Modelling movement of stock market indexes with data from emoticons of Twitter users

Alexander Porshnev, Ilya Redkin, Nikolay Karpov

National Research University Higher School of Economics,
Nizhny Novgorod, Russia

aporshnev@hse.ru, ilya-redkin@yandex.ru, nkarpov@hse.ru

Abstract. The issue of using Twitter data to increase the prediction rate of stock price movements draws attention of many researchers. In this paper we examine the possibility of analyzing Twitter users' emoticons to improve accuracy of predictions for DJIA and S&P500 stock market indices. We analyzed 1.6 billion tweets downloaded from February 13, 2013 to May 19, 2014. As a forecasting technique, we tested the Support Vector Machine (SVM), Neural Networks and Random Forest, which are commonly used for prediction tasks in finance analytics. The results of applying machine learning techniques to stock market price prediction are discussed.

Keywords: prediction; emoticons; DJIA; S&P500; Twitter; mood; Support Vectors Machine; Neural Networks; Random Forest; behavioral finance

1. Introduction

Moods and emotions influence our decisions; in psychological experiments Johnson and Tversky report that psychological states invoked by reading stories can affect the evaluation of risk level [1]. While in a good mood, an individual tends to make decisions expecting positive consequences and, vice versa, bad moods lead to pessimistic choices [2–4]. Moods and emotions determine the choice of basic heuristics, which can be done unconsciously. For example, individuals in a good mood are more likely to eliminate alternatives that fail to meet a criterion for an important dimension, which leads to increased efficiency[5].

It should be mentioned that other people's states play a big role not only in shaping individual moods, but also influence decision making. Nofsinger suggests the idea that the general level of optimism/pessimism shared in society is connected with economic activity [6]. Whereas Nofsinger supposes that the stock market itself can be a direct measure of social mood [6], in our research we decided to focus on an additional measure of shared emotions in Twitter.

Following Nofsinger, we will regard the economy as a complex system of human interactions, in which moods and irrationalities can play a significant role. This point can be supported by observing the informational cascades phenomenon in the stock market [7–9].

Twitter sentiment analysis gained in popularity in the financial domain thanks to the works by Bollen and his colleagues [10]. However, the possibility of predicting the stock market by means of analysis based on the wisdom of the crowds still triggers questions.

There are three main approaches to the use of Twitter data for financial forecasting. The first one is based on news analysis. For example, Reuters data shows that even fake news from a reliable source (Twitter account of Associated Press) can change the market, which means that information published in Twitter was used in a real trading strategy [11, 12]. The second approach is to analyze positive or negative sentiments about a company or a company's stock prices[13]. The third approach focuses on measuring the public mood and following the logic of behavioral finance used to improve stock market price forecasts. We know that the first and second approaches can be used in a trading strategy, but as far as the third approach is concerned, the situation is still unclear. There are several works on this topic and results vary from 87.4% of accuracy in works by Bollen and his colleagues [10] to 51.8% in those by Ding, Fang, and Zuo[14]. In our research using the same methodology we found S&P500 data from Twitter to be capable of significantly improving forecast accuracy to 68.63%[15]. Thus we chose to follow the third approach by testing the hypothesis within a bigger time frame and tried to change the methodology and concentrate on emoticons rather than on words.

In our research we regarded the amounts of Twitter emoticons as possible measures of social mood, and tested the hypothesis that it would be possible to use the analyses of moods expressed in tweets to increase prediction accuracy for stock market indicators.

The article has the following structure. The introduction is followed by Section 2 that describes the main design decisions and overviews the prediction system methodology. Section 3 contains a description of the dataset used in our research. Section 4 provides analysis of DJIA prediction and S&P500 indexes using additional information from Tweets. Section 5 compares the findings of our method and the approaches applied in the previous research. Section 6 concludes the work defining open research issues for further investigation.

2. Methodology

While analyzing online social networks using emoticons, Boia, Faltings and others found emoticons to closely coincide with the sentiment of the entire message. Tweets and their evaluation show that emoticons have a very good classification power[16] and that accuracy of emoticon-based sentiment classification is higher than 90% (for tweets with emoticons) [16]. Impressed by the results obtained by Boia and coauthors, we decided to analyze amounts of different emoticons as a measurement of public mood.

