and a definitive investment strategy. Stock prediction, with the aim at predicting future price trend of stocks, plays one of the key foundation techniques and has attracted increasing attention from brilliant minds [25, 27, 29]. Most of the traditional efforts on stock prediction rely on time-series analysis models, such as Autoregressive models [17], Kalman Filters [3], technical analysis [6], etc. In general, these solutions create dynamic stock indicators, based on stock prices and volumes, as stochastic inputs and take the historical data of indicators to fit the stochastic trends. However, such traditional solutions yield apparent drawbacks as they lack the capability to model dynamic validity of indicators, highly volatile market, and complex correlation between stocks and the market. With recent rapid development of deep learning, deep neural networks, especially recurrent neural networks (RNN), have been introduced as a promising substitute since its ability to model the sequential nature and non-linear structure within the stock prediction task [2, 10, 26, 31].

While deep neural networks have demonstrated a remarkable potential to boost stock prediction, human investors still retain the lead position, because they make the prediction on stock trend by considering intrinsic difference between stocks [4, 5, 7], which, however, is overlooked by most of existing deep learning approaches. In particular, experienced investors tend to distinguish a variety of stocks into several categories based on their intrinsic properties, based on which they could follow quite different prediction principles in various categories, respectively. For example, to invest a stock with the '*income*' property, meaning it has steady streams of revenue with low level of volatility, practiced human investors are inclined to hold the stock for a longer while. In contrast, a stock

with the '*cyclical*' property, indicating its price rises and falls with the business cycle, prompts skilled investors to time the market, i.e. buy the stock at the low point in the business cycle and sell it at the high point, in a more frequent way compared to investing income stocks.

Therefore, to pursue more accurate stock prediction, it is quite beneficial to take into account each stock's intrinsic properties. In this paper, we propose to integrate the stock's intrinsic properties into the existing deep neural networks approach with widely-used dynamic stock indicators. To this definitive end, however, it is requisite to answer two challenging questions: 1) how to extract and represent the stock intrinsic properties? 2) how to integrate the *static* intrinsic properties into the existing deep neural networks approach to enhance the *dynamic* stock prediction.

Due to the highly abstract nature of stock intrinsic properties. There is hardly a comprehensive and clear definition of such intrinsic properties in human investors' minds. Moreover, such intrinsic properties could cover a couple of diverse aspects of human investors' intuition and knowledge. Therefore, investment behaviors of human investors could be an indispensable information to recognize such intrinsic properties. Fortunately, mutual fund portfolio data, consisting of investment behaviors of expert investors, forms a gold mine for mining stock intrinsic properties.

In general, mutual funds are operated by professional fund managers, who allocate the fund's investments and attempt to produce capital gains for the fund's investors, and a mutual fund's portfolio is structured and maintained to match the investment objectives. Therefore, the history of a mutual fund's portfolios, consisting of collective investment behaviors, can reflect the fund manager's common beliefs on stock intrinsic properties. For instance, the portfolios of aggressive managers usually contain a large portion of stocks yielding higher rate of return together with a greater volatility; while those of conservative managers typically include many income stocks. These observations raise an important principle that stocks held by the same fund managers are more likely to share common intrinsic properties. Based on this principle, in this paper, we formulate the mutual fund portfolio data into a matrix of fund managers and stocks and propose to extract and represent stock intrinsic properties using a Matrix Factorization (MF) approach. The extracted vector for each stock can be viewed as its latent representation that reflect its intrinsic properties.

After obtaining the representations of stock intrinsic properties, the next question remains as how to integrate the static stock intrinsic properties into the existing deep neural networks approach. In other words, it is vital to design a proper structure of deep neural networks to effectively exploit static stock intrinsic properties into the task of stock prediction, which is dynamic in nature and yields inputs of widely-used dynamic stock indicators.

A straightforward method is to directly append the representations of stock properties with dynamic stock indicator as the unified input into the deep neural networks, especially RNN similar to [8, 31]. However, this method overlooks the incompatibility between the *static* stock properties and the *dynamic* prediction, which hinders the potential role of such properties in dynamic prediction. To address this challenge, in this paper, we propose to model the dynamic market state using stock representations so as to compute the dynamic correlation between the stock and the market state. Specifically, at a certain time point, the whole market can be described by the rank list of all stocks ordered by their recent profitable performance, based on which we can aggregate the representations of top ranked stocks to characterize the market state and take advantage of sequential modeling to generate the dynamic representations for the market state trend. After that, we can compute the dynamic correlation between the stock and the market state based on their respective representations, which enables us to leverage static stock properties to boost the dynamic stock prediction.

Table 1: An example of a mutual fund, named '*ChinaAMC Growth Fund*' with code 000001 in Chinese stock market, and its sampled half-year portfolios.

publish time	fund code	stock code	proportion
2013/06/30	000001	600887	3.15%
2013/06/30	000001	600837	3.09%
2013/06/30	000001	601633	2.19%
2013/06/30	000001	...	...
2013/06/30	...	...	...
2013/12/31	000001	600887	4.25%
2013/12/31	000001	601633	2.54%
2013/12/31	000001	002643	1.17%
2013/12/31	000001	...	...
2013/12/31	...	...	...
2014/06/30	000001	600887	3.04%
2014/06/30	000001	600270	2.28%
2014/06/30	000001	002279	2.24%
2014/06/30	000001	...	...
2014/06/30	...	...	...
...	...	...	...

[1] The specific names of funds and stocks can be found in https://www.msn.com/en-us/money

To validate the effectiveness of our approach, in this paper, we perform extensive experiments on real-world data which contains more than 2000 stocks in Chinese stock market from 2013 to 2016. The experiment results demonstrates the effectiveness of stock properties extracted from mutual fund portfolio data in the task of stock prediction. The further market trading simulation illustrates that our proposed approach can result in a significant increment of profit. To sum up, the contributions of this paper include:

- We mine the representations for stock intrinsic properties from mutual fund portfolio data, under the principle that stocks held by the same fund manager are likely to share common properties.
- We develop a novel deep learning framework to integrate *static* stock intrinsic properties into the *dynamic* stock prediction task by modeling dynamic market state/trend.
- We empirically demonstrate the effectiveness of extracted stock intrinsic properties and corresponding dynamic market state for stock trend prediction on real-world data.

