# Personalized Dynamic Recommender System for Investors

Takehiro Takayanagi
m2021ttakayanagi@socsim.org
The University of Tokyo
Tokyo, Japan

Chung-Chi Chen
c.c.chen@acm.org
National Institute of Advanced
Industrial Science and Technology
Tokyo, Japan

Kiyoshi Izumi
izumi@sys.t.u-tokyo.ac.jp
The University of Tokyo
Tokyo, Japan

## ABSTRACT

With the development of online platforms, people can share and obtain opinions quickly. It also makes individuals' preferences change dynamically and rapidly because they may change their minds when getting convincing opinions from other users. Unlike representative areas of recommendation research such as e-commerce platforms where items' features are fixed, in investment scenarios financial instruments' features such as stock price, also change dynamically over time. To capture these dynamic features and provide a better-personalized recommendation for amateur investors, this study proposes a Personalized Dynamic Recommender System for Investors, PDRSI. The proposed PDRSI considers two investor's personal features: dynamic preferences and historical interests, and two temporal environmental properties: recent discussions on the social media platform and the latest market information. The experimental results support the usefulness of the proposed PDRSI, and the ablation studies show the effect of each module. For reproduction, we follow Twitter's developer policy to share our dataset for future work.

## CCS CONCEPTS

• **Information systems → Recommender systems**; **Information systems applications**.

## KEYWORDS

stock recommendation; investor modeling; financial data mining

## 1 INTRODUCTION

Dynamic recommender system aims to capture the change in user, item, or environment spheres, and provide recommendations based on the latest features [14]. Although some previous studies considered the change of user's preference when recommending emoji [22] and news articles [12, 13], few of them dynamically track the change

of all spheres, i.e., user, item, and environment. It is reasonable because the application scenarios explored in most previous studies only have one or two dynamic features. For example, users' interests and preferences may change on e-commerce platforms but items' information, such as Pixel 7 Pro, is fixed. However, we notice that a dynamic recommender system for investors should consider the change in the investor's interests (user), the latest stock information (item), and the public discussions of the stock (environment). That is, in investment decision makings, individual investors consider their portfolios based on numerous factors such as their interests, financial market information, technical indicators of an individual stock, and discussions on social media. As a result, it takes quite an amount of time for an investor to properly choose a portfolio that suits their interest given the fact that there are several thousand stocks with their context dynamically changing in the financial market. The financial stock recommender system that recommends suitable stocks for an investor is crucial as it helps investors discover investment opportunities among thousands of traded stocks in the dynamically changing market. Following this line of thought, we propose a new recommendation task, which aims to predict the next stock an investor will be interested in on social media platforms. In the proposed task, we attempt to predict the stock that the investor will mention based on given previous tweets of the same investor, the latest stock information, and other investors' recent discussions, corresponding to the dynamic user, item and environment aspects in the financial stock marke.

The stock recommender systems in the previous studies can be separated into two groups, profitability-oriented [6, 10, 19] and personalization-oriented [16, 20]. The former has a significant amount of research, and it is also related to stock price movement prediction tasks. In contrast, the latter attracts less interest and has a crucial problem: lack of benchmark dataset for investors' preferences [23]. To fill this gap, we propose a D3 dataset, which contains **D**ynamic investors' interests, **D**ynamic financial market information, and **D**ynamic social media discussions. The proposed D3 dataset contains over 180K tweets related to constituents of the S&P 500 Index, and will be published based on Twitter's developer policy for further explorations.

In order to address the proposed task, we design a Personalized Dynamic Recommender System for Investors, PDRSI, as shown in Figure 1. The proposed PDRSI models the dynamic user preference with the stock preference and personal interest modules. At the same time, it incorporates dynamic market information with the discussion and market modules.

