An NLP-PCA Based Trading Strategy On Chinese Stock Market

Zhao Liu[1,a], Huiying Zhu[2,b,*] and Tham Yew Chong[3,c]

[1 2 3] National University of Singapore, Institute of System Science, Singapore 119615

[a]liuzhao@u.nus.edu, [b]zhuhuiying@u.nus.edu, [c]erictham@nus.edu.sg

*Tham Yew Chong

Keywords：Chinese stock analysis, Financial sentiment analysis, NLP, PCA.

Abstract. The stock market is a barometer of a country's economy. However, the stock market is significantly affected by policies, news, and public opinion, and it is prone to volatility. Compared with the already mature financial securities market in foreign countries, China's stock market is still in the exploratory stage. There are more individual stock speculators in short-term trading. They will search for news through various channels to make decisions. Behavioral finance has created a theoretical basis for the mining of stock reviews. The rise of technologies such as text mining, machine learning, and time series models has made stock review mining possible. In this paper, we extract 28 kinds of financial sentiment features from thousands of Chinese news and social media outlets by NLP. Compared with popular methods where they only use positive or negative sentiment to predict stock price, we find out five more specific information categories from the news, which is SSE rising or dropping expectation, macro-public finance, bond sentiment, bond price forecast, the buzz on bond, rate and stock index. Finally, we predict stock price using these primary factors, providing a specific predictive ability to the stock market trend.

1. Introduction

1.1 Literature Review

As the size of stock market investors continues to grow, scientific and detailed research and forecasting of the Chinese stock market has become an increasingly important issue. Along with the development of the stock market, several stock market theory research emerges endlessly, and the market indicators are also varied. Experts and scholars from all walks of life try to analyze the stock market from different angles. We conclude the main achievements in table 2-1.

Table 1. Conclusion of research on a stock prediction by analyzing social media

Author	Data	Research Method	Conclusion
Tetlock [1]	The Wall Street Journal	PCA and build media pessimistic index and autoregression	A high pessimistic index can lead to a decline in stock prices, and an abnormal index points to a short high transaction volume.
Schumaker	news articles from Yahoo Finance	BoW, NER,SVM	Forecast the stock price after 20 minutes of the news release, the result is better than linear regression
Fung	Reuters Market 3000 Extra	Piecewise segmentation algorithm, hierarchical clustering algorithm, SVM	Market simulations based on forecasts from news categories can be profitable.
Zimbra [2]	Wal-Mart-related Web forum	Topic and Sentiment Analysis, Regressions	The emotions and topics reflected in the forum comments can predict the stock price
Koski [3]	RagingBull and Yahoo message boards	Cross-sectional and Time-Series Granger-causality regressions	Online commentary and stock market volatility have Granger causality

1.2 Financial Sentiment

Financial sentiment analysis [4] is an area of behavioral nance which studies the effect of market psychology on market practitioners and its impact on the market. In academic literature, sentiment has been broadly understood from two angles - optimism and over-confidence. Optimism occurs when investors over-estimate future returns, while over-confidence under-estimates future returns volatility. The psychological definition of negative emotions refers to the feelings that individuals are affected by internal or external factors that are not conducive to the individual's continued work or thinking. In behavioral finance, negative investor sentiment is expressed as a pessimistic expectation of the future. Investors in a negative mood will become conservative, which will reduce the expectation of future asset returns and increase the expectation of risk. In the stock market, investors with negative emotions can accept small spreads and buy, and the number of stocks they are willing to buy is small. In the face of falling stock markets, negative sentiment will cause investors to quickly sell shares in the short term, causing stock prices to fall.

Thomson Reuters MarketPsych Indices (TRMI) are available at MarketPsych Data LLC, including asset-class, and market levels for daily frequency. TRMI evaluates the data from news and social media contents and identifies both, the entities from the article (companies, countries, currencies, commodities, etc.) and macroeconomic, financial and emotions related words that are relevant for the entity. Subsequently, the volume and tone of phrases and words are converted into measurable variables. TRMI data are provided for three source sets: traditional news, social media, and the combined content.