2 MINING STOCK INTRINSIC PROPERTIES FROM INVESTMENT BEHAVIORS

2.1 Fund Manager Portfolios and Stock Properties

A mutual fund is an investment vehicle made up of a pool of money collected from many investors for the purpose of investing in various assets such as stocks. Typically, mutual funds are operated

by professional fund managers, who create the certain investing strategies and manage the portfolio so as to seek capital gains for the fund's investors. To demonstrate the fund's performance and allow investors to check the fund manager's investing strategies, mutual funds need to regularly (e.g. every half-year in Chinese stock market) release their portfolios to their investors. Table 1 illustrates an example of a mutual fund in Chinese stock market with part of its respective half-year portfolios. Every half year, all mutual funds must release their investment reports, containing totally hundreds of fund managers' investment behaviors over thousands of stocks within the latest half year.

Such regularly-published portfolio reports indeed comprise an invaluable repository for mining investment-related knowledge from a crowd of outstanding investing minds. Due to the professional understanding of the stock market and definitive investment strategies, investment behaviors of fund managers can indeed reflect their proficient investment knowledge. Furthermore, different fund managers usually lead to diverse investment preference on stocks, reflecting their separate attentions on various stock intrinsic properties. It inspires us to mine both stock properties and fund managers' preference from their investment behaviors. To the best of knowledge, this work introduces a very early attempt to explore the value of investment behaviors in mutual fund portfolio data. Meanwhile, this paper will focus on disclosing stock intrinsic properties from such data and leave the mining of its other values to future work.

2.2 Stock Intrinsic Properties Inside Investment Behaviors

With the definite goal of producing superior rate of return consistently, different fund managers, in the meantime, may create their respective portfolios consisting of stocks of quite diverse properties. Such divergence is mainly caused by that fact that each fund manager tends to embrace her own investment preference and specialization in terms of concerned intrinsic properties. For instance, fund managers, who possess magisterial knowledge economic cycles, would prefer to manage their portfolios more concentrated to cyclical stocks; while another group of fund managers, who seek steady streams of revenue with constrained volatility, are more likely to focus their portfolios on income stocks. Note that, the stock intrinsic properties could cover a variety of investment dimensions. For example, different fund managers may generate portfolios with explicit preference on certain industrial sectors, e.g. coal and oil, bank and finance, IT, etc., which implies that these fund managers hold respective expertise on these diverse domains.

While stock intrinsic properties play quite an important role in explaining and even predicting investment behaviors, it is uneasy to obtain such information. A straightforward method is to introduce human labeling, which however requires experts with domain knowledge and leads to very high cost. Moreover, stock intrinsic properties can be highly abstract with no clear definition in investors' minds, which makes it more than difficult to leverage human labeling.

Fortunately, inspired by observations about fund managers' different preferences on various stock intrinsic properties, we propose

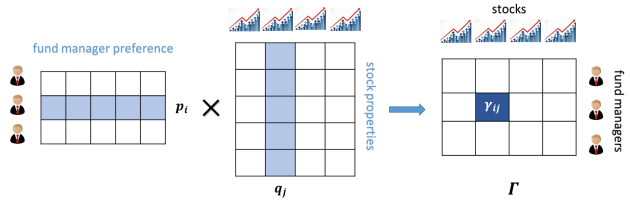


Figure 1: Fund manager i's overall preference on stock j, i.e., $\hat{\gamma}_{i,j}$, is approximated by aggregating respective preferences with respect to various latent property dimensions.

to learn latent and compact representations of stock intrinsic properties by mining the collection fund managers' investment behaviors in mutual fund portfolio data. In particular, based on aforementioned observations, it is expected that those stocks included in the portfolios by the same fund manager are more likely to share common intrinsic properties. Accordingly, we can formulate the mutual fund portfolio data into a matrix of fund managers and stocks and take advantage of a Matrix Factorization (MF) approach to extract a latent vector of each stock, viewed as the representation of the stock intrinsic properties.

It is worth mentioning that fund managers' investment behaviors do not only depend on their attention on certain stock intrinsic properties but also the dynamic stock trend. In other words, no fund manager is willing to invest a stock with an obvious declining trend even though it yields the certain properties attracting the fund manager. Furthermore, in practical investment, fund managers may invest other diverse stocks in order to reduce risk exposing to limited stocks. Therefore, besides the fund manager's inherent preference, the semi-annual fund manager portfolio is also determined by the stock dynamic trend as well as risk-averse diversity. On the other hand, as long as we observe the semi-annual mutual fund portfolio data within the period broad enough, the accumulated investment behaviors can magnify the long-term preference of fund managers as well as alleviate the effects of short-term trend dynamics or diversified investing for risk-reduction. In this way, it reassure us to disclose stock intrinsic properties by mining the collection of mutual fund portfolio data covering a broad time range.

只要我们在足够宽的时间段内观察半年度基金组合数据，累积的投资行为可以放大基金经理的长期偏好，并减轻短期趋势动态或分散投资降低风险的影响。通过这种方式，我们可以揭示股票的内在属性。

2.3 Learning Representations of Stock Intrinsic Properties by Matrix Factorization

As aforementioned, we propose to extract stock intrinsic properties by mining accumulated mutual fund portfolios due to its implication of fund managers' long-term preference on stock properties. Before stepping into the specific process, we make a more formal descriptions on accumulated mutual fund portfolios as well as stock intrinsic properties and corresponding fund managers' preference.

Let Γ represent the accumulated mutual fund investment portfolios, specified by each fund's average invested proportion on each stock over past several years in this paper. Then, $\gamma_{i,j}$ denotes the investment behavior of fund manager i on stock j, where $i = 1, 2, \cdots, M$, $j = 1, 2, \cdots, N$, and M, N denote the total number of fund managers and stocks, respectively. More formally,

$$\Gamma = \begin{pmatrix} \gamma_{1,1} & \gamma_{1,2} & \cdots & \gamma_{1,N} \\ \gamma_{2,1} & \gamma_{2,2} & \cdots & \gamma_{2,N} \\ \cdots & \cdots & \cdots & \cdots \\ \gamma_{M,1} & \gamma_{M,2} & \cdots & \gamma_{M,N} \end{pmatrix}$$

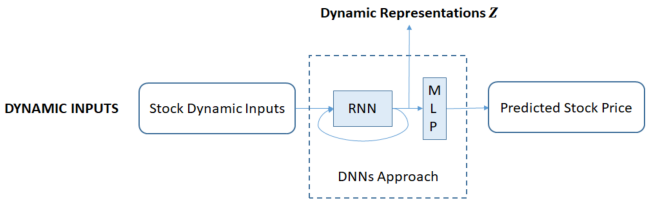


Figure 2: A Deep Neural Networks approach with dynamic inputs (Section 3.1.2). The representation Z contains the dynamic stock information that are different from the static properties.

