Scientific Research Publishing

Journal of Data Analysis and Information Processing, 2019, 7, 295-305
https://www.scirp.org/journal/jdaip
ISSN Online: 2327-7203
ISSN Print: 2327-7211

Predicting the Stock Price Movement by Social Media Analysis

Sitong Chen[1*#], Tianhong Gao[2*], Yuqi He[3*], Yifan Jin[4*]

[1]Capital University of Economic and Business, Beijing, China
[2]Beijing University of Posts and Telecommunications, Beijing, China
[3]Sichuan University, Chengdu, China
[4]Zhejiang University, Hangzhou, China
Email: #15001099795@163.com

How to cite this paper: Chen, S.T., Gao, T.H., He, Y.Q. and Jin, Y.F. (2019) Predicting the Stock Price Movement by Social Media Analysis. *Journal of Data Analysis and Information Processing*, **7**, 295-305.
https://doi.org/10.4236/jdaip.2019.74017

Received: September 6, 2019
Accepted: November 17, 2019
Published: November 20, 2019

Abstract

Prediction of stock trend has been an intriguing topic and is extensively studied by researchers from diversified fields. Machine learning, a well-established algorithm, has been also studied for its potentials in prediction of financial markets. In this paper, seven different techniques of data mining are applied to predict stock price movement of Shanghai Composite Index. The approaches include Support vector machine, Logistic regression, Naive Bayesian, K-nearest neighbor classification, Decision tree, Random forest and Adaboost. Extracting the corresponding comments between April 2017 and May 2018, it shows that: 1) sentiment derived from Eastmoney, a social media platform for the financial community in China, further enhances model performances, 2) for positive and negative sentiments classifications, all classifiers reach at least 75% accuracy and the linear SVC models prove to perform best, 3) according to the strong correlation between the price fluctuation and the bullish index, the approximate overall trend of the closing price can be acquired.

Keywords

Social Media, Investor Sentiment, Machine Learning

1. Introduction

With the rapid development of economy-oriented society, investor sentiment has received more and more attention. The efficient market hypothesis has been at the core pillar of modern financial theory since the 1960s. According to Fama, in an efficient market, the price fully reflects all the information it can get [1].

*These are co-first authors, sorted alphabetically by last name.

However, financial markets are considered to be a complex non-linear system, and it is very challenging to predict stock prices in a technical way [2] [3] [4] [5]. Market anomalies were observed which contradict the EMH basic assumptions according to which the prediction of share prices should not be possible [6] [7] [8] [9]. In recent years, financial economists have been trying to study the financial behavior of investors from the perspective of human science, which has also spawned a new field of financial research—behavioral finance tracing back to the early 1990s [10]-[15]. The important branch with investor sentiment as the research object is gradually emerging as the information technology has witnessed an unprecedented boom. Single events (e.g., sport results, daylight saving anomaly) or continuous effects (e.g., weather effect, air pollution) influence people's emotions [16] [17] [18] [19]. The prediction of share returns based on mood states can be seen as market anomaly contradicting the efficient market hypothesis [20]. These mood-related anomalies can be explained by the misattribution bias according to which people make risky decisions depending on mood states [21]. The Affect Infusion Model (AIM) can explain the relationship between positive and negative mood states and the risk-taking tendency which postulates that people in positive mood rely on positive cues to make decisions [22] [23] [24] [25].

Considering individual emotion is a vague concept, previous research made significant progress on various sentiment techniques after tracking indicators of public mood directly from social media content, such as Facebook and Twitter feeds [26] [27] [28] [29] [30]. In a seminal work, harnessing the cross-validate time series, Bollen *et al.* compared the ability of two mood tracking tools, namely OpinionFinder and Google-Profile of Mood States, to detect the public response on daily Twitter feeds to Dow Jones Industrial Average during Presidential election and Thanksgiving day [31].

