The 12th International Conference on
Computer Science & Education (ICCSE 2017)
August 22-25, 2017. University of Houston, USA

L1-regularized Logistic Regression for Event-driven Stock Market Prediction

Si-Shu Luo
College of Mathematics
Sichuan University
Chengdu, China

Yang Weng
College of Mathematics
Sichuan University
Chengdu, China

Wei-Wei Wang
Automation Department
Xiamen University
Xiamen, China

Wen-Xing Hong
Automation Department
Xiamen University
Xiamen, China

***Abstract*—This paper presents a machine learning method for event-driven stock prediction, using L1 regularized Logistic regression model. It studies the stock price movement after listed companies make announcements. The model uses specific events extracted from these announcements and combine with financial indicators of listed companies, macro indicators, and technical indicators as dependent variables. The listed companies are divided into sample sets based on market value size and industry. Experiments show that this model can be a good predictor of stock within one week after events occur. In addition, compared with commonly used machine learning methods, our model has a better overall ability.**

***Index Terms*—stock, event-driven, L1 regularized logistic regression**

I. Introduction

Financial markets are driven by information. By integrating information, we can understand the market better. Based on data structure, the data in the financial market is divided into structured data and unstructured data.

Structured data has been fully utilized in financial sector, such as time series trading data, etc.. For example, some traditional methods predict the stock price trend by fundamental analysis and technical analysis. The fundamental analysis uses financial statements to describe the company's basic situation and use macroeconomic data to describe the overall economic environment, which reflects the intrinsic value of the stock. Technical analysis [1] is a long-term observation, which used by investors to find a number of laws on the stock. The common method starts from the K line chart and technical indicators, that is, using the direct performance of historical data to predict the future trend of stock price.

Unstructured data also contains huge information. For instance, investors sometimes make decisions based on the announcements released by listed companies. However, the information that one person can receive and process is too limited, so we hope to use machine to extract core value and associate it with the stock price. However, how to effectively extract the information from text data has always been a problem [2], [3], which leads to a large amount of value in the text data not being explored. Fama first proposed event study [4] to examine whether the market responded quickly and reflected it into the stock price when new information appeared. The specific approach was to examine whether the stock price fluctuated abnormally before and after an event occurred. This method is proposed based on the efficient market hypothesis [5], [6] that the current price has adequately reflected all available public information. Nuij et al. [7] and Ding et al. [8] predicted the stock based on event study too. The former extracted events from the news and analyzed the impact on stock price, then established the optimal trading strategy from the events and the technical indicators. The latter extracted events from the news and used tensor neural network to express the event [9]. Finally, events were directly used as inputs of the convolution neural network [10], [11] to predict the stock. it aimed to find the direct relationship between events and the stock price [12]. Sometimes, public opinions will be produced after the news releases, which make associated stock price fluctuate. There are various methods to determine the market sentiment using emotional analysis of text, then predict stock by emotional analysis. In [13], authors predicted the stock through public sentiment embodied in the media. And in [14], authors predicted the stock by analyzing the emotional colors hidden in the text.

Stock market is very sensitive to information, especially in China's stock market. A large part of the stock price volatility is caused by the spread of the news. Therefore, it's meaningful to study the change of stock price after information disclosure. We can aquire stock news through a lot of channels, including the major financial media, stocks forum, etc., from which the information is so complex that exist a lot of noise. However, announcements of listed company have advantages of real content and timely release because of the disclosure of the obligation. So using announcements can not only decrease the SNR, but also improve the relevance of text contents with listed companies. This paper focuses on the impact of events from the announcements in the stock market.

We use as many input variables as possible since there are huge number of factors that determine the price of financial market, and aim to predict the stocks rise or fall after events occur. Therefore, it requires a classification model that can handle high-dimensional variables. The L1 regularized penalty [15] adds one norm to the classical Logistic regression loss function, which makes the model have ability to deal with high-dimensional features. It can automatically remove unimportant variables, which reduces the complexity of the model and improves the accuracy as well as the robustness of the model.