Inspired by the principle that stocks invested by the same fund manager tend to share common intrinsic properties, we learn the representations of stock properties by modeling fund managers' preference on different stocks. Specifically, we associate a vector $q_j \in \mathbb{R}^K$ for each stock j and $p_i \in \mathbb{R}^K$ for each fund manager i, where q_j and p_i represent stock intrinsic properties and fund manager's preferences on stock properties, respectively, and K is the dimension size of both the stock representation and the fund manager's preference representation. Note that, we expect such vector representation, though in the form of latent vector, can imply stock properties covering multi-aspects, such as value, growth, cyclicity, volatility, section, etc. In this way, the investment behavior of fund manager i on stock j can be approximated by a similarity function of q_j and p_i in such low dimension space, which can be realized as

$$\hat{\gamma}_{i,j} = {p_i}^T q_j \tag{1}$$

where the inner product of q_j and p_i, as shown in Figure 1, denotes fund manager i's overall preference on stock j by aggregating respective preferences with respect to various latent property dimensions.

In this paper, we employ the matrix factorization approach to estimate the stock intrinsic properties q_j and fund manager's preference p_i. Matrix factorization has been wildly used in many scenarios, including recommender system [15], text mining [24], computer vision [28] .etc., as it can be applied to learn latent representation vectors of two types of entities underlying the interactions between each other. In our task, given the set of known investment behaviors γ, the parameters q_j and p_i, i.e. the latent representation of the stock and that of the fund manager, can be estimated through fitting the training data by solving the following optimization problem

$$\min_{p,q} \sum_{i,j} (\gamma_{i,j} - {p_i}^T q_j)^2 \tag{2}$$

In reality, there could exist prior investment bias of stocks as well as prior preference of fund managers. Hence, we introduce three respective bias items, μ_j^s, μ_i^f, and μ, to represent the bias of stock j, fund manager i, and the overall bias in the learning model, respectively. As a result, the investment behavior of fund manager i on stock j can be approximated by

$$\hat{\gamma}_{ij} = \mu + \mu_i^f + \mu_j^s + {p_i}^T q_j \tag{3}$$

In this way, all parameters, $\theta_{i,j} = \{ q_j, p_i, \mu_j^s, \mu_i^f \}$, can be estimated through fitting the training data by solving the following optimization problem:

$$\min_{\theta} \sum (\gamma_{i,j} - (\mu + \mu_i^f + \mu_j^s + {p_i}^T q_j))^2 + \lambda ||\theta|| \tag{4}$$

where $||\theta|| = (||p_i||^2 + ||q_j||^2 + {\mu_i^f}^2 + {\mu_j^s}^2)$ denotes the regularization term on introduced parameters to avoid over-fitting.

Note that stock properties learned by matrix factorization include two parts, i.e., q_j and μ_j^s, where the stock bias item μ_j^s represents the popular degree of stocks, and the stock vector q_j indicates latent stock intrinsic properties. We use $Q^j \in \mathbb{R}^{n+1}$ to denote the representation of stock properties of stock j in future sections.

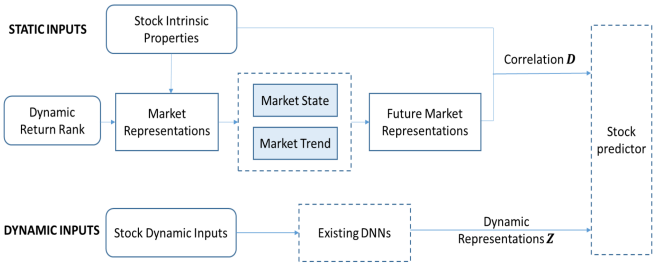


Figure 3: The framework of exploring *static* stock intrinsic properties for *dynamic* stock prediction, together with dynamic inputs (Section 3.2). The bottom part represents any existing deep neural networks approach with dynamic inputs, while the upper part depicts the main process of integrating static properties for the dynamic prediction.

3 EXPLORING STOCK INTRINSIC PROPERTIES FOR STOCK PREDICTION

3.1 Stock Prediction

3.1.1 The Problem Statement. With the objective to predict the future price trend, stock prediction can be formalized as a typical machine learning problem, either price movement classification or price return ratio regression, which is particularly to learn a prediction function $\hat{y}_{t+1}^i = f(X_{\leqslant t}^i)$ mapping a stock i from its feature space $X_{\leqslant t}^i$ to the target label $\hat{y}_{t+1}^i$.

While such typical problem setting merely treats different stocks as independent sequences, it neglects the relative comparison between stocks. Therefore, we re-formalize the stock prediction problem into learning a ranking function, $\hat{\mathbf{r}}_{t+1} = f(X_{\leqslant t})$, which projects a bunch of stocks into a ranking list in which stocks with higher ranking scores are expected to bring higher investment profits. To this end, we can use the return ratio of a stock, $r_{t+1}^i = (p_{t+1}^i - p_t^i)/p_t^i$, as the ground-truth, where p_t^i is the closing price of stock i at day t. In this way, $r_t^i < r_t^j$ means the return ratio of stock i is less than stock j at time t. To learn this model, we propose to combine both pointwise regression loss and pairwise ranking-aware loss[18] into a unified objective function:

$$\begin{aligned} l(\hat{\mathbf{r}}_t, \mathbf{r}_t) =& ||\hat{\mathbf{r}}_t - \mathbf{r}_t||^2 \\ &+ \alpha \sum_i \sum_j \frac{1}{2} (\max\{0, -(\hat{r}_t^i - \hat{r}_t^j)(r_t^i - r_t^j) + \tau\})^2 - \lambda \tau^2 \end{aligned} \tag{5}$$

where $\mathbf{r}_t$ and $\hat{\mathbf{r}}_t$ denote the ground-truth and predicted ranking list at time point t, within which r_t^i and $\hat{r}_t^i$ denote the ground-truth and predicted ranking scores of stock i, respectively; α is used to balance the two loss terms; and the regularization item τ is used to avoid obtaining an optimal constant ranking score.