## 2 RELATED WORK

There is a growing demand for recommender systems for investors as the number of retail investors using online brokers has rapidly
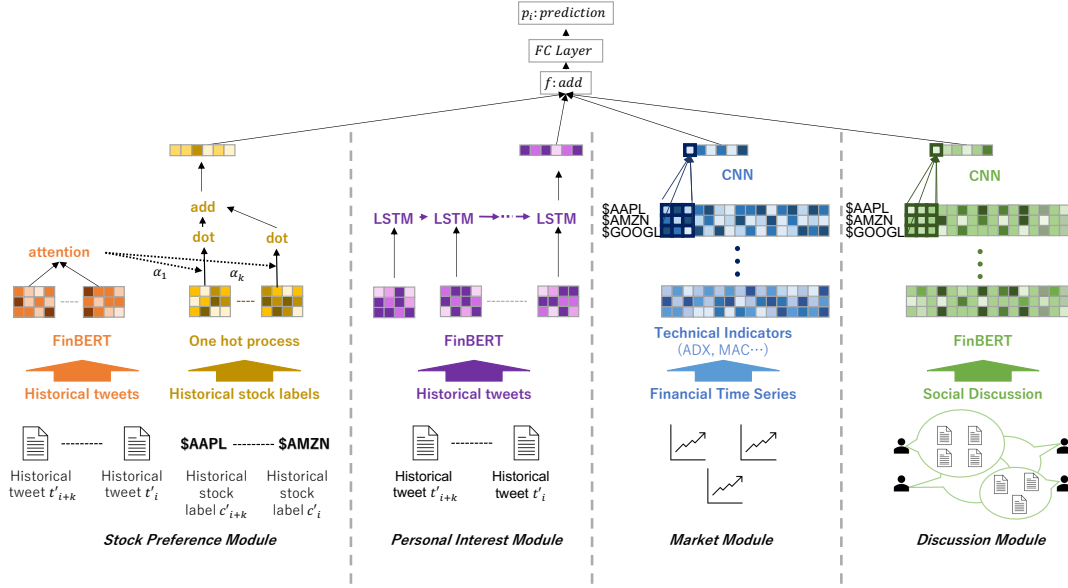
**Figure 1: Overview of the proposed PDRSI.**

increased. Accordingly, many studies have tackled stock recommendation tasks. The approach of stock recommendations can be classified into two approaches: non-personalized stock recommendations and personalized stock recommendations. Most works in stock recommendation fall within the scope of non-personalized stock recommendation which focuses on identifying optimal strategies for selecting stocks or portfolios that are likely to be more profitable in the future [18]. On the other hand, collaborative filtering has been used for personalized stock recommendation, oftentimes combined with other recommendation approaches such as order book analysis, and multiple criteria decision analysis [15, 17, 21]. For instance, Swezey and Charron [15] estimates the investor's risk tolerance from users' portfolios and recommends stock based on the relevance of the stock's risk-return with the user's risk tolerance combined with a collaborative filtering method. These personalized stock recommendation studies are different from this work as they deal with transaction data from a private company and they personalize the recommendation with investors' risk tolerance. While some studies have studies the investors' personalized preferences in stock recommendations [3–5], the scarcity of open data has led to a significantly less comprehensive body of research in comparison to non-personalized stock recommendations, as argued in [15]. To fill this gap, we propose a D3 dataset for exploring personalized stock recommendations.

With abundant open data, the research on the sequential recommendation in areas such as emoji and news developed various methods to incorporate the dynamic preference of users into recommendations [12, 13, 22]. For example, Zheng et al. [22] developed PERD (personalized emoji recommendation with dynamic user preference). PERD proposed a novel personalized attention module to learn dynamic user preference representations from text and emojis in historical tweets and achieved the highest performance in the emoji recommendation task. However, different from emoji

prediction tasks or news recommendation tasks, the property of the item changes dynamically in the financial market. To deal with all dynamically-changing properties simultaneously, this study proposes a new framework, PDRSI, and obtains a good performance in the personalized dynamic recommendation task.

## 3 DATASET

### 3.1 Problem Definition

We define user set $U = [u_1, u_2, ..., u_g]$ with size g, tweets data $T = [t_1, t_2, ..., t_n]$ with size n , corresponding stock in the tweet $C = [c_1, c_2, ..., c_n]$, and stock set $S = [s_1, s_2, ..., s_m]$ with size m. The i-th tweet $t_i = [w_1, w_2, ..., w_o]$ is composed of a sequence of words. Based on the above definitions, our stock recommendation task can be formulated as a multi-class classification. Our task is to recommend the next stock that investors are interested in from the candidate stocks $S$. The ground truth of predictions can be denoted as $Y = [y_1, y_2, ..., y_n]$. $y_i = \{0, 1\}^m$ represents whether a stock is used in the user's next tweet. $y_{i,j} = 1$ means that the j-th stock $s_j$ is included in the user's next tweet.

