Proceedings of the International MultiConference of Engineers and Computer Scientists 2019
IMECS 2019, March 13-15, 2019, Hong Kong

Stock Market Trend Prediction with Sentiment Analysis based on LSTM Neural Network

Xu Jiawei, Tomohiro Murata

Abstract—**This paper aims to analyze influencing factors of stock market trend prediction and propose an innovative neural network approach to achieve stock market trend prediction. With the breakthrough of deep learning recently, there occurred lots of useful techniques for stock trend prediction. This thesis aims to propose a method of feature selection for selecting useful stock indexes and proposes deep learning model to do sentiment analysis of financial news as another influencing factor influencing stock trend. Then it proposes accurate stock trend prediction method using LSTM (Long Short-term Memory).**

Index: Stock trend prediction, LSTM, Sentiment Analysis, Deep learning, Chinese Stock market, Feature Selection…

I. Introduction

Stock market trend prediction plays a significant role in investment field. A lot of technique analysis methods occurred to solve this problem. With the stock market developing, traditional techniques can hardly achieve better performance now. With deep learning developing, a lot of techniques like LSTM, neural network proved to be effective in finance field. We aim to use deep learning method on stock trend prediction and analysis the influencing factors of stock trend prediction method based on LSTM neural network. This study mainly focused on feature selection, sentiment analysis of financial news and neural network structure.

II. Background / Related Literature / Datasets

A. Background

Prediction of stock market has attracted attention from industry to academia. Various machine learning algorithms such as neural networks, genetic algorithms, support vector machine, and others are used to predict stock price.

However, accuracy is unsatisfied, because of the reasons:

1.**Data Noise**:

There are lots of unprocessed factors. It causes some problems like data redundancy, data noise and overfitting.

2.**Market Emotion**:

Stock market is a stochastic field. Various aspects influent investors' emotion. Market emotion strongly affects the stock market trend. And investors' emotion is usually affected by financial news.

3.**Time series Information**:

Traditional methods can handle time series data, but with limited performance. And some methods like ARIMA have poor effect on big data.

B. Related Literature

Huynh, Huy D., L. Minh Dang, and Duc Duong introduced a new prediction model depend on Bidirectional Gated Recurrent Unit (BGRU)[3]. Sun, Haonan[1] developed a predictive model to improve the accuracy by enhancing the denoising process which includes a training set selection based on four K-nearest neighbors (KNN) classifiers to generate a more representative training set and a denoising autoencoder-based deep architecture as kernel predictor. This paper has a shortcoming that he didn't straightly use the time series data. Some researchers attempt using financial news data to solve some stock selection problem[4].

C. Datasets

This project mainly uses 2 different kinds of datasets: Financial news data for sentiment analysis, stock data. We got this data from web crawler and did lots of data preprocessing of them to be suitable format.

C.1 **Sentiment analysis data**

When an investor plan to do investment on some stock, he always read some financial news from website to get some advice from experts or reports. Proposed method is to do sentiment analysis of news to help investors modify their investment strategy. Following is financial news data preprocessing workflow:

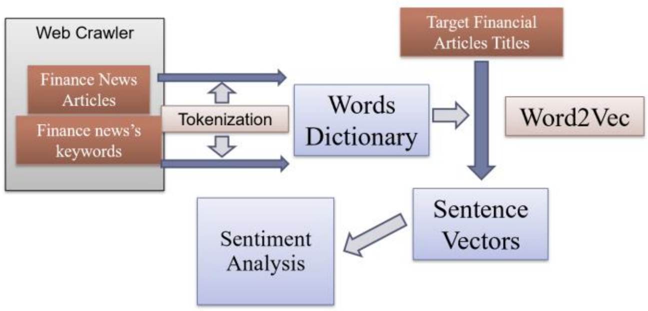


Fig. 1. Financial news processing diagram.

After web crawler processing, financial news title data becomes like this:

2016-04-01 00:03:56,等待非农就业数据 美股午盘维持小幅上扬

F. Xu Jiawei is graduate student of the Graduate School of Information, Production and Systems, Waseda University, Fukuoka, Japan 808-0135 (corresponding author to provide phone: +86-185-8871-8785; +81-070-4477-9045; e-mail: xujiawei@akane.waseda.jp) .