978-1-5090-2508-4/17/$31.00 ©2017 IEEE

This paper combines three aspects of the news, fundamental information and historical market information to make a comprehensive description of the stock market. Regarding the amount of text information, events will be extracted only from announcements and study the change of the stock price based on the date when events occur. For the fundamentals, it will use listed companies' financial situation. All information directly reflects the stock's "intrinsic value", which is the most basic reason to determine the share price. This paper will also proceed from the macroeconomic data, which is the performance of economic trends of the country and even the world. In the long run and fundamentally, the trend of the stock is determined by the whole economic situation. In the technical level, this paper will use the historical market data to calculate technical indicators to describe different aspects of the historical trend of stock prices.

Economic structures vary from different industries, so the responses to the same event are different. Apparently, different market values of listed companies make the sensitivity of the incident different. Therefore, this paper divides the listed companies into different categories based on the industry and the number of current capital stock.

II. L1 Regularized Logistic Regression for Event-Driven Stock Prediction

Logistic regression is a classical two-class model, which limits the output variable to [0,1] by the sigmoid function to represent the classification probability. One difficulty of this traditional model is feature selection. Because these selected features not only need to cover the relevant factors as much as possible, but also need to avoid the collinearity. To overcome this difficulty, one norm can compress some coefficients to zero, which plays the role of screening variables. At the same time, deleting some variables can also prevent the model from over-fitting.

This section will propose a L1 regularized logit model to predict the relative return and excess return after one kind of event occurs.

A. L1 Regularized Logistic Regression

In order to illustrate the principle of regularized logical regression, this paper will introduce the following two models. There are N pairs of observations (x_i, y_i), $x_i \in \mathbb{R}^p$ $i = 1, 2, ..., N$ are predictive variables and $y_i \in \mathbb{R}$ $i = 1, 2, ..., N$ whose value is taken in $\{0, 1\}$ are response variables. And the parameters $\beta_j \in \mathbb{R}$ $j = 1, 2, ..., p$ are what we need to estimate. In particular, we assume that x_{ij} are standardized, which is $\frac{1}{N}\sum_{i=1}^{N} x_{ij}^2 = 1$ $j = 1, 2, ..., p$.
The logistic regression is a binary classification model. The class-conditional probabilities of the logit model is shown below

$$P(y_i = 0|x_i) = \frac{1}{1 + e^{-(\beta_0 + x_i^\top \beta)}} \tag{1}$$

$$P(y_i = 1|x_i) = \frac{1}{1 + e^{(\beta_0 + x_i^\top \beta)}} \tag{2}$$

Hence we have the link function

$$\frac{P(y_i = 0|x_i)}{P(y_i = 1|x_i)} = \beta_0 + x_i^\top \beta \tag{3}$$

And the log-likelihood function is

$$\mathcal{L}(\beta_0, \beta) = \sum_{i=1}^{N} [y_i(\beta_0 + x_i^\top \beta) - ln(1 + \exp^{\beta_0 + x_i^\top \beta})] \tag{4}$$

Maximizing the log-likelihood function to estimate β_0 and β is logistic regression.

The l1-regularized logistic regression is to maximize log-likelihood function with L1 penalty, which is equivalent to

$$\min f(\beta_0, \beta) = \frac{1}{2N} l(\beta_0, \beta) + \lambda \|\beta\|_1 \tag{5}$$

where $\|.\|_1$ is one norm that is $\|\beta\|_1 = \sum_{j=1}^{p} |\beta_j|$, λ is the regularization parameter, and $l(\beta_0, \beta) = -\mathcal{L}(\beta_0, \beta)$.

From (5) we can see that, what we need to minimize is a non-smooth convex function, so the gradient descent is no longer feasible. In this paper, the method proposed by [16] is used to solve the problem. The method is solved by transforming the objective function into a lasso problem and then solved by the coordinate descent method [17].

B. Prediction Model

This model extracts the date when the event occurs to predict the stock movement.