Scholars' research on sentiment analysis is not limited to processing text, but extends to machine learning and achieves high accuracy. Data mining techniques have been introduced for prediction of movement sign of stock market index by Leung *et al.* and Chen *et al.*, Schumaker *et al.* predicted the S&P 500 index through SVM technology and used four text eigenvectors to represent the emotional dimension of the entire text, with an accuracy of 58.2% in the prediction results [32] [33] [34]. Hassan, Nath, and Kirley proposed and implemented a fusion model by combining the Hidden Markov Model (HMM), Artificial Neural Networks (ANN) and Genetic Algorithms (GA) to make financial market behavior forecast [35]. Kumar & Thenmozhi collected five different approaches including SVM, Random forecast, Neural network, Logit and LDA to predict Indian stock index movement based on economic variable indicators [36].

In this paper, we aim to analyze individual sentiment by addressing the accuracy of using seven machine learning algorithms in classifying financial stock comments into positive as well as negative classes. Platform Eastmoney is China's most popular exclusive community for financial professionals with daily

average flow exceeding 200 million, making it the preferred platform for domestic investors to interact. We compare the accuracy of these classifiers using the feature model: unigram TF-IDF. We assess the effects of including public mood information on the accuracy of a "baseline" prediction model rather than proposing an optimal prediction model.

2. Methods and Materials

2.1. Methods

In terms of the methodology as shown in Figure 1, we totally proceed in three phases.

2.1.1. Data Preparation and Feature Engineering

In the first phase, after data pre-processing, including word segmentation, pause word removal and tokenization, we leverage the unigram TF-IDF metric, a feature for word importance in a document that takes the product of term frequency (TF) and inverse document frequency (IDF). TF-IDF for a certain term t is defined as the multiplication of TF(t) by IDF(t). TF measures how frequently a term (feature) occurs in a comment. Since every comment may have different length, it is possible that a term would appear much more times in long blogs than shorter ones. Thus, the term frequency is often divided by the length as a way of normalization. Normalized TF for a given term t is defined as (formula 1):

$$\mathrm{TF}(t) = \frac{n}{N} \tag{1}$$

where n = Numbers of term t occurs in the comments, N = Total numbers of the terms in the comments.

In contrast, IDF measures the importance of terms based on how frequently they appear across multiple comments. Intuitively, a term appears frequently in

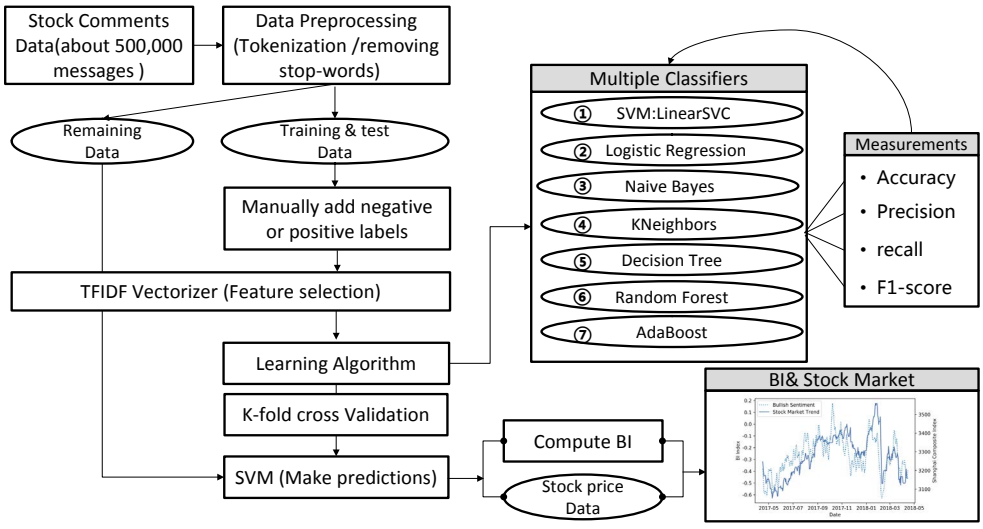

Figure 1. Diagram outlining the methodology overview.

a single comment is important and gets a high weight. However, if the term appears in many blog posts, then it becomes less discriminative; hence, IDF deemphasizes its weight. IDF for a term t is given by:

$$\mathrm{IDF}(t) = \log \frac{Q}{q} \tag{2}$$

where q = Numbers of comments with term t in it, Q = Total numbers of comments.