3.1.2 Deep Neural Networks Approaches with Dynamic Stock Inputs. Considering the strong temporal dynamics of stock markets, it is intuitive to regard the historical status of a stock as the most influential factors to predict its future trend. Accordingly, most traditional methods feed dynamic inputs, such as daily price and various indicators [13], to time-series analysis models, such as Autoregressive models [17], Kalman Filters [3], technical analysis [6], etc. Recently, with rapid development of deep learning, deep neural networks (DNNs), especially recurrent neural networks (RNNs), have been applied and generate state-of-the-art performance on the stock prediction task [8, 31]. Figure 2 illustrates a generic DNN approach, especially with an RNN specification, for stock prediction. From this figure, we can abstract that, with no loss of generality, the DNN approach essentially first projects the dynamic inputs of each stock at time t, i.e., X_t^j, into a dynamic stock representation Z_t, and then conducts prediction based on such higher level representations.

3.2 Exploring Stock Intrinsic Properties

While deep learning has given rise to significant improvement, human investors still retain the leading position due to their understanding on stock intrinsic properties. Therefore, it is quite valuable to incorporate stock properties into the current stock prediction framework to pursue more accurate stock prediction. A straightforward way is to combine the representations of stock properties together with the dynamic representations, i.e., formally,

$$\hat{r}_{t+1}^j = MLP([Z_t^j, Q^j]) \tag{6}$$

where Z_t^j denotes the representation of the dynamic inputs of stock j, Q^j represents the representation of stock j's intrinsic properties, and $[\cdot,\cdot]$ means a direct concatenate. MLP represents the multilayer perceptron.

This straightforward integration, nevertheless, overlooks the incompatibility between the static stock properties and the dynamic stock prediction. In other words, the static stock properties cannot explicitly predict the stock trend in dynamic market. To address this problem, we propose to leverage stock properties in a more dynamic way.

Inspired by the fact that the market keeps its dynamics in terms of the changing inclination to various stock properties within different time period, we propose to model the representation for dynamic market and compute the correlation between the market representation and the stock representation. Given the market's dynamic preference, such correlation between stock properties and market dynamics can indeed provide more valuable information than merely the stock properties to enhance the stock prediction. Figure 3 summarize the whole framework of exploring static intrinsic properties for dynamic stock prediction, together with traditional dynamic inputs. Essentially, it is crucial to generate appropriate market representation to reflect the market's preference on various stock properties. In this paper, we propose two methods to obtain

Figure 4: The dynamic market representation is generated on a set of stocks with top-ranked return ratios at the certain day because the stocks with highest return ratios can reflect the market preference.

market representations by characterizing dynamic market state and dynamic market trend, respectively.

3.2.1 Dynamic Market State. Motivated by the intuition that the market representation should reflect the market's current preference on various stock properties, we propose to model the daily market representation based on a set of stocks with top-ranked return ratios at the certain day, considering that those stocks with highest return ratios can reflect the latest market preference. In particular, the market representation is computed by averaging the representations of those top-ranked stocks, considering that those stocks with highest return ratios can reflect the latest market preference. Figure 4 illustrate the general generation process of such market representation. More formally, we can compute the market state S_t at time t based on the representations of stocks whose return ratio is ranked within top-K_r at time t:

$$S_t = \frac{1}{K_r}\sum_{i=1}^{i=K_r} Q^i \qquad (r_t^{Q^1} > r_t^{Q^2} > \cdots > r_t^{Q^{K_r}} > \cdots) \tag{7}$$

where Q denotes the representation of stock intrinsic properties, and the index i of each stock is decided based on their respective return ratio rank at t.

Given this market representation, we can compute the correlation between the stock properties for each stock and current market state as:

$$D_t^j = S_t Q^j \tag{8}$$

Inspired by the assumption that the market preference is likely to be consistent in two consecutive days, we integrate the correlation $\hat{D}_t^j \approx D_{t-1}^j$ into the existing sequential model to predict the stock ranking at time $t+1$.

$$\hat{r}_{t+1}^j = MLP([Z_t^j, \hat{D}_t^j]) \tag{9}$$

where $[Z_t^j, \hat{D}_t^j]$ means a concatenate of stock dynamic representations and dynamic correlation between the stock j and the market state at time t.

While this approach can generate the representations of dynamic market state to enhance the stock prediction, the assumption that the market preference will be consistent in two consecutive days could not be always true and only using the last market state may still suffer from the suddenly high volatility in the market. To address this challenge, we propose to learn the sequential patterns of

historical market states and predict market trend representations accordingly in the next section.

3.2.2 Dynamic Market Trend. To address the limitation of the assumption that the market state will be consistent in two consecutive days, it is essential to model future market trend from historical market states rather than merely using the market state of the previous day for stock prediction. To this end, we propose to replace the representation of the last market state with the future market trend representation. With sequential market states for future state prediction, it is natural to take advantage of LSTM models, which have been widely used to conduct sequential prediction in many applications [2, 10, 26, 31]. LSTM is essentially a special type of Recurrent Neural Networks (RNNs) [19] that use hidden states (memory) to model sequential patterns of input data. A definite advantage of LSTM over vanilla RNN is its introducing of "forget" gates to store long-term memory and avoid vanishing gradients [11]. Then we make a formal description as,

$$\hat{S}_t = LSTM(S_{<t}) \tag{10}$$

where $\hat{S}_t$ represents the predicted market trend. In our method, the vector of the last hidden layer in the LSTM is used to represent the state of market trend. Similar to the method aforementioned in section 3.2.1, the predicted correlation $\hat{D}_t^j$ is computed based on the predicted market trend $\hat{S}_t$ and the stock properties Q^j, and then the combination of the correlation and the dynamic representation is feed to the MLP as Eq. 9.

Comparing two proposed methods for combining the market representation with the stock representation, the most essential difference lies in that one measures the correlation between the stock and the market based on the market's current state, while the other compute the correlation based on the market's future trend.