### 3.2 Investor Behavior Analysis

In order to prepare a dataset suitable for the proposed recommendation task, we retrieved 10 million English tweets containing cashtags of S&P 500 companies, where a cashtag is a company ticker symbol preceded by a dollar sign (e.g., $TWTR for Twitter, Inc.'s stock, and $AAPL for Apple Inc.'s stock). Some examples of the retrieved tweets are provided in Table 1. Although T1 is a targeted tweet in our experiment because an investor shows interest in an individual stock, not all tweets, such as T2 and T3, are similar to T1. Therefore, we filtered the data in two ways to ensure that the investors in the dataset are appropriate for our experiment. The first method filters out users who have a tweet that includes more

**Table 1: Examples for behavior analysis.**

| | |
|---|---|
| T1 | $MSFT earnings were ok. 11% rev growth (16% in cc) not bad, and runway is long. Azure +42% cc ($52bn run rate!) |
| T2 | These 7 stocks make up 48.6% of $QQQ and 24.7¥% of $SPY. It's going to be hard for markets to rally if they're all losing momentum (except for $AAPL) and only 3 are gaining relative strength ($AAPL, $TSLA & $AMZN). $MSFT and $GOOGL are losing both and $NVDA/ $META losing momentum |
| T3 | $AAPL Trading Ideas \| Awaiting Buy signal. 83.33% Profitability based on 6 trades. Profit factor is 3.78 |

than two cashtags because such users often discuss macro market information rather than the individual stocks of interest as in T2. The second method extracts users who have tweeted more than 10 stocks and less than 100 stocks during the sampling period because most investors who tweeted more than 100 stocks are likely to be auto-generated bots. For instance, T3 contains only one cashtag in the tweet itself, but the user tweets similar content for more than 500 stocks with the same format, indicating that the user is probably an auto-generated bot. Additionally, we selected users with more than 10 stocks because we wanted to observe changes in investors' interest and ensure that they had an interest in multiple stocks in the past.

### 3.3 Dataset Statistics

In the proposed dataset, D3, there are 184,370 tweets posed by 2,168 users, and the sampling period is from 1st August 2022 to 30th November 2022. The average number of stocks a user mentioned during the sampling period is 21. As we mentioned, we target the constituents of the S&P 500 Index, and thus there are 502 stocks mentioned in the collected tweets. Some notable stocks attract investors' interest such as Tesla, Meta, and Apple which amount to 10,694, 8,261, and 7,322 tweets, respectively. Following the Twitter developer guideline, we release our dataset in the following ways. We release the TweetIDs and UserIDs. We also share the query of the way how we collect the dataset using Twitter API, and release codes of data preprocessing. [1]

### 4 METHOD

Figure 1 illustrates the proposed PDRSI. We provide the details of each module in the following sections.

### 4.1 Stock Preference Module

We capture the users' dynamic preferences on an individual stock with the self-attention mechanism. Different from the emoji prediction task where the task is to predict the suitable emoji for the current tweet [22], our task is to predict the next stock a user is interested in based on the historical tweets. Therefore, we use all the stock labels from the historical tweet as input for the model, while the latest historical information is not utilized in Zheng et al. [22].

$$q = W_q h_i + b_q \tag{1}$$

[1]https://github.com/TTsamurai/PDRSI_public_code

$$K = W_k H' + b_k \tag{2}$$

$$V = W_v C' + b_v \tag{3}$$

$$a_j = k_j^T q \tag{4}$$

$$\alpha_j = \frac{exp(a_j)}{\sum_{j=1}^{k} exp(a_j)} \tag{5}$$

$$a_i = \sum_{j=1}^{k} \alpha_j v_j^T \tag{6}$$

where $q$ is a query vector, $K$ is a key matrix, $V$ is a value matrix, $W_q$, $W_k$, $W_v$ are corresponding weights and $b_q$, $b_k$, $b_v$ are corresponding bias. $\alpha_j$ is the attention weight of j-th stock label. $k_j$ and $v_j$ are j-th row if $K$ and $V$.