1.3 Prospect Theory and Social Media Sentiment

Recently, an increasing amount of literature has been devoted to obtaining financial sentiment from social media. Financial sentiment from social media can be seen in the context of the Prospect Theory in Tversky and Kahneman (1979) which is a seminal paper in behavioral nance. As formula (1) shown, prospect theory states that people go through two different stages when deciding between choices that carry uncertainty. First, there is the editing phase. Here people simplify the complicated decisions into simpler ones by just making a belief of gains versus losses per option. In the second valuation phase, people examine the edited options and act on the one they believe has the highest value.

$$\mathrm{V} = \sum_{i=1}^{n} \pi(p_i) v(x_i) \tag{1}$$

- $\pi(p_i)$ as beliefs function
- $\mathrm{v}(x_i)$ as value fuction

The success of social media and Big Data allows distinguishing between the two phases. Investors reflect their beliefs on the stock market by writing about it on social media. This is the first editing phase. However, they might not necessarily act on their feelings; they may have just lost their job or need to save their money to keep their kids in college. This gives insights into the second valuation phase. These two phases are distinct although dependent on each other.

2. Data preprocessing

2.1 data description

In our study, sentiment scores are obtained from the news headlines and stock forums from the Internet. We also collect daily stock price as a research target. This data is between the period 01 January 2017 and 30 April 2018. The news headlines come from the major Chinese news media including SINA, Yunvs, Eastmoney, jqka, Hexun and Caijing. These websites carry more objective economic releases and fundamental company news reports. These news and stock forum discuss many kinds of topics, such as bond, stock, central bank and so on. Moreover, people express their

attitude, like surprise, Optimism, trust, fear and so on. We use NLP technology to extract 28 significant features from these raw data.

2.2 Chinese Natural Language Processing

Chinese Natural Language Processing (NLP) [5] is used to quantify the sentiment and topics from forums and news headlines. There are a few steps to carrying out the NLP analysis successfully:

- Filtration of noisy posts
- Word segmentation
- Removal of stop words
- Name Entity Recognition for weights

The filtration of noisy posts in the forum to remove irrelevant comments is necessary given the laissez-faire nature of social media. These noisy posts constitute almost 50% of the total posts and are generally identified by trivial word features and the shortness of the forum. Further, a major difference between English and Chinese NLP is the need for word segmentation as Chinese words are not separated from one another by space. This can be done through
machine learning techniques like conditional random fields referenced in Tseng et al. (2005). Some open source NLP tools available for this purpose include the Jieba [6] and the excellent Stanford NLP program. After the Chinese phrases are extracted by word segmentation, typical stop words are removed referencing a standard corpus. Examples of stop words include '是', '以' which correspond to English words like 'is' and 'are'. The removal of stop words is necessary as they impair the learning process of the classification algorithm next described.

2.3 Named Entity Recognition (NER)

In this study, related stock symbols and names, and market words are identified as the named entities [7] which are maintained in a separate corpus. A regular expression search through each news and forum is done to filter out these entities. The named entities are weighted to reflect the relevance of the news and forum content to the market index sentiment.

Strong words are over-weighted and weak words have less weighting. We only deploy the variable weighting for extremely powerful words such as "terrified" (2) versus weak words such as "cautious" (0.5). We also use Maximizing and Minimizing modifiers to change word strength, such as "A little afraid" (0.5) versus "a lot of fear" (2).

3. Empirical Analysis

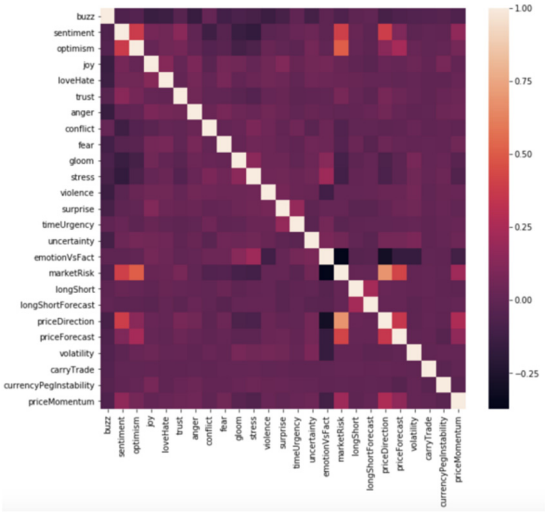
Experiments data is from TRMI(Thomson Reuters MarketPsych Indices), consisting of 28 features. Before achieving sentiment scores, some of the data are unstructured which have similar means before being extracted from the news or social media context, which need to be further explored by the correlation matrix first.