- Model: L1 regularized logistic regression
- Input: Financial indicators of listed companies such as P/E ratio, etc. Macro data such as GDP, etc. And technical indicators calculated by historical market data.
- Output: Relative trading signal y_{it}^r and excess trading signal y_{it}^a after t days the event occurs. The relative trading signal and the excess trading signal are calculated by the relative return r_{it} and the excess return a_{it}, respectively. And i is the date when the event occurs and t is the days after an event occurs. The formula is the following

$$r_{it} = \frac{p_{i+t} - p_i}{p_i} \tag{6}$$

$$a_{it} = r_{it} - \beta_0 - \beta r_{it}^m \tag{7}$$

r_{it} is the relative stock return, r_{it}^m is the relative industry return which calculated by industry indicators, and p is the closing price. β_0 and β are estimated by $r_{it} = \beta_0 + \beta r_{it}^m + \epsilon_{it}$. Hance, y_{it}^r and y_{it}^a are calculated as follows

$$y_{it}^r = \begin{cases} 1 & r_{it} \geq 0 \\ 0 & r_{it} \leq 0 \end{cases} \quad y_{it}^a = \begin{cases} 1 & a_{it} \geq 0 \\ 0 & a_{it} \leq 0 \end{cases} \tag{8}$$

III. EXPERIMENT

Firstly, we choose the shareholding increase of executives [18], [19] as the study object. This event shows the companys executives are optimistic about the company's development and have confidence in their business. So it's usually a signal of the improvement of company fundamentals. It is appropriate to select the the shareholding increase of executives as a specific event to study.

Secondly, we choose four industries, real estate, medicine, nonferrous metals and computer, as our study objects. These four industries have certain representation. The real estate industry is closely linked with the macroeconomic and policy; The "safe haven" nature of medicine industry makes this industry not always consistent with the macroeconomic trends; Computer industry has its own characters as a new industry; The last non-ferrous metal industry plays a role of verifier as a triditional industry. Then, we divide each industry by the current capital stock. In each industry, we selected those listed companies whose current capital stocks are more than ten thousand and less than one hundred thousand as samples.

A. Experimental Data

We acquired the shareholding increase of executives data from Eastern Fortune website and the rest of the data from the qutke website, including 165 financial indicators, 26 macro data and all historical market data. We use the historical market data to calculate the 51 technical indicators and use closing price to calculate $y^r_{it}, i = 1, 3, 5, 10, 30, 60$,$y^a_{it}, i = 1, 3, 5, 10, 30, 60$. See detailed data set at table I

TABLE I
STATISTICAL OF DATASETS

Industry	Training set	Time span
Medicine	660	2012-12-25∼2016-01-27
Computer	478	2010-03-03∼2016-05-18
Real estate	420	2007-05-09∼2015-08-28
Nonferrous metals	326	2007-03-07∼2015-12-10
Industry	**Back-Test set**	**Time span**
Medicine	166	2016-01-29∼2017-05-11
Computer	120	2016-05-18∼2017-05-11
Real estate	106	2015-08-28∼2017-05-09
Nonferrous metals	82	2015-12-11∼2017-05-10

B. Evaluation criteria

This paper uses two evaluation criteria. The first criterion is to evaluate the quality of the model, we chooses two indicators as model evaluation criterion, one of which is the 10 fold cross validation prediction error err^{cv} calculated in training set , and the other is the prediction error err^{test} calculated in test set which is one tenth of training set.

Although the first evaluation criterion has been predicted error in the test set, the stock market in the different periods of time the price trend may vary widely. In order to show that the model has the ability to predict the stock, it is necessary to test the trained model on the data set on another time period, so the second evaluation criterion is to describe if the model has the ability to predict the stock, so we calculate the back-test prediction error $error^{bt}$ on the back-test set consisting of historical data of different time intervals from the training set.