### 4.2 Personal Interest Module

We propose a personal interest module to account for a temporal change in personal interest expressed in the tweets such as the trading strategy or the reasons they mentioned the stock. While the investors' interest is directly shown in the stock that the investors mentioned in their tweets, their interest is also implied in their text. For each user, we select their target tweet $t_i$ and their $k$ recent historical tweets to form a historical tweet list $T' = [t_i', t_{i+1}', ..., t_{i+k}']$ . The stock label $C'$ corresponding to $T'$ can be written as $C' = [c_i', c_{i+1}', ..., c_{i+k}']$. The tweet representations $H'$ of $T'$ are $[BERT(t_i'), BERT(t_{i+1}'), ...BERT(t_{i+k}')] = [h_i', h_{i+1}', ..., h_{i+k}']$ , where $h_i'$ is a tweet representation of $t_i'$. We use Long Short Term Memory (LSTM) to extract a temporal feature vector, $e_i$.

### 4.3 Market Module

We propose a market module to capture the dynamic market information. Due to the dynamic nature of the financial market, investors' interest is likely to be influenced by the recent price movement. Therefore, recognizing recent market information is useful in predicting investors' interest. In the market module, we first calculate the representative technical indicators for each stock as follows.

$$[p_{\tau_i,s_j} \| p_{\tau_i-1,s_j}, ..., \| p_{\tau_i-u+1,s_j}] = TC_u(\tau_i, s_j) \tag{7}$$

where $\tau_i$ is the date when a tweet $t_i$ is created, $TC_u$ is a function to calculate technical indicators of stock $s_j$ for the previous $u$ days from day $\tau_i$, $p_{\tau_i,s_j}$ is a vector of technical indicators of stock $s_j$ on day $\tau_i$, and $\|$ shows concatenation operation. In our experiments, we use the most representative technical indicators such as Bollinger Bands, average directional movement index (ADX), moving average convergence/divergence (MACD), relative strength index (RSI), and stochastics. We calculate $P_{\tau_i}$ the matrix of technical indicators for all the stocks on the day $\tau_i$, where $P_{\tau_i} = [TC_u(\tau_i, s_1) \| TC_u(\tau_i, s_2), ..., \| TC_u(\tau_i, s_m)]^T$ Then, we use CNN to extract the market features, $m_{\tau_i}$.

### 4.4 Discussion Module

The discussion module incorporates the discussion on a social platform. In social networking platforms like Twitter, investors exchange information and formulate opinions. It is natural that the previous discussion on the social platform influences investors'

**Table 2: Experimental results and ablation studies.**

|  | N@5 | N@10 | P@5 | P@10 | R@5 | R@10 | F@5 | F@10 |
|---|---|---|---|---|---|---|---|---|
| *Previous Mentioned Stock* | 0.1387 | 0.1397 | 0.0341 | 0.0173 | 0.1176 | 0.1734 | 0.0569 | 0.0315 |
| *Most Popular Mentioned Stock* | 0.1685 | 0.1813 | 0.0533 | 0.0308 | 0.2666 | 0.3077 | 0.0889 | 0.0559 |
| B-LSTM | 0.2115 | 0.2482 | 0.0673 | 0.0449 | 0.3367 | 0.4488 | 0.1122 | 0.0816 |
| PERD | 0.2555 | 0.2936 | 0.0740 | 0.0488 | 0.3700 | 0.4871 | 0.1234 | 0.0886 |
| PDRSI | **0.3068** | **0.3405** | **0.0880** | **0.0543** | **0.4396** | **0.5429** | **0.1466** | **0.0988** |
| -Stock Preference Module | -0.0052 | -0.0024 | -0.0254 | -0.0240 | -0.0085 | -0.0045 | -0.0337 | -0.0335 |
| -Personal Interest Module | -0.0337 | -0.0334 | -0.0051 | -0.0024 | -0.0253 | -0.0239 | -0.0085 | -0.0044 |
| -Market Module | -0.0132 | -0.0134 | -0.0046 | -0.0022 | -0.0226 | -0.0226 | -0.0076 | -0.0042 |
| -Discussion Module | -0.0070 | -0.0051 | -0.0013 | 0.0000 | -0.0064 | 0.0000 | -0.0022 | 0.0000 |

interest. Thus, we incorporate the previous discussion in the following way. Given day $\tau_i$, we denote $D_{\tau_i, s_j} = [t'_1, t'_2, ..., t'_d]$ as the set of tweets created on the previous day of $\tau_i$ about stock $s_j$. We encode tweets, average them for each stock, concatenate them, and use CNN layers to extract discussion features, $d_{\tau_i}$.