2.1.2. K-Fold Cross Validation with Multiple Machine Learning Algorithms

In the second phase, we deploy bag-of-words technique by manually sorting out positive and negative messages respectively. We apply K-fold cross validation to train the models where we divide the data into 5 splits and harness the first 80% for observations and the remaining 20% for test. We leverage multiple machine learning algorithms for analyzing the emotional polarity (Table 1).

Table 1. Machine learning classifiers overview.

Algorithms	Explanation
LinearSVC	Support vector machine (SVM) have two main categories: support vector classification (SVC) and support vector regression (SVR). SVM is a learning system using a high dimensional feature space. The main objective of support vector machine is to identify maximum margin hyper plane as the final decision boundary.
Logistic Regression	Logistic regression predicts the probability of an outcome that can only have two values (*i.e.* a dichotomy). A logistic regression produces a logistic curve, which is limited to values between 0 and 1. Logistic regression is similar to a linear regression, but the curve is constructed using the natural logarithm of the "odds" of the target variable, rather than the probability. Moreover, the predictors do not have to be normally distributed or have equal variance in each group.
Naive Bayesian	The Naive Bayesian classifier is based on Bayes theorem with the independence assumptions between predictors. Bayes theorem provides a way of calculating the posterior probability. A Naive Bayesian model is useful for very large datasets. Despite its simplicity, the Naive Bayesian classifier often does surprisingly well and is widely used because it often outperforms more sophisticated classification methods.
K neighbors Classifier	K nearest neighbors is a simple algorithm that stores all available cases and classifies new cases based on a similarity measure. KNN has been used in statistical estimation and pattern recognition already in 1970's as a non-parametric technique. A case is classified by a majority vote of its neighbors, with the case being assigned to the class most common amongst its K nearest neighbors measured by a distance function.
Decision Tree	Decision tree builds classification or regression models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes.
Random Forest	The training algorithm for random forests applies the general technique of bootstrap aggregating, or bagging, to tree learners.
AdaBoost	Adaptive boosting machine learning meta-algorithm used for enhancing performance and classifier accuracy by means of adding more weight to previously misclassified instances.

Since the particular problem is classification-based in nature, we test out the efficacy of each classifier. Accuracy and f-score are used to evaluate the performance of proposed models. Computation of these evaluation measures requires estimating Precision and Recall which are evaluated from True Positive (TP), False Positive (FP), True Negative (TN) and False Negative (FN). These parameters are defined in Equations (3)-(8). Since the prediction model has two dimensions, *i.e.*, true of false and negative or positive, we have the verification matrix (Table 2):

$$P_{pos} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} \tag{3}$$

$$P_{neg} = \frac{\mathrm{TN}}{\mathrm{TP}+\mathrm{FP}} \tag{4}$$

$$R_{pos} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \tag{5}$$

$$R_{pos} = \frac{\mathrm{TN}}{\mathrm{TP}+\mathrm{FN}} \tag{6}$$

where P is the precision of the model, R is recall, TP is the number of true positive, TN is the number of true negative, FP is the number of false positive, and FN is the number of false negative.

Taking the product of the two, we calculate the F1-score which is defined as:

$$\mathrm{F1} = 2 * \frac{P * R}{P + R} \tag{7}$$

$$\mathrm{A} = \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{FP}+\mathrm{TN}+\mathrm{FN}} \tag{8}$$

where F1 is the F1-score of the model and A is the accuracy of the model.

2.1.3. Bivariate Correlation Analysis for the Two Time Series

In the third phase, we select the model with the best accuracy and conduct the relationship between bullish sentiment and stock market trend. The bull/bear ratio is a market-sentiment indicator which reflects how these professionals are feeling about the market, and how they are likely advising their clients to invest based on those feelings. In this paper, we define the bullish indicator as:

$$BI^{*} = \ln\left[\frac{1+M^{\mathrm{Bull}}}{1+M^{\mathrm{Bear}}}\right] \tag{9}$$

Table 2. Positive and negative-accuracy verification matrix.

		Prediction		
		1	0	Total
Actuality	1	True Positive	False Negative	Actual Positive (TP + FN)
	0	False Positive	True Negative	Actual Negative (FP + TN)
	Total	Predicted Positive	Predicted Negative	TP + FP + FN + TN

Additionally, for bivariate correlation analysis, we usually have three methods of analysis which are the Pearson coefficient, Spearman coefficient and Kendall coefficient. Among them, we choose the Pearson correlation coefficient method measuring the linear relationship between two variables.