C. Experimental results and analysis

1) Statistical description: In order to get an description of the impact of the event on the stock, we select these relative returns $r_{it}, i \in D, t = 1, 3, 5, 10, 30, 60$ and excess returns $a_{it}, i \in D, t = 1, 3, 5, 10, 30, 60$ on the date $D = (d_1, d_2, ..., d_n)$ when the event occurred to form sample sets. We get the general situation of the data set by basic statistical method and we describe the sample sets from three aspects. Firstly, use the significance test to test whether the stock rise after the event occurs. And then reflect the overall trend of the data set through the average. Finally, count the number of stock rises or falls after the event occurs. The statistical results are shown in the table II

TABLE II
STATISTICAL DESCRIPTION FOR RELATIVE RETURNS AND EXCESS RETURNS AFTER AN EVENT OCCURRED x DAYS

	Medicine			**Real estate**		
$r_x\&a_x$	p	avrg	rate	p	avrg	rate
r_1	0.231	0.001	55.0%	0.140	0.002	58.1%
r_3	0.080	0.005	53.0%	0.021	0.003	70.0%
r_5	0.444	0.000	54.6%	0.013	0.004	60.3%
r_{10}	0.127	0.001	56.8%	0.623	0.000	60.6%
r_{30}	0.383	0.000	55.2%	0.141	0.001	64.3%
r_{60}	0.000	0.004	59.6%	0.266	0.001	60.0%
a_1	0.618	0.000	51.6%	0.099	0.001	55.1%
a_3	0.153	0.001	53.0%	0.036	0.002	56.8%
a_5	0.295	0.000	52.7%	0.020	0.002	58.6%
a_{10}	0.664	0.000	50.5%	0.001	0.003	56.5%
a_{30}	0.004	0.001	53.8%	0.223	0.001	54.4%
a_{60}	0.576	0.000	50.6%	0.790	0.001	49.0%
	Computer			**Nonferrous metals**		
$r_x\&a_x$	p	avrg	rate	p	avrg	rate
r_1	0.486	0.000	55.7%	0.011	0.006	59.3%
r_3	0.726	-0.001	52.0%	0.224	0.002	56.4%
r_5	0.532	0.000	53.7%	0.104	0.003	55.9%
r_{10}	0.772	-0.001	52.8%	0.154	0.002	57.1%
r_{30}	0.661	-0.001	57.7%	0.032	0.003	61.5%
r_{60}	0.153	0.001	59.2%	0.013	0.004	63.4%
a_1	0.201	0.001	50.5%	0.462	0.000	51.0%
a_3	0.433	0.000	51.0%	0.589	0.000	54.4%
a_5	0.722	0.000	48.0%	0.564	0.000	53.4%
a_{10}	0.539	0.000	50.5%	0.357	0.000	52.7%
a_{30}	0.999	-0.004	43.0%	0.367	0.000	55.9%
a_{60}	0.171	0.001	52.7%	0.130	0.001	53.9%

$r_x\&a_x$: relative returns and excess returns
p: one tailed t-test significance the 95% level
avrg: the average of sample sets
rate: the rate of the stock didn't fall

Thus we can draw the following three conclusions

- After the event occurred, the stock price rise is not a high probability event, so predict the stock after the event occurred is meaningful.
- Different industries (Each industry divides the number of current capital stock in the same way) have different degrees of sensitivity to the same event. In our case, the impact of the real estate industry is more pronounced, and the mean of $r_3, r_5, a_3, a_5, a_{10}$ is significantly greater

than zero. Followed by the medicine industry and the nonferrous metals industry, the mean of r_3, r_{60}, a_{30} and r_1, r_{30}, r_{60} is significantly greater than zero respectively.
- Different industries react at different rates to the same event. In our case, the real estate industry respond in short term, but with the dissemination of information, the impact of the event gradually disappeared. But for the medicine industry and non-ferrous metals industry, after a long period of time this event still has an impact on the stock.

In general, the relative returns is more affected by the event than the excess returns. This may be because the impact of the industry index on the stock is greater than the impact of the event on the stock.

2) Experimental Results and Analysis: This subsection will show the experimental results, we mainly analyze the model established by L1 regularized logistic regression. And in order to understand the problem more deeply, we select two common machine learning models which are the decision tree and random forest to compare the three methods to draw conclusions.