4 EXPERIMENTAL SETUP

4.1 Datasets

For the stock prediction model, we collect the Chinese stock data including time series of daily stock price and volume from 2012 to 2016 [1]. There are totally more than 2000 stocks, covering the vast majority of Chinese stocks. To further generate dynamic indicators, we follow the previous study [13] and compute totally 101 trading indicators. In order to effectively extract the stock intrinsic properties, we also collect the semi-annual Chinese mutual fund portfolio reports from 2012 to 2016 [2]. To guarantee the quality of stock properties, we keep the funds that always exist from their born to 2016. Table 2 shows the statistics in terms of the number of funds and stocks after filtering in the semi-annual mutual fund portfolio reports. When predicting stock trends, we filter out stocks under suspended trading status on more than 2% of trading days within the collection period, with concerns that their intermittent sequences may bring abnormal patterns. For those stocks never invested by any funds, zero vectors are regarded as their stock representations.

[1]We collect daily stock price and volume data from http://xueqiu.com/ and https://finance.yahoo.com/

[2]We collect mutual fund portfolios from https://www.morningstar.com/

Table 2: The statistics of mutual fund portfolios data.

time	#funds	#stocks	time	#funds	#stocks
2012/06/30	558	2035	2012/12/31	593	1980
2013/06/30	625	1981	2013/12/31	669	2004
2014/06/30	741	2172	2014/12/31	741	2350
2015/06/30	741	2612	2015/12/31	741	2638
2016/06/30	741	2688	2016/12/31	741	2832

4.2 Evaluation Metrics

In our experiments, we evaluate the ranking score of stock returns with information retrieval metrics including MAP and MRR. We introduce them as follows.

- **Mean Average of Precision(MAP)**: Average precision for each query is defined as the mean of the precision at n values calculated after each ranking list was retrieved. This metric measures the overall quality of ranking list. Formally, $MAP@K = \frac{1}{|S|}\sum_{s\in S}\frac{1}{K}\sum_{k=1}^{k=K} P(k)_s \cdot rel(k)_s$, where S denotes samples in test data, $P(k)_s$ is the precision measuring overall hits of the top-K results in the s-th sample. $rel()$ is a binary function on the relevance of a given rank.
- **Mean Reciprocal Rank (MRR)**: MRR evaluates the predicted rank of the top-return stock in ground-truth. It can be defined as $MRR = \frac{1}{|S|}\sum_{s\in S}\frac{1}{rank_s}$.

4.3 Compared Methods

The recent efforts on stock trend prediction focused on leveraging deep neural networks with dynamic inputs in terms of price and volume indicators. For example, LSTM based approach [22] and, most recently, SFM based approach [31] can obtain more competitive performance with the comparison of the other traditional methods. Therefore, we take advantage of LSTM and SFM as the RNN module, as shown in Figure 3, to model dynamic inputs. Note that, among these state-of-the-art stock prediction approaches, none of them, to our best knowledge, ever leveraged the intrinsic properties extracted from mutual fund portfolios. To evaluate the effectiveness of our models that explore such important information, we compare the following methods:

- **A LSTM with dynamic stock inputs (stock_LSTM)**: This model is proposed by [22] and operates on the sequential indicators. This method can be viewed as a specification of the method introduced in Section 3.1.2.
- **A SFM with dynamic stock inputs (stock_SFM)**: Zhang *et.al* proposed the SFM [31] and applied it in the stock prediction task. Compared to LSTM, SFM decomposes the hidden states of memory cells into multiple frequency components, each of which models a particular frequency of latent trading patterns underlying the fluctuation of stock price. This method can also be viewed as a specification of the method in Section 3.1.2.
- **Directly appending stock representation (DASR)**: Corresponding to the approach described at the beginning of Section 3.2, we combine the stock dynamic inputs and the stock intrinsic properties. The representations of stock properties extracted from the mutual fund portfolio data are directly appended on the last hidden state of SFM with dynamic inputs.
- **Integrating market state representations (IMSR)**: Corresponding to the approach described in Section 3.2.1, the correlation of

Table 3: Three examples of stock clusters generated based on stock representation extracted from mutual fund portfolios.

Cluster 1

Stock Code and Name	Sector
000878.SZ Yunnan Copper Co., Ltd.	Non-ferrous metal
000937.SZ Jizhong Energy Resources Co., Ltd.	Coal
000983.SZ Shanxi Xishan Coal and Electricity Power Co.,Ltd	Coal
002440.SZ Zhejiang Runtu Co., Ltd.	Chemical
002559.SZ Jiangsu Yawei Machine Tool Co., Ltd.	Machinery
600161.SH Beijing Tiantan Biological Products Co., Ltd.	Medical
600166.SH Beiqi Foton Motor Co., Ltd.	Car
600188.SH Yanzhou Coal Mining Company Limited	Coal
600348.SH Yangquan Coal Industry (Group) Co., Ltd.	Coal
600497.SH Yunnan Chihong Zinc & Germanium Co., Ltd.	Non-ferrous metal
600549.SH Xiamen Tungsten Co., Ltd.	Non-ferrous metal
600660.SH Fuyao Glass Industry Group Co., Ltd.	Car
600688.SH Sinopec Shanghai Petrochemical Company Limited	Petrochemical
601958.SH Jinduicheng Molybdenum Co., Ltd.	Non-ferrous metal

Cluster 2

Stock Code and Name	Sector
002079.SZ Suzhou Good-Ark Electronics Co., Ltd.	Electronics
002335.SZ Xiamen Kehua Hengsheng Co., Ltd.	Telecommunications
002436.SZ Shenzhen Fastprint Circuit Tech Co., Ltd.	Electronics
002618.SZ Shenzhen Danbond Technology Co.,Ltd.	Electronics
300010.SZ Beijing Lanxum Technology Co., Ltd.	Electronics
300219.SZ Hongli Zhihui Group Co., Ltd.	Electronics
600679.SH Shanghai Phoenix Enterprise (Group) Co., Ltd.	Light manufacturing
600749.SH Tibet Tourism Co.,Ltd.	Tourism
600892.SH Dasheng Times Cultural Investment Co., Ltd.	Cultural
600978.SH Yihua Lifestyle Technology Co., Ltd.	Light manufacturing

Cluster 3

Stock Code and Name	Sector
002033.SZ LiJiang YuLong Tourism Co., LTD.	Tourism
002035.SZ Vatti Corporation Limited	Electronics
002157.SZ Jiangxi Zhengbang Technology Co.Ltd.	Electronics
002234.SZ Animal Husbandry Co., Ltd.	Livestock
002441.SZ Zhongyeda Electric Co., Ltd.	Electronics
002458.SZ Livestock & Poultry Breeding Co., Ltd.	Livestock
002714.SZ Muyuan Foods Co.,Ltd.	Agriculture
600371.SH WanXiang Doneed Co., Ltd.	Agriculture
600593.SH Tourism Holding CO.,LTD	Tourism

the stock and the market state at the last time step is integrated into the last hidden state of SFM with dynamic inputs.