The Pearson correlation coefficient formula is as follows:

$$P(x,y)=\frac{E(xy)-E(x)E(y)}{\sqrt{E(x^2)-E^2(x)}\sqrt{E(y^2)-E^2(y)}} \tag{10}$$

2.2. Data

We perform analysis on the Shanghai Composite Index. All price data and comments data are drawn from the period between April 2017 and May 2018, totaling 266 trading days. Two main datasets were used.

2.2.1. Comments Data

Comments data is collected from the financial forum of Eastmoney (http://guba.eastmoney.com/) in CSV format, containing over 480,000 messages. Besides, we manually sort out about 5000 positive messages and 5000 negative messages.

2.2.2. Price Data

Daily split-adjusted stock price data of Shanghai Composite Index is collected via Tushare, a Python module which provides stock price data in dataframe format. We focus only on the closing price data.

3. Results and Discussions

As shown in Table 3 and Figure 2, the results indicate that the chosen algorithms are clearly indicators of both the positive and negative sentiments classifications with worst case accuracy of 75% and SVC yielded the best accuracy of 88%.

We choose SVM as the basic classification algorithm for our prediction model. We calculate the time series data of sentiment indicators through the bullish

Table 3. The test accuracy for each of the learning models.

Algorithms	Accuracy	Positive			Negative		
		Precision	Recall	F1-score	Precision	Recall	F1-score
LinearSVC	**0.8816**	**0.8805**	**0.8825**	**0.8815**	**0.8824**	**0.8801**	**0.8813**
LogisticRegression	0.8809	0.8791	0.8832	0.881	0.8828	0.8782	0.8804
Multinomial NB	0.8796	0.8821	0.876	0.879	0.8767	0.8832	0.8799
KNN	0.8201	0.8071	0.8404	0.8234	0.8336	0.7991	0.8159
DecisionTree	0.7994	0.8169	0.7711	0.7933	0.7833	0.8272	0.8046
RandomForest	0.8137	0.8662	0.742	0.7992	0.7739	0.885	0.8256
AdaBoost	0.7719	0.7973	0.7989	0.7654	0.8253	0.7594	0.7666

index. We combine it with the time series of stock prices in a single picture, (Figure 3). As shown in Figure 3, BI index and Shanghai composite index were selected as variables and Pearson coefficient was used for correlation test. The two series yielded statistically significant Pearson correlation coefficient of 0.689 (as shown in Table 4).

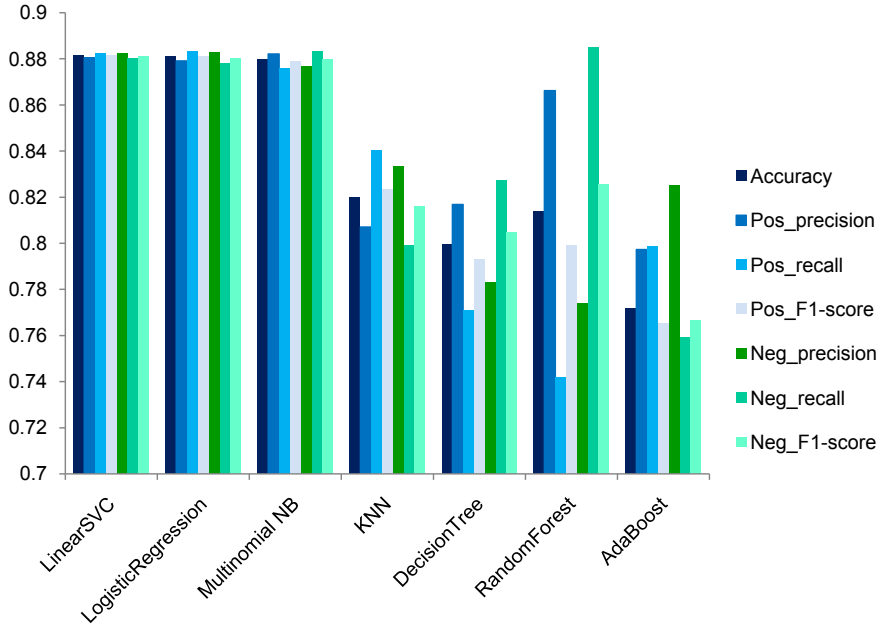

Figure 2. Diagram showing the test accuracy according to the four measurements.