3.1 Factor Analysis on News Topics

3.1.1 Correlation

Correlation matrix means how a change happened on a sentiment feature related to another one. Low correlations mean the pairs are independent, and high correlation means two factors can represent each other on some extent. After checking the results, what is shown is that some features relationship are stronger than others (Ratesbuzz&Bondbuzz=0.41, stock index stress&stock index market-risk=0.56, etc).

Table 2. Corelation Matirx



3.1.2 Bartlett test

Then the following experiment, KMO and Bartlett's Test [8] is undertaken on standardized processing of the data to check whether there are latent features. From the table below what can be seen is, KMO value reached 0.662 which> 0.5, Bartlett ball test significance < 0.01 illustrates indicator variables being studied has a strong correlation and factor analysis is necessary. This means Factor analysis is applicable to those sentiment features.

Table 3. KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		0.662
Bartlett's Test of Sphericity	Approx. Chi-Square	3209.271
	df	435
	Sig.	.000

3.1.3 Factor Analysis

The basic idea of factor analysis [9] is to transform many variable indexes into a few comprehensive indexes utilizing dimensionality reduction, so as to replace the original variables for multivariate statistics. In general, the final synthesis indexes are called as the factor analysis, and each factor is a linear combination of some original variables, and there is no correlation between these factors. So the factors have a more significant representative significance compared with the original variable.

With the use of factor analysis, researchers can find some critical factors among many variables involved in a certain problem, thus effectively carrying out quantitative analysis on a large number of statistical data, mining the internal relations between various variables, and improving the research work. Scree plot shows after 5 factors, scree plot become more smoothing which is not significant. So top 5 factor are selected for later using.

Table 4. Factor Analysis for news sentiment

Number	EigenValue	Percent	Cum percent
1	4.065	13.56%	13.56%
2	2.046	6.82%	20.37%
3	1.953	6.51%	26.88%
4	1.573	5.24%	32.12%
5	1.433	4.78%	36.90%
6	1.347	4.50%	41.39%

The first factor is similar to the Market-Risk factor according to CAPM. That means Chinese market is unlike other stock markets, which mainly based on 3-5 factors and can explain most of the variance. That also states Chinese financial market players is motivated more by nature, which corresponding to large based individual investors in China.

Table 5. Factor Loadings Matrix（Loadings>0.3）

Name	Factor1	Factor2	Factor3	Factor4	Factor5
StockPriceMarketRisk	0.9305	*	*	*	*
StockIndexSentiment	0.8149	*	*	*	*
StockIndexOptimism	0.7090	*	*	*	*
StockIndexTrust	0.3333	*	*	*	*
StockIndexPriceForecast	0.3095	*	*	*	*
debtDefault	*	0.9360	*	*	*
bondSentiment	*	*	0.948	*	*
bondOptimism	*	*	0.425	*	*
bondPriceForecast	*	*	*	0.995	*
bondPriceDirection	*	*	*	0.405	*
ratesBuzz	*	*	*	*	0.694
bondBuzz	*	*	*	*	0.693
stockIndexFear	-0.4820	*	*	*	*
centralBank	*	-0.782	*	*	*
stockIndexPriceDirection	0.5278	*	*	*	*
stockIndexBuzz	*	*	*	*	*
stockIndexStress	-0.6188	*	*	*	*

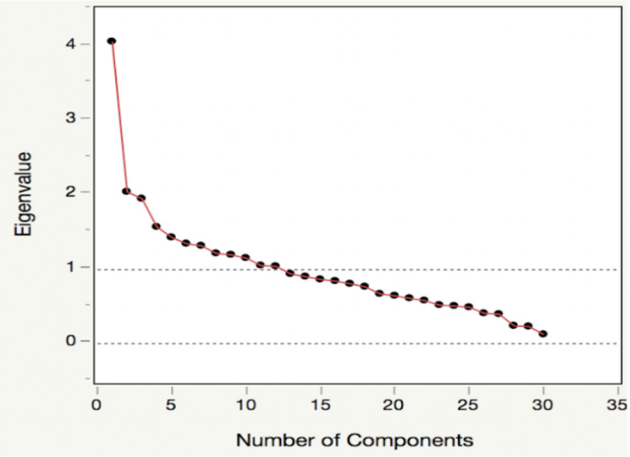


Fig. 1. Scree Plot

Top 8 variables eigenvalue above 1 and top 5 variances from scree plot gives sharp slope in those. Based on this, the top 5 factors were selected first and to be named as followings.PCA1: Stock index rising or dropping related news, PCA2: Central bank financing news, PCA3: bond sentiment news, PCA4: bond price forecasting news, PCA5: buzz on bond, interest rate, and stock index.