S. Tomohiro Murata is the professor of the Graduate School of Information, Production and Systems, Waseda University, Fukuoka, Japan 808-0135 (e-mail: t-murata@waseda.jp)

ISBN: 978-988-14048-5-5
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

2016-04-01 01:01:06,4月1日股市早盘内参：上交所决定4月8日起暂停*S
2016-04-01 01:19:11,净买入再超20亿元 沪股通扫货四大券商股
2016-04-01 01:22:06,专家:三月反攻圆满收官 四月行情值得期待
2016-04-01 01:38:17,梧桐树扎根A股“一箭双雕”
016-04-02 09:08:18,跨境电商的最后狂欢：红利消失 行业面临大洗牌
2016-04-02 09:10:22,大盘有望冲破强压力线
2016-04-02 09:14:12,四川鼓励农民进城购房 除成都外全面放开落户限
2016-04-02 09:14:16,站上三千 看高一线
2016-04-02 09:17:19,大突破节后将至目标剑指3500点以上
2016-04-05 07:16:46,年初A股市场暴跌影响难消 一季度股基平均亏损17

After preprocessing, we got finance news and transform them into sentence vectors as the input of sentiment analysis neural network model [10].

C.2 **Stock data**

Stock data mainly has 3 categories: Stock fundamental Index, Stock technique index, financial macro index. And they can be calculated for inventing higher-level technique indexes like moving average, KDJ. These are input of our neural network.

Category	Explaination
Open Price	The first price of a stock traded at the beginning of a specified trading day.
Close Price	The last price of a stock in the last transaction on a specified trading day.
Adjusted Close Price	The close price adjusted based on the reflection of dividends and splits.
Highest Price	The highest price when a stock traded on a specified trading day.
Lowest Price	The lowest price when a stock traded on a specified trading day.
Percentage Change of Adjusted Close Price	The percentage change of adjusted close price on ***i*** th trading day to the previous (i-1) th trading day
Trading Volume	Total amount of shares of a stock traded on a trading day

Table. 1. Stock Fundamental Index

Stock Index: Technical Index	TR	Turnover Rate
	MA	Moving Average
	MACD	Moving Average Convergence and Divergence
	DIF	Difference
	DEA	Difference Exponential Average
	KDJ_K	Stochastics Index
	KDJ_D	
	KDJ_J	
	BIAS	Bias
	RSI	Relative Strength Index
	ROC	Rate of Change
	PSY	Psychological line

Table. 2. Stock Technical Index

Category	Index	shorter form
Financial Index	Price earnings ratio	PE
	Price-to-sales Ratio	PS
	Price cash flow ratio	PC
	Price to book ratio	PB
A-share Macro Index	Open Price today	Open_A
	Highest Price today	High_A
	Lowest Price today	Low_A
	Close Price today	Close_A
	Volume Price today	Volume_A
	Change Price today	Change_A

Table. 3. Stock Macro Index
(A-share is the name of Chinese stock market. These indexes related to the composite index of whole market)

III. Approach

Main approach contains 3 parts: feature engineering, sentiment analysis of financial news and Deep Learning based trend prediction methods on stock data. In the feature engineering it has 2 steps: using auto encoder to do feature dimension reduction and comparing different feature combinations to find optimized features. In the sentiment analysis method of financial news, we labeled lots of financial news title text with pos/neg labels to represent their sentiment and then predict the newly published news data. In trend prediction model, I build LSTM type neural network for improving time series processing ability.

1.*Feature Engineering*

1) SDA (Stacked Denoising Auto Encoder) is applied to reduce the dimension of features which is not sensitive to the noise. [1] An autoencoder is a type of artificial neural network used to do unsupervised learning of data coding. The aim of an auto encoder is to learn higher-level representation for a set of data, typically for dimension reduction.