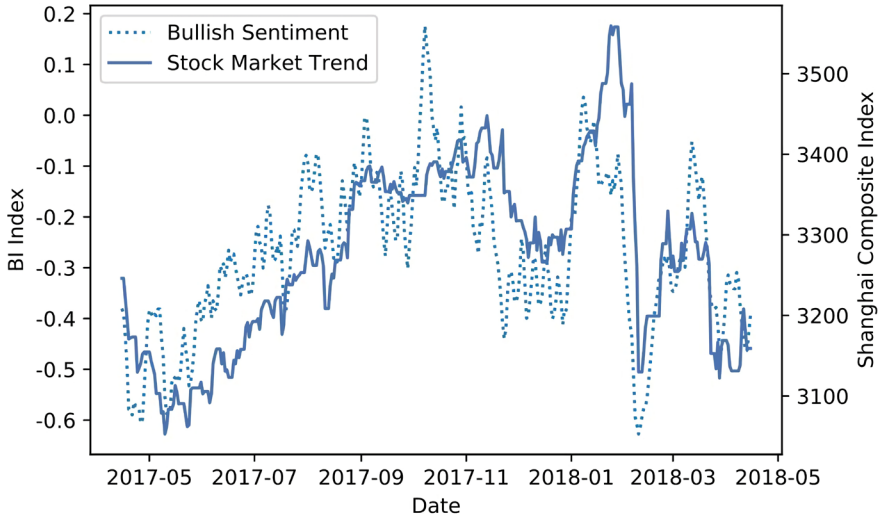

Figure 3. The two merged time series graph consisting of bullish sentiment and stock market trend.

Table 4. Correlation test result (**: Correlation is significant at the 0.01 level).

Pearson correlation (BI*/Stock Close Price)	Sig.	N
0.689**	0.000	392

4. Conclusions

This research focused on predicting the direction of stocks and stock price indices. Prediction performances of seven models namely SVM, Logistic regression, Naive Bayesian, KNN, Decision tree, Random forest and Adaboost are compared based on one year of historical data of Shanghai Composite Index from the Platform Eastmoney.

Experiments with continuous-valued data show that Adaboost model exhibits least performance with 77.2% accuracy and SVM with highest performance of 88.16% accuracy. SVM classifier has a better fitting degree for dichotomies. We divide emotions into positive emotions and negative emotions, so SVM is the most suitable classifier. Although these seven classification algorithms have achieved good fitting results, none of them is more than 90 percent accurate. On the one hand, Chinese words are more complex than English. On the other hand, most of natural language processing is mainly aimed at English, but not suitable for Chinese.

Further research will focus on extending the technical indicator's opinion about stock price movement as "highly possible to go up", "highly possible to go down", "less possible to go up", "less possible to go down" and "neutral signal" are worth exploring. This may give more accurate input to inference engine of the sentiment analysis algorithms. Besides calculating the correlation coefficient of the two time series, the research will be conducted to predict long term analysis of stock's quarterly performance involved the ARIMA model based on exogenous variables for empirical test.

Acknowledgements

Upon the completion of the thesis, we would like to take this opportunity to express my sincere gratitude to my supervisor, Professor Patrick Houlihan, who has given us important guidance on the thesis. Without his help and encouragement, our thesis would have been impossible. Besides his help with our thesis, he has also given us much advice on the methods of doing research, which is of great value to our future academic life.

We are also obliged to the authors in the references whose thesis have broadened my scope of vision in data science and help us lay a necessary foundation for the writing of the thesis.