- **Integrating market trend representations (IMTR)**: Corresponding to the approach described in Section 3.2.2, this method models the market trend by LSTM with the historical market states, rather than merely using the market state of the previous day in IMSR. Then the correlation of the stock and the predicted market state trend is computed and integrated into the last hidden state of SFM.

由于中国共同基金投资组合报告每半年发布一次，因此我们每半年更新一次股票表示。（股票特征）

4.4 Parameter Setting

Because Chinese mutual fund portfolio reports are published every half year, we update stock representations in the frequency of every six-months. For stock trend prediction, we model and predict the stock trend every half year using the last updated stock representations. In the remaining part, we simply represent the first and the second half of a certain *year* as *year*(1) and *year*(2), respectively. In each six-month dataset, we split it into training set (3 months), validation set (1 month) and test set (2 months). When extracting stock intrinsic properties, we use the investment for three consecutive reports as the accumulated investment behaviors γ.

In this paper, we employ grid search to select the optimal hyper-parameters that can result in maximized MAP@50 on the validation sets for all methods. Specifically, for the LSTM and SFM parts in models, we search the number of LSTM/SFM cell within {5, 10, 20}; for learning stock intrinsic properties, we tune the size of stock representations within {32, 64, 128, 256}; for constructing the market state, we tune the number of top-return stocks K_r in Equation 7 within {1, 5, 10, 15, 20, 25, 30}. In addition, we further tune α which balances the pointwise and pairwise terms within {0.1, 1, 10, 100}, and τ which avoid an optimal constant ranking score in loss function within {0.1, 0.5, 1} in Equation 5. We also tune the λ of the regularization term in Equation 4 and 5 within {0.1, 1, 10}.

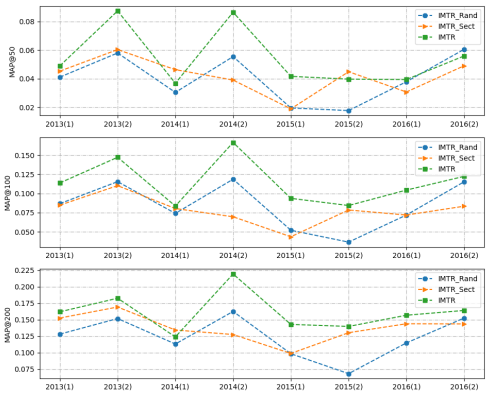


Figure 5: Model performance comparison on MAP@50, 100 and 200 against varying types of stock representations.

5 EXPERIMENT RESULTS

定性分析，以评估学习到的股票表示是否能够捕捉到内在属性

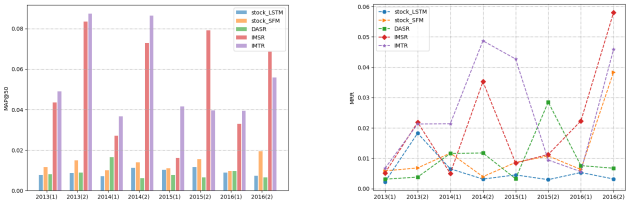
5.1 Learned Stock Representations

To examine the quality of stock representations learned from mutual fund portfolios, we take some qualitative analysis to assess if the learned stock representations can capture the intrinsic properties. In particular, we cluster all the stocks based on their respective learned representations by the Affinity propagation [9]. Table 3 shows three examples of obtained stock clusters in the second half year of 2015. From this table, we can find that all the stocks in the first cluster belong to the basic industry while those in the second cluster are much related to the light industry. Furthermore, most of the stocks in Livestock and Agriculture industries are clustered together into the third cluster. Such clustering results can clearly indicate that the stock representations extracted from mutual fund portfolios can carry certain intrinsic properties.

5.2 Effects of Learned Stock Representations

To demonstrate the effects of learned stock representations for stock prediction, we investigate the performance of IMTR by replacing the learned stock representations with random representations or pre-defined sector-based one-hot representations.

Figure 5 depicts the comparison performance of MAP@50, 100 and 200 on models with different stock representations. Specifically, IMTR_Rand assigns a random representation vector for each stock; as each stock belongs to one sector, IMTR_Sect uses a sector based one-hot vector to be the stock representation; and, IMTR employs the stock representations learned from mutual fund portfolios. Although IMTR_Rand does not have any meaningful properties in stock representations, the different representations distinguish stocks. From this figure, we can find that IMTR outperforms the other two. The better performance of IMTR over IMTR_Sect implies



(a) (b)

Figure 6: Performance comparison on MAP@50 and MRR among stock_LSTM, stock_SFM, DASR, IMSR and IMTR.

that the learned stock representations can contain more intrinsic properties than merely sectors. Furthermore, we can find that the stock representations in IMTR have better effect on the prediction with the incremental number of selected stocks compared to the other two methods, which indicates that more significant improvements are received on MAP@100 and MAP@200 than MAP@50.

5.3 Overall Performance

Figure 6a and 6b present a comprehensive analysis on the effectiveness of all compared methods in terms of MAP@50 and MRR for the stock prediction task. From the figures, we have following observations:

Firstly, IMSR and IMTR obviously outperform stock_LSTM and stock_SFM where only historical indicators of each stock are used. It indicates that stock intrinsic properties are useful and play an important role in stock prediction. Especially when they are used in a properly dynamic way.

Secondly, DASR has similar, even lower MAP@50 in some datasets, compared with stock_SFM. Although stock properties have been considered in DASR, the incompatibility between the static stock properties and the dynamic nature of both stocks and the market, has a bad effect on stock prediction. Compared with IMSR and IMTR which also take into account stock properties, lower performance of DASR implies the effectiveness of the market state which makes the static stock properties dynamic. On the other hand, higher MRR on DASR especially in 2015(2), when the stock trend is hard to be predicted, indicates that stock intrinsic properties play a more important role on highest-return stock selection than the dynamic prices and indicators because they are relatively static facing the stock market turbulence.