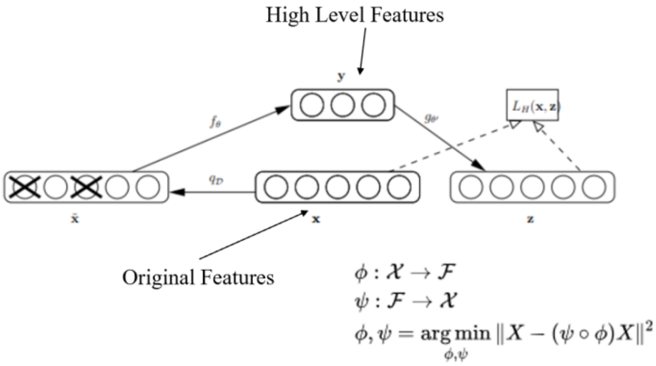


Fig. 2. Financial news processing diagram.

2) Then different sequence length and different sample amount are applied to experiment for finding an optimal sequence length [9].

2.*Sentiment Analysis of financial news*

LSTM neural network is applied for sentiment analysis. Input is sentence vectors. Then split data into several parts by different companies. Put them into LSTM layer. Then add a dense layer. Output layer will output sentiment analysis result, value range from 0.0~1.0. And then a 4-quantile value is used. If value <25%quantile as -1(negative); 25%quantile =< value <= 75%quantile as 0(even); value > 75%quantile as 1(positive). Model design is showing below:

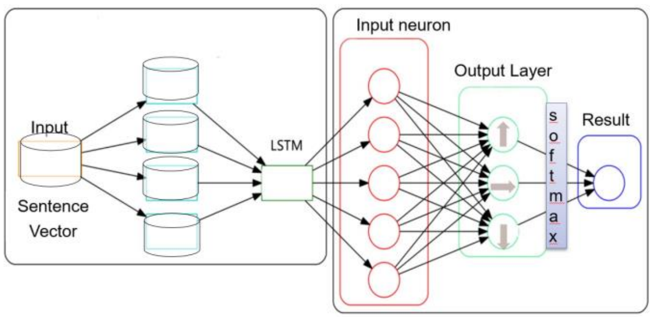


Fig. 3. Sentiment Analysis Tool (based on neural network)

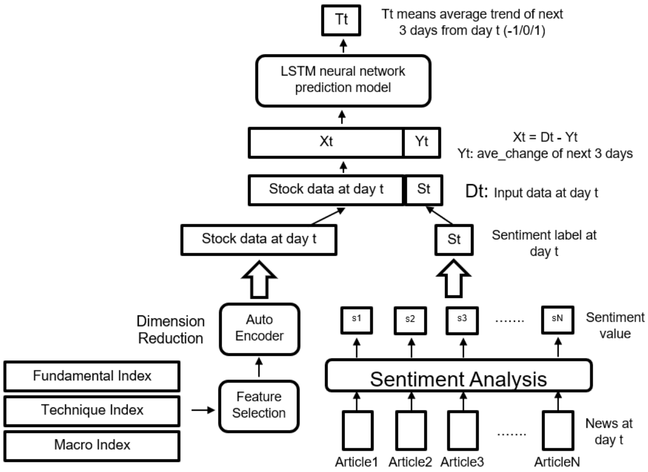
3.*Stock trend prediction*

LSTM is also used to build trend prediction model.

ISBN: 978-988-14048-5-5
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

Sentiment analysis results + stock data are input of LSTM Model [2]. And output is the trend of stock price movement.

After feature selection and sentiment analysis, we got stock data at day t and sentiment label at day t. Combine them as the input. Output is trend prediction result -1/0/1 representing negative/even/positive. The framework shows below:



IV. Experiments

It mainly contains 4 experiments: Accuracy Evaluation, Feature Selection Effect, Sentiment Analysis Performance Evaluation, and Stability and Generality Evaluation of proposed method.

Experiment 1: Accuracy Evaluation

Objective: Compare proposed method with other commonly used prediction methods using same experiment environment.

Process: 1. Using different prediction methods to predict stock trend and compare with proposed model. 30 days stock data as input and predict average stock trend of next 3 days.

2. Calculate average change price of next 3 days using change price:

3. Get 3 quantile value 0.33AC, 0.67AC , and convert average change price into category labels .

4.Training model, stock data as input matrix, *real_trend* as output and predict *pred_trend*.