Last but not least, we would like to express our gratitude to all the friends and family members who have previously offered their help.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Fama, E.F. (1970) Efficient Capital Markets: A Review of Theory and Empirical

Work. *The Journal of Finance*, **25**, 383-417. https://doi.org/10.2307/2325486

[2] Hommes, C.H. (2001) Financial Markets as Nonlinear Adaptive Evolutionary Systems. *Quantitative Finance*, **1**, 149-167. https://doi.org/10.1080/713665542

[3] Sornette, D. (2017) Why Stock Markets Crash: Critical Events in Complex Financial Systems. Volume 49, Princeton University Press, Princeton, NJ. https://doi.org/10.23943/princeton/9780691175959.001.0001

[4] Marschinski, R. and Matassini, L. (2001) Financial Markets as a Complex System: A Short Time Scale Perspective (No. 01-4). Research Notes.

[5] Peinke, J., Böttcher, F. and Barth, S. (2004) Anomalous Statistics in Turbulence, Financial Markets and Other Complex Systems. *Annalen der Physik*, **13**, 450-460. https://doi.org/10.1002/andp.200410088

[6] Thaler, R.H. (1987) The January Effect. *Journal of Economic Perspectives*, **1**, 197-201. https://doi.org/10.1257/jep.1.1.197

[7] Jegadeesh, N. and Titman, S. (1993) Returns to Buying Winners and Selling Losers: Implications for Stock Market Efficiency. *The Journal of Finance*, **48**, 65-91. https://doi.org/10.1111/j.1540-6261.1993.tb04702.x

[8] Chan, K., Hameed, A. and Tong, W. (2000) Profitability of Momentum Strategies in the International Equity Markets. *The Journal of Financial and Quantitative Analysis*, **35**, 153-172. https://doi.org/10.2307/2676188

[9] Grinblatt, M., Titman, S. and Wermers, R. (1995) Momentum Investment Strategies, Portfolio Performance, and Herding: A Study of Mutual Fund Behavior. *American Economic Review*, **85**, 1088-1105.

[10] Barberis, N. and Thaler, R. (2003) A Survey of Behavioral Finance. *Handbook of the Economics of Finance*, **1**, 1053-1128. https://doi.org/10.1016/S1574-0102(03)01027-6

[11] Shiller, R.J. (2003) From Efficient Markets Theory to Behavioral Finance. *Journal of Economic Perspectives*, **17**, 83-104. https://doi.org/10.1257/089533003321164967

[12] Ricciardi, V. and Simon, H.K. (2000) What Is Behavioral Finance? *Business, Education & Technology Journal*, **2**, 1-9.

[13] Olsen, R.A. (1998) Behavioral Finance and Its Implications for Stock-Price Volatility. *Financial Analysts Journal*, **54**, 10-18. https://doi.org/10.2469/faj.v54.n2.2161

[14] Statman, M. (1995) Behavioral Finance versus Standard Finance. *Behavioral Finance and Decision Theory in Investment Management*, **1995**, 14-22. https://doi.org/10.2469/cp.v1995.n7.4

[15] Chan, W.S., Frankel, R. and Kothari, S.P. (2004) Testing Behavioral Finance Theories Using Trends and Consistency in Financial Performance. *Journal of Accounting and Economics*, **38**, 3-50. https://doi.org/10.1016/j.jacceco.2004.07.003

[16] Edmans, A., Garcia, D. and Norli, Ø. (2007) Sports Sentiment and Stock Returns. *The Journal of Finance*, **62**, 1967-1998. https://doi.org/10.1111/j.1540-6261.2007.01262.x

[17] Chang, S.-C., Chen, S.-S., Chou, R.K. and Lin, Y.H. (2012) Local Sports Sentiment and Returns of Locally Headquartered Stocks: A Firm Level Analysis. *Journal of Empirical Finance*, **19**, 309-318. https://doi.org/10.1016/j.jempfin.2011.12.005

[18] Levy, T. and Yagil, J. (2011) Air Pollution and Stock Returns in the US. *Journal of Economic Psychology*, **32**, 374-383. https://doi.org/10.1016/j.joep.2011.01.004

[19] Raj, M. and Kumari, D. (2006) Day-of-the-Week and Other Market Anomalies in

the Indian Stock Market. *International Journal of Emerging Markets*, **1**, 235-246. https://doi.org/10.1108/17468800610674462