## 4.5 Prediction and Model Training

To predict the probability $p_i$ for tweet $t_i$ and define the loss function to optimize the model, we add the market vector $m_{\tau_i}$, discussion vector $d_{\tau_i}$, temporal vector $e_i$, and preference vector $a_i$ to get a vector $z_i$. We then feed the vector $z_i$ to a linear layer, and softmax layer to obtain probability $p_i$. We use the cross-entropy function for model optimization. Adam [9] is used for optimizer with its initial learning rate $5e - 5$. In the market module, we set the length of technical indicators $u$ to 7 days. We set the length of historical tweets $k$ to 4, following Zheng et al. [22].

## 5 EXPERIMENT

### 5.1 Experimental Setting

We follow previous work [8] to use the leave-one-out evaluation approach. That is, we held out the latest interaction as each investor's test data. We randomly chose one interaction for each investor and used it as validation data, and we utilized all the remaining data for training data. The model with the best performance on the validation set is used for the test. We used nDCG@K, Precision@K, Recall@K, and F1@K for evaluation.

### 5.2 Experimental Results

We compare the proposed PDRSI with two statistical baselines and two neural-based baselines. The *Previous Mentioned Stock* is based on the user's previous tweet, and recommends the same stock that is mentioned in the latest tweet. The *Most Popular Mentioned Stock* is based on the number of times the user has mentioned the stock in the sequence of historical tweets. These statistical baselines assume that users' interests and items' properties will not dynamically change and can be predicted based on the sequence of historical tweets. On the other hand, the neural-based baselines including B-LSTM [2], and PERD [22] are proposed to capture the dynamic change in users' interests.

Table 2 shows the experimental results. The proposed PDRSI outperforms all baselines regardless of the evaluation metrics used. Compared with PERD, the proposed PDRSI achieves 20.1% and 16.0%

**Table 3: Comparison of different text encoders.**

|  | N@10 | P@10 | R@10 | F@10 |
|---|---|---|---|---|
| PDRSI (BERTWEET) | 0.3279 | 0.0529 | 0.5286 | 0.0961 |
| PDRSI (BERT) | 0.3395 | 0.0535 | 0.5351 | 0.0973 |
| PDRSI (FinBERT) | **0.3405** | **0.0543** | **0.5429** | **0.0988** |

improvement on nDCG@5 and nDCG@10. The comparison of the NN-based methods and statistical methods shows that investors' interest is not fixed, and dynamically capturing and updating user properties are needed. We further perform an ablation analysis, and the results show that the stock preference module and personal interest module have a higher influence on the performance than other modules, and the influence of the discussion module is the least. We consider the discussion module does not perform well as it contains noisy information by averaging all the discussions on social media. Therefore, we proceed to extract important discussions in the future work.

We further provide a comparison among different text encoders. We use the proposed PDRSI with BERT [7], BERTWEET [11], and FinBERT [1]. Table 3 shows the experimental results. The results imply that the domain-specific LM performs better than the general LM and source-oriented LM in the proposed task.

## 6 CONCLUSION

This study proposes a new task about personalized dynamic recommendations for investors. We share a new dataset, D3, and design a well-performing model, PDRSI, for the proposed task. We hope this study can raise more attention to personalized dynamic recommender systems in investment scenarios, in which all features dynamically and rapidly change over time. We plan to extend our results by extracting important information from social discussions and adopting follow-follower network information to model how the information is conveyed to an investor.

# REFERENCES

[1] Dogu Araci. 2019. Finbert: Financial sentiment analysis with pre-trained language models. *arXiv preprint arXiv:1908.10063* (2019).