3.2 Stationarity Test for Time Series

3.2.1 Unit Root test for news

Check time series for PCA sentiments and log return based on following unit root tests with to make sure their time series process is significant. The unit root test is about the stationarity test of variables. Only the stationary variables can be estimated by OLS model [10] or VAR model [11].

Table 6. Unit root test for PCA from news and Ln(Return)

Varibale	Difference	(C,T,K)	DW	ADF	1%	5%	Conclusion
Factor1	0	(c,t,0)	1.997	-4.947	-3.771	-3.419	I(0)*
Factor2	0	(0,t,0)	1.983	-12.304	-2.570	-1.941	I(0)*
Factor3	0	(0,t,0)	2.004	-11.980	-2.570	-1.941	I(0)*
Factor4	0	(c,t,0)	1.990	-13.218	-3.977	-3.419	I(0)*
Factor5	0	(c,t,0)	1.989	-12.760	-3.977	-3.419	I(0)*
LnReturn	0	(0,t,0)	2.003	-15.66	-2.57	-1.941	I(0)*

The table shows there is no intercept but the only trend for the series and difference are 0 means there the original series is already stationary. The Durbin-Watson [12] are all close to 2, which means all residuals obey the normal distribution and suit assumption of time series. All the PCA variables and lnReturns are significant.

3.2.2 Vector autoregression(Var) model [13] to check Lag order

Check which is a suitable period for the series. Based on the standard to choose the minimum AIC [14] from lag0 to lag3, it is shown that lag 2nd is the best that shows the minimum AIC 10.334.

Table 7. VAR Lag Order Selection Criteria

Lag	AIC
0	10.819
1	9.868*
2	9.969
3	9.974
4	10.023

3.2.3 Autoregression Graph to further check stationary

Then, the stability of the VAR model is tested. The AR root graph of VAR is directly used to verify whether it satisfies the stability. If the reciprocal of all roots of the equation are in the unit circle, it indicates that the VAR system is stable. From given results and all points are in the circle, which means the VAR model is stable.

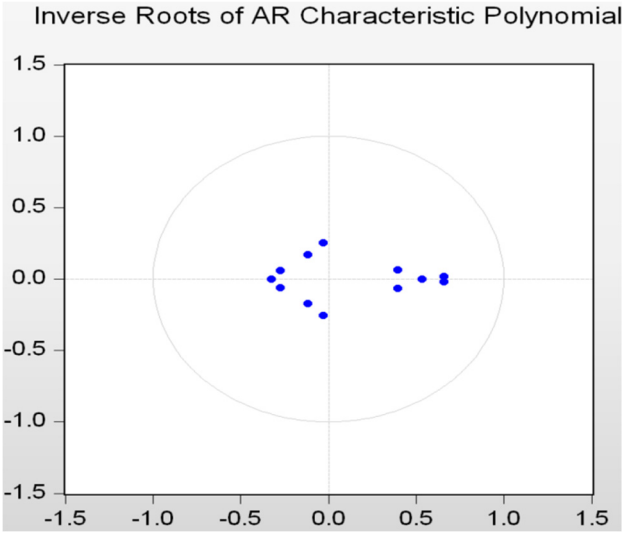


Fig. 2. PCA variables influence In return root

3.2.4 Pairwise Granger Causality Tests

Granger causality test [15] can be used to analyze whether there is a causal relationship between two variables. Granger causality test also needs to determine the lag order, which generally needs to be confirmed by the VAR model of the AIC criterion. So the lag order is also 2 in Granger Causality Tests.

The left side is the null hypothesis, and the right side is the p-value. If the p-value is less than 0.05, the null hypothesis will be rejected.