- From the table III we can see that, in general, the predicted results within one week after the event occurs are the best, which means that the model has chosen various technical indicators which are frequently used to predict the short-term trend of the stock, result in short-term stock price forecast more accurate. We will discuss it latter.
- The prediction of the relative trading signal is better than the prediction of the excess trading signal. Because we use the industry indicators to calculate the excess signal, which means that the excess signal and the industry trend is closely related. But in our model, the input variables cover too little industry information, resulting in poor predictions.
- The L1 regularized logical regression has the best comprehensive results. Although the decision tree model performs best on the training set, the performance on back-test set is poor, which indicates that this model has poor robustness on such data sets. And for the random forest model, both the training error and the test error are worse than the L1 regularized logistic regression.

Next, we will discuss the coefficients selected by the L1 regularized logistics regression. Since this model has ability to automatically select important variables, we can further understand the model based on the variables the model chooses. Table IV shows the number of selected variables from financial indicators of listed companies, macro indicators and technical indicators, from which we can draw some conclusions.

- Overall, short-term stock movements are more likely to rely on technical indicators and long-term stock movements tend to dependent on the financial indicators of listed companies and macro indicators, which is also in line with the actual situation.

TABLE III
MODEL COMPARISON TABLE

Medicine						
	L1 log		**Decsion Tree**		**Random Forest**	
$r_x\&a_x$	err^{cv}	err^{bt}	err^{cv}	err^{bt}	err^{test}	err^{bt}
r_1	0.021	0.090	0.000	0.000	0.000	0.000
r_3	0.270	0.337	0.428	0.355	0.364	0.386
r_5	0.355	0.410	0.418	0.482	0.333	0.361
r_{10}	0.382	0.434	0.509	0.476	0.394	0.440
r_{30}	0.444	0.524	0.460	0.548	0.470	0.506
r_{60}	0.453	0.512	0.404	0.500	0.333	0.470
a_1	0.442	0.422	0.382	0.470	0.409	0.458
a_3	0.459	0.554	0.528	0.452	0.379	0.464
a_5	0.485	0.434	0.467	0.536	0.561	0.452
a_{10}	0.476	0.506	0.495	0.458	0.576	0.446
a_{30}	0.409	0.566	0.509	0.572	0.561	0.566
a_{60}	0.465	0.614	0.454	0.506	0.515	0.530
Computer						
	L1 log		**Decsion Tree**		**Random Forest**	
r_1	0.021	0.008	0.000	0.000	0.000	0.000
r_3	0.272	0.358	0.294	0.425	0.333	0.408
r_5	0.328	0.358	0.369	0.575	0.333	0.475
r_{10}	0.387	0.483	0.354	0.508	0.333	0.417
r_{30}	0.458	0.533	0.403	0.483	0.479	0.558
r_{60}	0.389	0.442	0.382	0.467	0.146	0.517
a_1	0.481	0.467	0.437	0.467	0.500	0.433
a_3	0.483	0.517	0.377	0.492	0.563	0.458
a_5	0.477	0.517	0.387	0.483	0.479	0.550
a_{10}	0.469	0.592	0.368	0.508	0.479	0.500
a_{30}	0.402	0.467	0.448	0.500	0.479	0.558
a_{60}	0.418	0.425	0.388	0.500	0.521	0.500
Real estate						
	L1 log		**Decsion Tree**		**Random Forest**	
r_1	0.045	0.066	0.000	0.000	0.000	0.000
r_3	0.305	0.340	0.401	0.368	0.381	0.396
r_5	0.371	0.358	0.345	0.491	0.357	0.500
r_{10}	0.381	0.519	0.365	0.462	0.571	0.377
r_{30}	0.440	0.528	0.442	0.557	0.476	0.415
r_{60}	0.467	0.462	0.385	0.566	0.524	0.500
a_1	0.452	0.472	0.485	0.481	0.357	0.443
a_3	0.445	0.368	0.403	0.575	0.476	0.387
a_5	0.417	0.396	0.434	0.462	0.429	0.349
a_{10}	0.414	0.519	0.429	0.481	0.381	0.425
a_{30}	0.450	0.538	0.365	0.491	0.452	0.538
a_{60}	0.431	0.491	0.380	0.519	0.548	0.443
Nonferrous metals						
	L1 log		**Decsion Tree**		**Random Forest**	
$r_x\&a_x$	err^{cv}	err^{bt}	err^{cv}	err^{bt}	err^{test}	err^{bt}
r_1	0.040	0.024	0.000	0.000	0.000	0.000
r_3	0.347	0.195	0.320	0.378	0.455	0.390
r_5	0.359	0.268	0.365	0.378	0.364	0.329
r_{10}	0.393	0.500	0.399	0.488	0.485	0.476
r_{30}	0.436	0.573	0.289	0.390	0.515	0.463
r_{60}	0.402	0.451	0.361	0.439	0.394	0.427
a_1	0.454	0.451	0.346	0.439	0.515	0.585
a_3	0.454	0.451	0.393	0.427	0.455	0.415
a_5	0.463	0.488	0.316	0.537	0.636	0.598
a_{10}	0.466	0.561	0.435	0.463	0.606	0.610
a_{30}	0.411	0.463	0.454	0.524	0.455	0.476
a_{60}	0.423	0.549	0.418	0.488	0.303	0.537