Test data: I use stock data of Ping-an Finance Group (stock code: 601318.SH) from 2011/01/01~2018/11/18 :

ts_code	trade_date	open	high	low	close	pre_close	change	pct_chg	vol	amount
601318.SH	20110104	56.85	57.60	56.50	56.91	56.16	0.75	1.34	245626.82	1400499.338
601318.SH	20110105	56.59	56.80	54.80	54.86	56.91	-2.05	-3.60	427554.25	2368305.774
601318.SH	20110106	54.94	54.94	51.45	52.59	54.86	-2.27	-4.14	947078.65	4971851.955

Fig. 4. Ping-an Finance Group stock data.

Then, I calculate average change price of next 3 days and named a new column '*ave_change*' as prediction target.

Experiment results shows below:

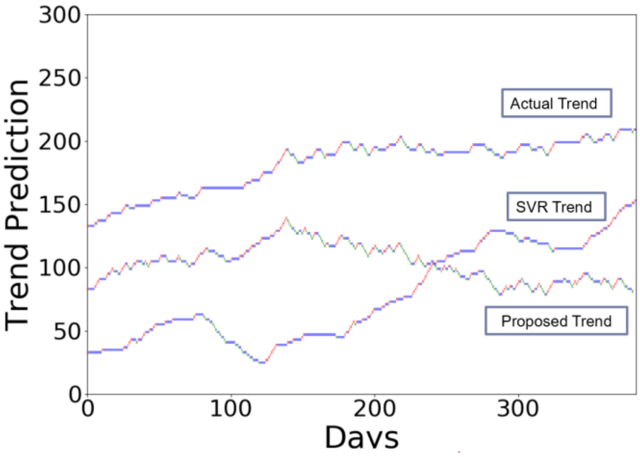


Fig. 5. Stock trend prediction of Proposed method and SVR

And then, I get 3 quantile value 0.33AC, 0.67AC , and convert ave_change price into category labels:

(1)If C < 0.33AC: label = -1 (Down)

If 0.33AC <C < 0.67AC: label = 0 (Even)

If C > 0.67AC: label = 1 (Up)

(2)Result_List = Real_List – Prediction_List ; if result = 0 , true , else false.

(3)Accuracy = 0 count / total count

After calculation, result shows below:

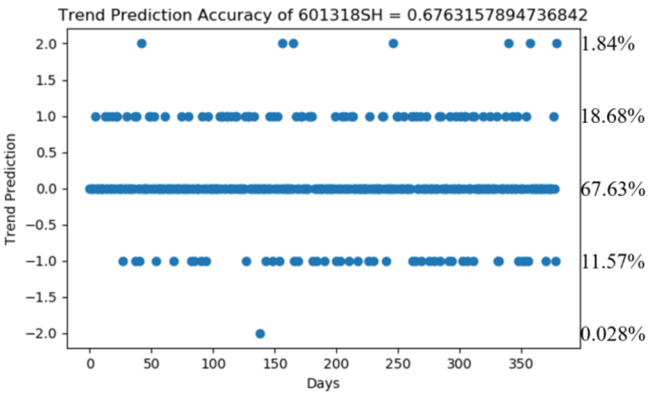


Fig. 6. Trend prediction Accuracy by Proposed Method

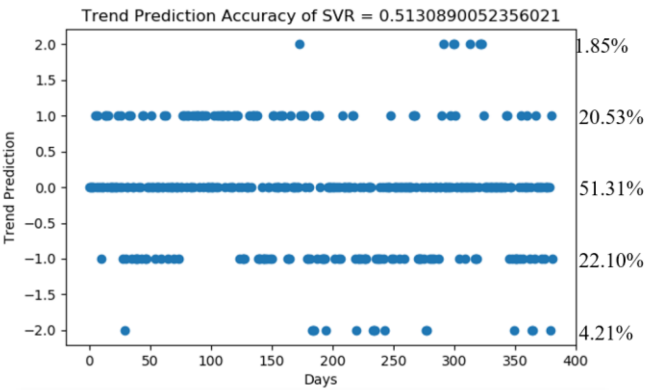


Fig. 7. Trend prediction Accuracy by SVR

Points in the graph means prediction is true or false.