Table 8. Granger Causality Test

Direction	F-value	P-value	Causality
Factor1→Daily Return	1.57	0.21	no
Daily Return→Factor1	19.02	1.E-08	yes
Factor2→Daily Return	0.61	0.54	no
Daily Return→Factor2	4.14	0.02	yes
Factor3→Daily Return	0.20	0.82	no
Daily Return→Factor3	2.62	0.07	no
Factor4→Daily Return	1.22	0.29	no
Daily Return→Factor4	0.6	0.54	no
Factor5→Daily Return	0.02	0.98	no
Daily Return→Factor5	0.15	0.86	no

This shows Daily Return is the cause of Factor 1. Daily Return is the cause of Factor 2. The results show that there is a close causal relationship between two main factors and SSEC performance.

3.3 Regression

3.3.1 train model

We use these five factors to train model, and target variable is Daily Return. It is defined as for formula 2.

$$\text{Daily Return} = \log \frac{Stock\ Close\ Price_t}{Stock\ Close\ Price_{t-1}} \quad (2)$$

We find that the performance of regression using the combination of the factors (like F1 *F2 * F3 * F4 * F5) is much better than linear regression between Daily Return and individual factor [16], F1, F2, F3, F4, F5. The regression result is shown in table 4-1.

Table 9. Regression model result

Variable	Coefficient	Std.Error	t-Statistic	Probability
F1	0.0017642	2.507e-04	3.519	0.000508
F5	0.002292	4.700e-04	2.438	0.015415
F1*F3*F4	-0.0009456	1.778e-04	-2.180	0.030118
F1*F2*F3*F4*F5	0.00019108	4.692e-05	2.036	0.042698
Intercept	0.0003418	4.547e-04	0.376	0.707243

In the regression model, R-square is 0.2107, the P value of F is 0.0001634 which is significant. And t values are significant too. As a result, the fitness of the model is good. The final regression model is shown in formula 3.

$$Y = 0.0003418 + 0.0017642 \times F1 + 0.002292 \times F5 - 0.0009456 \times F1 \times F3 \times F4 + 0.00019108 \times F1 \times F2 \times F3 \times F4 \times F5 \quad (3)$$

3.3.2 prediction

According to the obtained principal component regression model, we use Daily Return from April 1st to April 30th, 2018 to test the prediction results. The prediction effect is as follows:

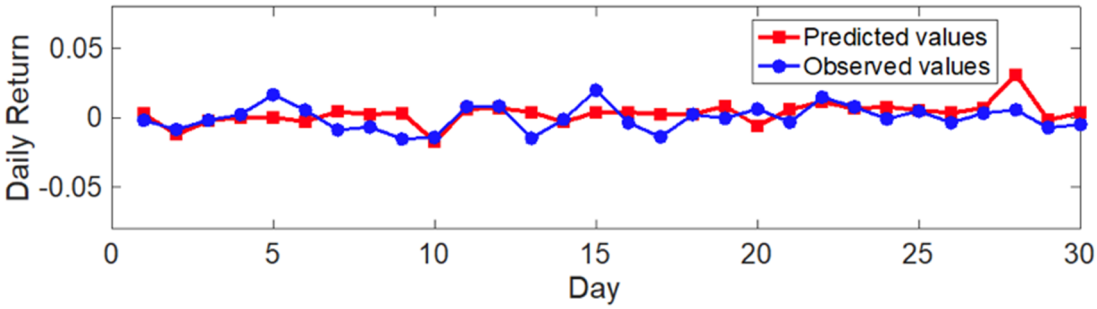


Fig. 3. line plot between day and Daily Return for predicted and Observed values.

The predicted Daily Return and actual Daily Return are nearly overlapped and MSE is 0.01, which means the prediction performance is satisfied.

4. Conclusion

Most researches were undertaken on sentiments of individual investors and the cause to returns about Chinese stock market mainly have three issues. First, most researches simply divided emotions into positive or negative from context, which ignores the inner quantitative relationship between the same polarity[17]. Second, up till now most research focusing on Chinese stocks and sentiments just crawled from limited information sources, making the information not diverse and consisting many unstructured context noise that causing a narrow sampling or range sampling error. Third, most researches ignore appraising qualitative information that which categories are more sensitive to the stock market.