Thirdly, IMTR has better performance than IMSR in most datasets because IMTR predicts the future market state while IMSR only considers the market state at the last time step as the future one. However, when the market falls into a hardly unpredictable state, for example the Chinese stock market turbulence which began with the popping of the stock market bubble on 12 June 2015 and ended in early February 2016, IMTR performs worse than IMSR because the patterns in the sequential market states are difficult to be recognized. It results in a bad prediction of market trend.

In a summary, all these results can indicate that, the learned stock intrinsic properties can help investors make more accurate stock selection; moreover, it is necessary to model dynamic market state based on relatively static stock properties for more accurate stock prediction.

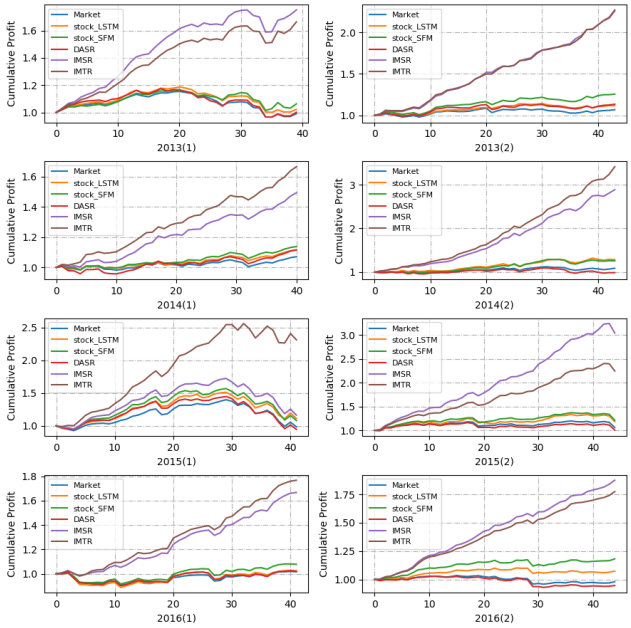


Figure 7: The cumulative profit curve of different methods with the portfolio of choosing top 50 stocks.

5.4 Market Trading Simulation

To further evaluate the effectiveness of our proposed models, we conduct back-testings by simulating the stock trading for each dataset. Our estimation strategy conducts the trading in the daily frequency. Given a certain principal at the beginning of each back-testing, investors invest in the top-K stocks in average with the highest predicted ranking score in each day. The selected stocks are hold for one day. The cumulative profit without consideration of transaction cost will be invested into the next trading day. We also calculate the average return on the stock market by evenly holding every stock as the baseline, indicating the overall market trend. Figure 7 shows the cumulative profit curve for each method with K as 50. In each back-testing, our models, IMSR and IMTR, are able to gain the best profit results among all the compared methods, even when the market falls into downturn such as the first half year of 2015.

6 RELATED WORK

Recent work on stock prediction relies on two kinds of information sources: indicators from stock price/volume data and text from news and social medias.

Technical indicators [6], i.e., mathematical calculations based on historical price, volume, or (in the case of futures contracts) open interest information [20], are proposed by financial experts at first for the purpose of discovering trading patterns of dynamic indicators. One of the most widely used models in the pattern recognition is Autoregressive (AR) model for linear and stationary time-series [17]. However, the non-linear and non-stationary nature of stock prices limits the application of AR models. Hence, substantial studies attempted to apply non-linear models to capture the complex dynamic market trend. With the development of deep learning, more scientists make efforts to exploit deep neural network for financial

prediction [1, 12, 14, 16, 23, 30]. To further model the long-term dependency in time series, recurrent neural networks (RNNs), especially Long Short-Term Memory (LSTM) network, have also been employed in stock prediction [2, 10, 26]. In most recent time, a new RNN named the State Frequency Memory (SFM) is proposed by Zhang *et.al* [31] to discover the multi-frequency trading patterns.

In order to seek more information from the outside market-historic-data, text from public news and social media is analyzed and explored for stock prediction. Public news of a company indicates the stock outlook and trend in advance which can be used in stock prediction. Nassirtoussi *et al.* [21] use the semantics and sentiment based on the text of breaking financial news-headlines and propose a multi-layer dimension reduction algorithm to predict intra-day directional-movements of a currency-pair in the foreign exchange market. In addition, investors' sentiments for stocks' trends can also help the prediction. Zhou *et al.* [32] studies particularly the Chinese stock market. They conduct a thorough study over 10 million stock-relevant tweets from Weibo, and find five attributes that stock market in China can be competently predicted by various online emotions. Nguyen *et al.* [29]explicitly consider the topics relating to the target stocks, and extracting topics and related sentiments from social media to make the prediction. However, textual analysis is a challenging task. It is hard to work out complete and accurate sentiments from the text.

Despite increasing efforts in stock prediction, few of them paid enough attention to stock intrinsic properties which are usually considered by expert investors. Inspired by this, this paper proposes a method to extract stock intrinsic properties and develops models to integrate them into existing deep neural networks approach.

7 CONCLUSION

In this paper, we propose to take into account stock intrinsic properties in stock prediction tasks in order to enhance existing model based on dynamic inputs. There are three contributions in our paper: firstly, we are the first to leverage stock intrinsic properties to help investors make stock selections. Secondly, we propose to extract stock intrinsic properties from mutual fund portfolios. Thirdly, we develop a novel model to use static stock properties in a dynamic way by measuring the correlation between the market and the stock. In the future, we plan to seek stock intrinsic properties from other valuable data and extend market state model in a dedicated way. Furthermore, we will explore more useful investment behaviors of fund managers to improve stock prediction models.

ACKNOWLEDGMENTS

Chi Chen and Chunxiao Xing are supported by NSFC 91646202 and National Key R&D Program of China SQ2018YFB140235.

REFERENCES

[1] Ayodele Ariyo Adebiyi, Aderemi Oluyinka Adewumi, and Charles Korede Ayo. 2014. Comparison of ARIMA and artificial neural networks models for stock price prediction. *Journal of Applied Mathematics* 2014 (2014).