(Point value = prediction trend result – real trend result)

Y axis means point value: 0 is true, 1 and -1 is slightly wrong, 2 and -2 is totally wrong.

Proposed method accuracy is 65.78% while SVR is 40.31%. Random selection accuracy is 33%. Both have better performance than random selection and proposed method is the best.

Experiment 2: Feature Selection Effect

Objective:
Compare different feature combination and find optimized feature selection. Furthermore, evaluate feature dimension reduction's performance. Stock data is from China Vanke Co. Ltd. (code: 000002.SZ) from 2011/07/01~2018/11/18. And A-share composite stock data (code: 000001.SH). We establish the following 5 comparison models experiments to find the optimal combination.
Technique Index:

data_dea	data_macd	kdj_k	kdj_d	kdj_j	bias	rsi	roc	psy	ave_change
0.000000	0.000000	68.292683	68.292683	68.292683	0.000000	0.00	0.000000	100.0	0.046667
0.000374	0.000598	72.682927	70.926829	76.195122	0.172513	100.00	0.344432	100.0	0.020000
0.002224	0.005329	89.088575	79.529762	108.206202	1.105604	100.00	1.809955	100.0	-0.006667
0.003491	0.004946	94.634146	85.803891	112.294658	0.656205	87.50	1.587302	75.0	-0.023333
0.003772	0.001324	92.948792	88.546720	101.752936	0.068462	56.25	1.026226	60.0	-0.090000

Macro Index (Shanghai Composite Stock price):

close_y	open_y	high_y	low_y	pre_close_y	change_y	pct_chg_y	vol_y	amount_y
2759.362	2767.834	2778.668	2752.966	2762.076	-2.714	-0.0983	92072878.0	1.054532e+08
2812.818	2770.940	2813.270	2770.940	2759.362	53.456	1.9373	121962260.0	1.438187e+08
2816.355	2812.722	2818.141	2799.110	2812.818	3.537	0.1257	110124224.0	1.286545e+08
2810.479	2811.815	2811.815	2780.728	2816.355	-5.876	-0.2086	103104393.0	1.220889e+08
2794.267	2813.193	2825.123	2793.892	2810.479	-16.212	-0.5768	116512143.0	1.400651e+08

Discussion:

No.	Model	Dimension	Feature combination	Accuracy
1	M1	9	1.Stock data	60.85%
2	M2	20	1.Stock data + 2. Technique index	60.55%
3	M3	18	1.Stock index + 3. Macro Index	66.36%
4	M4	29	1.Stock data + 2. Technique index + 3.Macro Index	61.46%
5	M5	18	1.Stock data + 2. Technique index + 3.Macro Index + 4.Auto-encoder	64.83%

Table. 4. Trend prediction accuracy after feature selection

1.M2 didn't increase accuracy compared with M1, but it causes more feature which will cause more computing cost. Technique index has little effect on trend prediction. Technique index is redundant to fundamental index in this task.
2.M3 is better than M1. Macro index has big effect. We can say individual stock price has positive correlation with whole stock price.
3.From M4 we find accuracy decreased. With too much features cause gradient vanishing and data redundancy.
4.Compare M5 with M4, we find using Auto-encoder for dimension reduction can solve gradient vanishing and improve prediction performance.

Experiment 3: Sentiment Analysis Performance Evaluation

Objective:
Compare with and without Sentiment analysis. I have established text sentiment analysis tool which is effective on finding sentiment of financial news. And then I calculated sentiment index of each day by sentiment news sentiment statistics. Stock data is China Composite Stock index. Shanghai Composite Stock Index(000001.SH) from 20160401 ~ 20171001. News data is from a stock forum named GUBA news data related to Shanghai composite index from 20160401~20171001.