$r_x\&a_x$: relative returns and excess returns
err^{cv}: ten fold cross validation error
err^{bt}: back test error
err^{test}: test error

TABLE IV
L1 REGULARIZED LOGISTIC REGRESSION COEFFICIENTS

	Medicine			**Real estate**		
r_x&a_x	fncl	macro	tech	fund	macro	tech
r_1	7	1	6	20	2	6
r_3	4	8	21	84	11	31
r_5	0	0	1	0	0	1
r_{10}	0	1	0	0	0	2
r_{30}	55	12	22	0	0	0
r_{60}	21	1	8	0	0	0
a_1	8	3	5	5	2	2
a_3	0	0	0	0	0	0
a_5	0	0	0	0	0	0
a_{10}	6	0	7	4	3	1
a_{30}	8	5	6	0	0	0
a_{60}	0	0	0	3	0	4
rate	50.5%	14.4%	35.2%	64.1%	9.9%	26.0%
	Computer			**Nonferrous metals**		
r_x&a_x	fncl	macro	tech	fund	macro	tech
r_1	1	0	2	0	0	2
r_3	36	6	21	45	10	21
r_5	0	0	1	0	0	1
r_{10}	22	6	13	0	0	1
r_{30}	7	1	6	0	0	0
r_{60}	35	8	13	53	7	18
a_1	9	0	4	72	11	30
a_3	14	4	4	0	0	0
a_5	0	0	0	75	16	34
a_{10}	0	0	0	0	0	0
a_{30}	0	0	0	0	0	0
a_{60}	42	6	10	0	0	0
rate	61.3%	11.4%	27.3%	61.9%	11.1%	27.0%

r_x&a_x: relative returns and excess returns
rate: $\frac{number\ of\ category}{(total\ number\ of\ coefficient)}$
fncl: The number of financial indicators of listed companlies
macro: The number of macro indicators
tech: The number of technical indicators

- The factors selected in this paper are not sufficient to predict the excess trading signal. We can see that the model compresses the coefficients of the unimportant factor into zero.
- From the aspect of industry, the medicine industry is mainly dependent on listed companies fundamentals and technical indicators, rather than relies on macro indicators, which is consistent with our understanding of the medicine industry. The remaining three industries are similar, the number of listed companies listed in the fundamentals are more than 60%, which means that the industry is very dependent on the situation of listed companies themselves.

IV. CONCLUSIONS AND FUTURE WORK

This paper studies the impact of specific events on stock prices. We describe the change of stock prices through relative returns and excess returns. And we use the shareholding increase of executives as a research object. This event is not only an improving signal of company's fundamentals, but also is what investors concerned about. After this, we choose medicine, computer, real estate and nonferrous metals these four industries and divide them into four categories by current capital stocks. We study stocks in these four categories.