[2] Ryo Akita, Akira Yoshihara, Takashi Matsubara, and Kuniaki Uehara. 2016. Deep learning for stock prediction using numerical and textual information. In *Computer and Information Science (ICIS), 2016 IEEE/ACIS 15th International Conference on*. IEEE, 1–6.

[3] Ranjeeta Bisoi and Pradipta K Dash. 2014. A hybrid evolutionary dynamic neural network for stock market trend analysis and prediction using unscented Kalman filter. *Applied Soft Computing* 19 (2014), 41–56.

[4] Louis KC Chan, Hsiu-Lang Chen, and Josef Lakonishok. 2002. On mutual fund investment styles. *The Review of Financial Studies* 15, 5 (2002), 1407–1437.

[5] Prasanna Chandra. 2017. *Investment analysis and portfolio management*. McGraw-Hill Education.

[6] Robert D Edwards, John Magee, and WH Charles Bassetti. 2007. *Technical analysis of stock trends*. CRC press.

[7] Eric G Falkenstein. 1996. Preferences for stock characteristics as revealed by mutual fund portfolio holdings. *The Journal of Finance* 51, 1 (1996), 111–135.

[8] Thomas Fischer and Christopher Krauss. 2018. Deep learning with long short-term memory networks for financial market predictions. *European Journal of Operational Research* 270, 2 (2018), 654–669.

[9] Brendan J Frey and Delbert Dueck. 2007. Clustering by passing messages between data points. *science* 315, 5814 (2007), 972–976.

[10] Qiyuan Gao. 2016. *Stock market forecasting using recurrent neural network*. Ph.D. Dissertation. University of Missouri–Columbia.

[11] Felix A Gers, Nicol N Schraudolph, and Jürgen Schmidhuber. 2002. Learning precise timing with LSTM recurrent networks. *Journal of machine learning research* 3, Aug (2002), 115–143.

[12] Mustafa Göçken, Mehmet Özçalıcı, Aslı Boru, and Ayşe Tuğba Dosdoğru. 2016. Integrating metaheuristics and artificial neural networks for improved stock price prediction. *Expert Systems with Applications* 44 (2016), 320–331.

[13] Zura Kakushadze. 2016. 101 formulaic alphas. *Wilmott* 2016, 84 (2016), 72–81.

[14] Kyoung-Jae Kim and Hyunchul Ahn. 2012. Simultaneous optimization of artificial neural networks for financial forecasting. *Applied Intelligence* 36, 4 (2012), 887–898.

[15] Yehuda Koren, Robert Bell, and Chris Volinsky. 2009. Matrix factorization techniques for recommender systems. *Computer* 42, 8 (2009).

[16] Leonel A Laboissiere, Ricardo AS Fernandes, and Guilherme G Lage. 2015. Maximum and minimum stock price forecasting of Brazilian power distribution companies based on artificial neural networks. *Applied Soft Computing* 35 (2015), 66–74.

[17] Lili Li, Shan Leng, Jun Yang, and Mei Yu. 2016. Stock Market Autoregressive Dynamics: A Multinational Comparative Study with Quantile Regression. *Mathematical Problems in Engineering* 2016 (2016).

[18] Tie-Yan Liu et al. 2009. Learning to rank for information retrieval. *Foundations and Trends® in Information Retrieval* 3, 3 (2009), 225–331.

[19] LR Medsker and LC Jain. 2001. Recurrent neural networks. *Design and Applications* 5 (2001).

[20] John J Murphy. 1999. *Technical analysis of the financial markets: A comprehensive guide to trading methods and applications*. Penguin.

[21] Arman Khadjeh Nassirtoussi, Saeed Aghabozorgi, Teh Ying Wah, and David Chek Ling Ngo. 2015. Text mining of news-headlines for FOREX market prediction: A Multi-layer Dimension Reduction Algorithm with semantics and sentiment. *Expert Systems with Applications* 42, 1 (2015), 306–324.

[22] David MQ Nelson, Adriano CM Pereira, and Renato A de Oliveira. 2017. Stock market's price movement prediction with LSTM neural networks. In *Neural Networks (IJCNN), 2017 International Joint Conference on*. IEEE, 1419–1426.

[23] Jigar Patel, Sahil Shah, Priyank Thakkar, and K Kotecha. 2015. Predicting stock market index using fusion of machine learning techniques. *Expert Systems with Applications* 42, 4 (2015), 2162–2172.

[24] V Paul Pauca, Farial Shahnaz, Michael W Berry, and Robert J Plemmons. 2004. Text mining using non-negative matrix factorizations. In *Proceedings of the 2004 SIAM International Conference on Data Mining*. SIAM, 452–456.

[25] G Preethi and B Santhi. 2012. STOCK MARKET FORECASTING TECHNIQUES: A SURVEY. *Journal of Theoretical & Applied Information Technology* 46, 1 (2012).

[26] Akhter Mohiuddin Rather, Arun Agarwal, and VN Sastry. 2015. Recurrent neural network and a hybrid model for prediction of stock returns. *Expert Systems with Applications* 42, 6 (2015), 3234–3241.

[27] Eberhard Schöneburg. 1990. Stock price prediction using neural networks: A project report. *Neurocomputing* 2, 1 (1990), 17–27.

[28] Amnon Shashua and Tamir Hazan. 2005. Non-negative tensor factorization with applications to statistics and computer vision. In *Proceedings of the 22nd international conference on Machine learning*. ACM, 792–799.

[29] Jianfeng Si, Arjun Mukherjee, Bing Liu, Qing Li, Huayi Li, and Xiaotie Deng. 2013. Exploiting topic based twitter sentiment for stock prediction. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Vol. 2. 24–29.

[30] Jonathan L Ticknor. 2013. A Bayesian regularized artificial neural network for stock market forecasting. *Expert Systems with Applications* 40, 14 (2013), 5501–5506.

[31] Liheng Zhang, Charu Aggarwal, and Guo-Jun Qi. 2017. Stock Price Prediction via Discovering Multi-Frequency Trading Patterns. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2141–2149.

[32] Zhenkun Zhou, Jichang Zhao, and Ke Xu. 2016. Can online emotions predict the stock market in china?. In *International Conference on Web Information Systems Engineering*. Springer, 328–342.