In this research, with multiple methods and statistical test, we finally found that certain category news that has best predictive power is factor 1: Stock index rising or dropping related news, factor 2: Central bank and government finance news and factor 3: bond sentiment news. After checking the predictive power with news factors for returns with a regression, results show the R-square is nearly 20% and the p-value is significant, which can give an instructive perspective to household investors that news which mentioned stock index changing, central bank and government finance and bond-related that most linked to returns on SSEC.

This paper is creative in following aspects. First, it based on sentiment data from Thomson Reuters MarketPsych Indices, which are widely derived from 2,000 premium news and 800 financial social media sources. Then, sentiment data are extracted with Nature Language Processing methods, which can provide scope to specific topics extraction. Sentiment score is standardized after extraction from context which is quantified. Third, the factor analysis method is used to make dimension reduction and explore their latent relationships.

Meanwhile, this research also has limitations. For example, due to the sample size and feature amounts, we only use regression to avoid overlapping, that is why final result R-square of regression is not high to give a perfect explanation by sentiments or topics arising from the news. Also, the result is only built from news sources data which show a stronger correlation in Granger Causality Test, which might lose the possibility of a combination of news and social media from both sides. However, it is still meaningful to provide a heuristic method for household investors to take it for reference.

References

[1] Tetlock, P.C., Giving content to investor sentiment: The role of media in the stock market. The Journal of Finance, vol.62, pp.1139-1168, 2007.

[2] Zimbra, D., Fu, T., and Li, X., Assessing public opinions through Web 2.0: a case study on Wal-Mart. ICIS 2009 Proceedings, p.67, 2009.

[3] Koski, J. L., Rice, E. M., & Tarhouni, A, Noise trading and volatility: Evidence from day trading and message boards. Available at SSRN 533943, 2004.

[4] Devitt, Ann, and Khurshid Ahmad. "Sentiment polarity identification in financial news: A cohesion-based approach." Proceedings of the 45th annual meeting of the association of computational linguistics. 2007.

[5] Che, Wanxiang, Zhenghua Li, and Ting Liu. "Ltp: A Chinese language technology platform." In Proceedings of the 23rd International Conference on Computational Linguistics: Demonstrations, pp. 13-16. Association for Computational Linguistics, 2010.

[6] Day, M.Y. and Lee, C.C., August. Deep learning for financial sentiment analysis on finance news providers. In 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM) pp. 1127-1134, 2016.

[7] Nadeau, David, and Satoshi Sekine. "A survey of named entity recognition and classification." Lingvisticae Investigationes 30, no. 1, 2007.

[8] Tobias, S. and Carlson, J.E.. Brief report: Bartlett's test of sphericity and chance findings in factor analysis. Multivariate Behavioral Research, 4(3), pp.375-377, 1969.

[9] Thompson, Bruce. "Factor analysis." The Blackwell Encyclopedia of Sociology, 2007.

[10] Horrace, W.C. and Oaxaca, R.L., Results on the bias and inconsistency of ordinary least squares for the linear probability model. Economics Letters, 90(3), pp.321-327, 2006.

[11] Dungey, M. and Pagan, A., A structural VAR model of the Australian economy. Economic record, 76(235), pp.321-342, 2000.

[12] Nerlove, Marc, and Kenneth F. Wallis. "Use of the Durbin-Watson statistic in inappropriate situations." Econometrica: Journal of the Econometric Society: 235-238, 1996.

[13] Stock, James H., and Mark W. Watson. "Vector autoregressions." Journal of Economic perspectives 15, no. 4,101-115, 2001.

[14] Akaike, Hirotugu. "Factor analysis and AIC." In Selected Papers of Hirotugu Akaike, pp. 371-386. Springer, New York, NY, 1987.

[15] Diks, C. and Panchenko, V., A new statistic and practical guidelines for nonparametric Granger causality testing. Journal of Economic Dynamics and Control, 30(9-10), pp.1647-1669, 2006.

[16] Gebka, B. and Wohar, M.E., Causality between trading volume and returns: Evidence from quantile regressions. International Review of Economics & Finance, 27, pp.144-159, 2013.

[17] Eric Tham, Trusting the Social Media. 31st Australasian Finance and Banking Conference 2018