[2] Francesco Barbieri, Miguel Ballesteros, and Horacio Saggion. 2017. Are Emojis Predictable?. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*. Association for Computational Linguistics, Valencia, Spain, 105–111. https://aclanthology.org/E17-2017

[3] Jun Chang and Wenting Tu. 2018. A Stock-Movement Aware Approach for Discovering Investors' Personalized Preferences in Stock Markets. In *2018 IEEE 30th International Conference on Tools with Artificial Intelligence (ICTAI)*. 275–280. https://doi.org/10.1109/ICTAI.2018.00051

[4] Jun Chang, Wenting Tu, Changrui Yu, and Chuan Qin. 2021. Assessing dynamic qualities of investor sentiments for stock recommendation. *Information Processing Management* 58, 2 (2021), 102452. https://doi.org/10.1016/j.ipm.2020.102452

[5] Chung-Chi Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2019. Next cashtag prediction on social trading platforms with auxiliary tasks. In *Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. 525–527.

[6] Vincent Cho. 2010. MISMIS–A comprehensive decision support system for stock market investment. *Knowledge-Based Systems* 23, 6 (2010), 626–633.

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. https://doi.org/10.18653/v1/N19-1423

[8] Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. 2017. Neural Collaborative Filtering *(WWW '17)*. International World Wide Web Conferences Steering Committee, 173–182.

[9] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization.. In *3rd International Conference on Learning Representations*, Yoshua Bengio and Yann LeCun (Eds.).

[10] Chi-Jie Lu, Tian-Shyug Lee, and Chih-Chou Chiu. 2009. Financial time series forecasting using independent component analysis and support vector regression. *Decision support systems* 47, 2 (2009), 115–125.

[11] Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. BERTweet: A pre-trained language model for English Tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. 9–14.

[12] Tao Qi, Fangzhao Wu, Chuhan Wu, and Yongfeng Huang. 2022. FUM: Fine-Grained and Fast User Modeling for News Recommendation. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Madrid, Spain) *(SIGIR '22)*. Association for Computing Machinery, New York, NY, USA, 1974–1978. https://doi.org/10.1145/3477495.3531790

[13] Tao Qi, Fangzhao Wu, Chuhan Wu, and Yongfeng Huang. 2022. News recommendation with candidate-aware user modeling. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1917–1921.

[14] Chhavi Rana and Sanjay Kumar Jain. 2015. A study of the dynamic features of recommender systems. *Artificial Intelligence Review* 43 (2015), 141–153.

[15] Robin M. E. Swezey and Bruno Charron. 2018. Large-Scale Recommendation for Portfolio Optimization. In *Proceedings of the 12th ACM Conference on Recommender Systems*. 382–386.

[16] Mona Taghavi, Kaveh Bakhtiyari, and Edgar Scavino. 2013. Agent-based computational investing recommender system. In *Proceedings of the 7th ACM conference on recommender systems*. 455–458.

[17] Mona Taghavi, Kaveh Bakhtiyari, and Edgar Scavino. 2013. Agent-based computational investing recommender system. In *Proceedings of the 7th ACM Conference on Recommender Systems*. 455–458. https://doi.org/10.1145/2507157.2508072

[18] Preeti Voditel and Umesh Deshpande. 2013. A stock market portfolio recommender system based on association rule mining. *Applied Soft Computing* 13 (2013), 1055–1063. https://doi.org/10.1016/j.asoc.2012.09.012

[19] Yumo Xu and Shay B. Cohen. 2018. 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)*. Association for Computational Linguistics, Melbourne, Australia, 1970–1979. https://doi.org/10.18653/v1/P18-1183

[20] Jungsoon Yoo, Melinda Gervasio, and Pat Langley. 2003. An adaptive stock tracker for personalized trading advice. In *Proceedings of the 8th international conference on Intelligent user interfaces*. 197–203.

[21] Yang Yujun, Li Jianping, and Yang Yimei. 2016. An Efficient Stock Recommendation Model Based on Big Order Net Inflow. *Mathematical Problems in Engineering* 2016 (2016), 1–15. https://doi.org/10.1155/2016/5725143

[22] Xuanzhi Zheng, Guoshuai Zhao, Li Zhu, and Xueming Qian. 2022. PERD: Personalized Emoji Recommendation with Dynamic User Preference *(SIGIR '22)*. Association for Computing Machinery, New York, NY, USA, 1922–1926. https://doi.org/10.1145/3477495.3531779

[23] Dávid Zibriczky12. 2016. Recommender systems meet finance: a literature review. In *Proc. 2nd Int. Workshop Personalization Recommender Syst.* 1–10.