[20] Kamstra, M.J., Kramer, L.A. and Levi, M.D. (2000) Losing Sleep at the Market: The Daylight Saving Anomaly. *American Economic Review*, **90**, 1005-1011. https://doi.org/10.1257/aer.90.4.1005

[21] Johnson, E.J. and Tversky, A. (1983) Affect, Generalization, and the Perception of Risk. *Journal of Personality and Social Psychology*, **45**, 20-31. https://doi.org/10.1037/0022-3514.45.1.20

[22] Forgas, J.P. (1995) Mood and Judgment: The Affect Infusion Model (AIM). *Psychological Bulletin*, **117**, 39-66. https://doi.org/10.1037/0033-2909.117.1.39

[23] Schwarz, N. (1990) Feelings as Information: Informational and Motivational Functions of Affective States. In: Sorrentino, R.M. and Higgins, E.T., Eds., *Handbook of Motivation and Cognition*: *Foundations of Social Behavior*, Guilford, New York, 527-561.

[24] Yuen, K.S. and Lee, T. (2003) Could Mood State Affect Risk-Taking Decisions? *Journal of Affective Disorders*, **75**, 11-18. https://doi.org/10.1016/S0165-0327(02)00022-8

[25] Chou, K.L., Lee, T. and Ho, A.H. (2007) Does Mood State Change Risk Taking Tendency in Older Adults? *Psychology and Aging*, **22**, 310. https://doi.org/10.1037/0882-7974.22.2.310

[26] Rao, T. and Srivastava, S. (2012) Using Twitter Sentiments and Search Volumes Index to Predict Oil, Gold, Forex and Markets Indices. Working Paper.

[27] Karabulut, Y. (2011) Can Facebook Predict Stock Market Activity? Working Paper, University of Frankfurt, Germany. https://doi.org/10.2139/ssrn.2017099

[28] Kramer, A.D., Guillory, J.E. and Hancock, J.T. (2014) Experimental Evidence of Massive-Scale Emotional Contagion through Social Networks. *Proceedings of the National Academy of Science of the United States of America*, **111**, 8788-8790. https://doi.org/10.1073/pnas.1320040111

[29] Houlihan, P. and Creamer, G.G. (2015) Leveraging Social Media to Predict Continuation and Reversal in Asset Prices. https://ssrn.com/abstract=2527968

[30] Pagolu, V.S., Reddy, K.N., Panda, G. and Majhi, B. (2016) Sentiment Analysis of Twitter Data for Predicting Stock Market Movements. 2016 *International Conference on Signal Processing*, *Communication*, *Power and Embedded System*, Paralakhemundi, India, 3-5 October 2016, 1345-1350. https://doi.org/10.1109/SCOPES.2016.7955659

[31] Bollen, J., Mao, H. and Zeng, X. (2011) Twitter Mood Predicts the Stock Market. *Journal of Computational Science*, **2**, 1-8. https://doi.org/10.1016/j.jocs.2010.12.007

[32] Leung, M.T., Daouk, H. and Chen, A.-S. (2019) Forecasting Stock Indices: A Comparison of Classification and Level Estimation Models. http://ssrn.com/abstract=200429

[33] Chen, A.-S., Daouk, H. and Leung, M.T. (2001) Application of Neural Networks to an Emerging Financial Market: Forecasting and Trading the Taiwan Stock Index. http://ssrn.com/abstract=237038 https://doi.org/10.2139/ssrn.237038

[34] Schumaker, R.P. and Chen, H. (2009) Textual Analysis of Stock Market Prediction Using Breaking Financial News: The AZFin Text System. *ACM Transactions on Information Systems* (*TOIS*), **27**, 12. https://doi.org/10.1145/1462198.1462204

[35] Hassan, M.R., Nath, B. and Kirley, M. (2007) A Fusion model of Hmm, Ann and Ga

for Stock Market Forecasting. *Expert Systems with Applications*, **33**, 171-180.
https://doi.org/10.1016/j.eswa.2006.04.007

[36] Kumar, M. and Thenmozhi, M. (2006) Forecasting Stock Index Movement: A Comparison of Support Vector Machines and Random Forest, Indian Institute of Capital Markets 9th Capital Markets Conference Paper.
http://ssrn.com/abstract=876544
https://doi.org/10.2139/ssrn.876544