After sentiment analysis, each news data will have a result like below:
2016-04-01 21:09:51,一周资金路线图：逾千亿资金出逃 沪股通连续买,0.978317175663178,1
2016-04-01 21:15:11,1日影响市场重大消息一览（附点评和后市策略）,0.3197187585638218,0
2016-04-01 21:15:25,标普下调中国三家政策性银行评级展望至负面,0.0393834433524691,-1
2016-04-01 21:33:12,创业板指3月涨幅全球NO.1！横空出世一批大牛股,0.898717404149230 67,1
2016-04-01 23:07:06,河北廊坊房地产调控措施出台：非本地户籍家庭限,0.9701718532792052,0
2016-04-02 00:32:48,五大银行高管连买房的税都交不起 怪不得都排队,0.008924845634290524,-1
2016-04-02 00:41:38,国家队持股全景图：8000亿家底大曝光 制造业最,0.9535020980791172,1
2016-04-02 00:51:30,水皮：世界同坐在一条船上 而且是一条和平之船,0.9806236622986924,1
2016-04-02 01:03:39,去年净利润逾6800亿元 股灾挡不住公募基金赚钱,0.7393016847392437,0
2016-04-02 02:46:00,A股市场迎来小阳春 分级基金B份额涨势喜人,0.9807725696686812,1
2016-04-02 03:43:49,证监会：新三板挂牌公司转板制度正在研究,0.9164444510908928,0

Then convert them into sentiment label for each day:

	trade_date	sentiment_label
0	20160401	24
1	20160402	25
2	20160403	8
3	20160404	17
4	20160405	17

Final experiment results showing below:

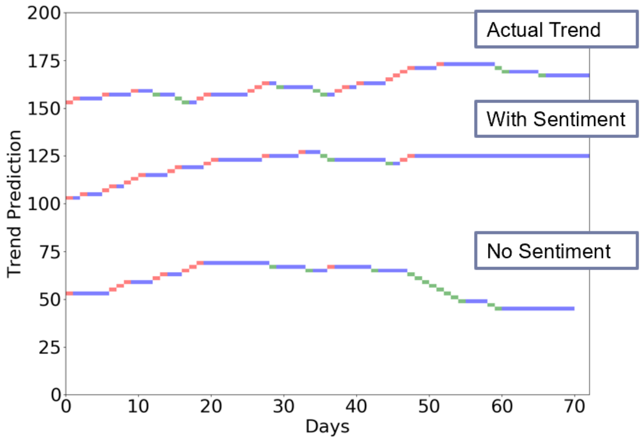


Fig. 8. Stock trend prediction with Sentiment analysis and without

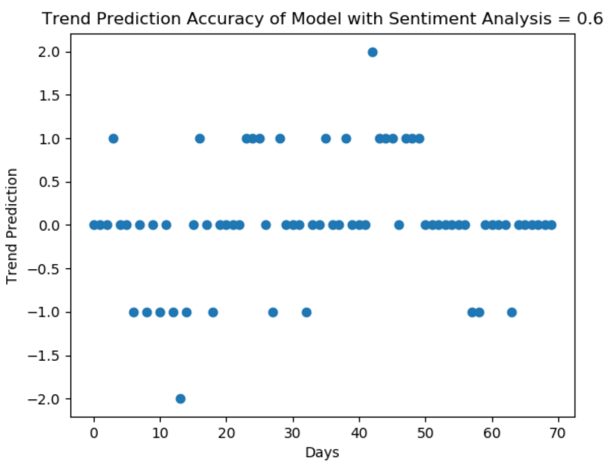


Fig. 9. Trend prediction Accuracy with sentiment analysis

ISBN: 978-988-14048-5-5
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

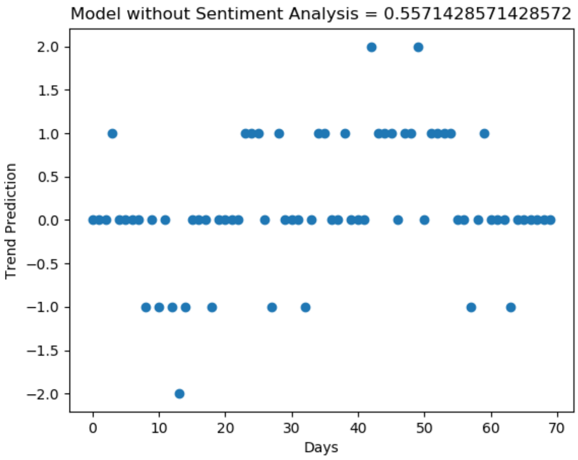


Fig. 10. Trend prediction Accuracy without sentiment analysis

In first time experiment, prediction accuracy increased to 62.5% from 54.28% with sentiment analysis. I repeat the experiment for 6 times. Generally, Sentiment analysis has positive effect.