Before building the model, we found out that the event is indeed affected by the stock price by analyzing data using statistics. And for different industries, the degree of sensitivity and speed of response are different. Also, the impact of events on relative returns is greater than the impact on excess returns. After establishing the model, we realize that the model is better at short-term stock forecasting, and the prediction effect of the relative returns is better than that of the excess returns. We also test two other machine learning models, and get a conclusion that L1 logistic regression is the optimal model from the comprehensive effect point of view.

Future research will focus on finding more events that are worthy of study and adding independent variables which contain industry information to improve the forecast of excess returns and affect the change of stock price in long term.

ACKNOWLEDGEMENTS

This work was supported in part by Fujian Shine Technology Limited Company and the National Natural Science Foundation of China (No.61032001).

REFERENCES

[1] Robert D. Edwards and John Magee. *Technical Analysis of Stock Trends*. CRC Press :, 2007.
[2] Alan Ritter, Mausam, Oren Etzioni, and Sam Clark. Open domain event extraction from twitter. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1104–1112, 2012.
[3] Yusuke Shinyama and Satoshi Sekine. Preemptive information extraction using unrestricted relation discovery. In *Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 4-9, 2006, New York, New York, USA*, 2006.
[4] Eugene F. Fama, Lawrence Fisher, Michael C. Jensen, and Richard Roll. The adjustment of stock prices to new information. *International Economic Review*, 10(1):1–21, 1969.
[5] Eugene F. Fama. The behavior of stock-market prices. *Journal of Business*, 38(1):34–105, 1965.
[6] Eugene F. Fama. Random walks in stock market prices. *Financial Analysts Journal*, 21(5):55–59, 1965.
[7] Wijnand Nuij, Viorel Milea, Frederik Hogenboom, Flavius Frasincar, and Uzay Kaymak. An automated framework for incorporating news into stock trading strategies. *International Journal of Research in Computer Applications And Robotics*, 2(11):823–835, 2014.
[8] Xiao Ding, Yue Zhang, Ting Liu, and Junwen Duan. Deep learning for event-driven stock prediction. In *International Conference on Artificial Intelligence*, pages 2327–2333, 2015.
[9] Richard Socher, Danqi Chen, Christopher D Manning, and Andrew Y Ng. Reasoning with neural tensor networks for knowledge base completion. In *International Conference on Neural Information Processing Systems*, pages 926–934, 2013.
[10] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. *Learning internal representations by error propagation*. MIT Press, 1988.
[11] Kuo Sheng Cheng, Jzau Sheng Lin, and Chi Wu Mao. The application of competitive hopfield neural network to medical image segmentation. *IEEE Transactions on Medical Imaging*, 15(4):560–7, 1996.
[12] Huong Thanh Nguyen, Tlis J Putni??, and Eliza Wu. What moves stock prices? *Social Science Electronic Publishing*, 2016.
[13] Johan Bollen, Huina Mao, and Xiaojun Zeng. Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1):1–8, 2011.
[14] Jianfeng Si, Arjun Mukherjee, Bing Liu, Qing Li, Huayi Li, and Xiaotie Deng. Exploiting topic based twitter sentiment for stock prediction. In *ACL*, pages 1779–1787, 2013.
[15] P Mccullagh and J. A. Nelder. Generalized linear models. *European Journal of Operational Research*, 16(3):285–292, 1984.
[16] J Friedman, T Hastie, and R Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1, 2010.

[17] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Pathwise coordinate optimization. *Annals of Applied Statistics*, 1(2):302–332, 2007.

[18] Michael C. Jensen. Agency costs of free cash flow, corporate finance, and takeovers. *American Economic Review*, 76(2):323–329, 1986.

[19] Terry D. Warfield, John J. Wild, and Kenneth L. Wild. Managerial ownership, accounting choices, and informativeness of earnings. *Journal of Accounting & Economics*, 20(1):61–91, 1995.