No.	No Sentiment	With Sentiment
1	54.28%	62.5%
2	56.13%	56.94%
3	52.09%	58.88%
4	53.66%	48.62%
5	48.28%	57.45%
6	54.25%	66.32%

Table. 5. Sentiment analysis effect experiment results.

Experiment 4: Stability and Generality of the propose method

Objective: Evaluate stability of proposed method on different market industry fields

Method: Stock data of different individual companies are used to do experiments separately and check the variance of prediction to find whether it’s a stable method.

Results showing below:

Stock Num	Acc	Stock Num	Acc
000002万科地产	62.96%	600000浦发银行	71.65%
600028中国石化	68.15%	601166兴业银行	68.60%
600036招商银行	63.75%	601901方正证券	76.82%
600519贵州茅台	53.42%	600104上汽集团	65.58%
601288农业银行	51.84%	601088中国神华	70.10%
601318中国平安	65.52%	601668中国建筑	57.48%
601398工商银行	62.36%	000333美的集团	65.00%
601628中国人寿	64.73%	601998中信银行	71.42%
601857中国石油	67.26%	601939建设银行	61.31%
601988中国银行	66.56%	601857中国石油	70.26%

Table. 6. Trend prediction accuracy of different individual stocks.

V. Conclusion

From the research, we proposed an accurate stock trend prediction method and convinced market emotion is a very important factor influencing stock market and can help improve prediction accuracy.

Proposed novel method with feature compression and sentiment analysis for stock trend prediction and it improved accuracy than SVR by about 20%.

Market emotion was caught by sentiment analysis and it is a very important factor influencing stock market and improve prediction accuracy by 5%.

Recurrent Neuron Network with LSTM (Long Short-term Memory) can handle financial time series data better than traditional time series prediction method.

My method is stable in different tasks like individual stock or composite stock prediction.

References

[1] Sun, Haonan, et al. "Stacked Denoising Autoencoder Based Stock Market Trend Prediction via K-Nearest Neighbour Data Selection." International Conference on Neural Information Processing. Springer, Cham, 2017.

[2] Pang, Xiongwen, et al. "An innovative neural network approach for stock market prediction." The Journal of Supercomputing (2018): 1-21.

[3] Huynh, Huy D., L. Minh Dang, and Duc Duong. "A New Model for Stock Price Movements Prediction Using Deep Neural Network." Proceedings of the Eighth International Symposium on Information and Communication Technology. ACM, 2017.

[4] Akita, Ryo, et al. "Deep learning for stock prediction using numerical and textual information." Computer and Information Science (ICIS), 2016 IEEE/ACIS 15th International Conference on. IEEE, 2016.

[5] Siami-Namini, Sima, and Akbar Siami Namin. "Forecasting Economics and Financial Time Series: ARIMA vs. LSTM." arXiv preprint arXiv:1803.06386 (2018).

[6] Song, Yuan. Stock Trend Prediction: Based on Machine Learning Methods. Diss. UCLA, 2018.

[7] Bai, Shaojie, J. Zico Kolter, and Vladlen Koltun. "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling." arXiv preprint arXiv:1803.01271 (2018).

[8] Yu, ShuiLing, and Zhe Li. "Forecasting Stock Price Index Volatility with LSTM Deep Neural Network." Recent Developments in Data Science and Business Analytics. Springer, Cham, 2018. 265-272.

[9] Zhan, Xukuan, et al. "Stock Price Prediction Using Time Convolution Long Short-Term Memory Network." International Conference on Knowledge Science, Engineering and Management. Springer, Cham, 2018.

[10] Vargas, Manuel R., Beatriz SLP de Lima, and Alexandre G. Evsukoff. "Deep learning for stock market prediction from financial news articles." Computational Intelligence and Virtual Environments for Measurement Systems and Applications (CIVEMSA), 2017 IEEE International Conference on. IEEE, 2017.

[11] Hu, Ziniu, et al. "Listening to chaotic whispers: A deep learning framework for news-oriented stock trend prediction." Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining. ACM, 2018.