In our research, we calculated the amount of different emoticons by days, the most frequent ones being ":(" and ":)" (see Table 1 in Appendix 1 for a complete list of emoticons analyzed in our study). Rare emoticons, for example, ":-c", with an average occurrence of less than 10 per day, were excluded from the analysis.

We created two datasets – Basic and Emoticons. The Basic Dataset contains stock prices data for three previous days. The Emoticons Dataset contains a normalized frequency of emoticons in tweets on each day in addition to the Basic Dataset.

The standard supervised machine learning techniques were used to classify days by shifts in stock market indices: Support Vector Machine, Neural Networks and Random Forest. These techniques were chosen as the most common ones with the best performance in the field. We trained a model with one part of data and tested the created model for prediction with another part of data.

We used RapidMiner (http://rapidminer.com/) for data handling, which is one of the key data mining tools according to www.KDnuggets.com poll in 2013.

In RapidMiner, the Support Vector Machine algorithm uses the Java implementation of the support vector machine mySVM by Stefan Rueping [17]. The SVM implementation in RapidMiner supports the following types of kernels: dot, radial, polynomial, neural, anova, epachnenikov, gaussian combination and multiquadric. We tested dot, radial, polynomial and neural kernels to establish the baseline.

We used the Neural Net operator to create the Neural Networks Model in RapidMiner. The Neural Net operator implemented a feed-forward neural network trained by a back propagation algorithm.

The Random Forests technique is implemented in the Random Forest operator of RapidMiner, but our preliminary tests showed that W-Random Forest from Weka (from the Weka extension of RapidMiner) provided better performance. Thus, we used this implementation of the Random Forest technique.

To establish the baseline for prediction rates, we trained models on historical index price data (information of three previous days).First, we used all data and made 1,000 cycles of validation using 90 randomly chosen days for training and one random day for testing (1,000 predictions in total). This validation allowed us to test the hypothesis about the predictability level of a stock market and to choose the technique that had demonstrated the best performance. Next, we ran the optimization parameters operator to establish the baseline performance.

Second, we used the best modeling technique, trained it on our sequence of 89 trading days and tested it on the next (90^{th}) day following the chosen period. Our dataset allowed us to carry out 189 experiments, which means that we validated our models on 189 days. The same type of parameters optimization was employed to establish comparativeness with the baseline performance. We used this validation as we intended this study to be one of the steps to devising a trading strategy. We wanted to model the actual situation in the stock market, where we made a prediction for tomorrow's stock price movement, based on 89 observed days.

According to our hypothesis that emoticons can provide additional information, we expect the techniques trained on the Emoticons Dataset to exhibit better accuracy.

3. Data description

By making use of Twitter API we managed to download more than 1.6 billion unfiltered tweets over the period from 13/02/2013 to 19/05/2014 (we downloaded an average of 3,483,642 tweets per day) and that is approximately 1% of the total amount, according to API limitations. All the English tweets (where the user's "lang" parameter value equals "en") were sorted by day and analyzed automatically according to the counts of the emoticons (the complete set of emoticons and their frequency are presented in Table 1, Appendix 1).

We chose two stock market indicators whose prediction accuracy could be improved. The first one is the Dow Jones Industrial Average (DJIA), one of the oldest US stock market indices. The second one is Standard & Poor's 500 (S&P500), a stock market index based on the market capitalizations of 500 large companies having common stock listed on the New York Stock Exchange[1] or in the National Association of Securities Dealers Automated Quotations System[2].

For the DJIA and S&P500 stock market prices data we used the Yahoo Finance website[3], which provides opening and closing historical prices as well as the volume for any given trading day.

To apply the machine learning techniques, we divided the days into two equal groups by the difference between closing and opening prices. If on a day the opening price minus the closing price exceeded 50% of all the differences for the period from 13/02/2013 to 19/05/2014, then "shift" was equal to 1, and when it was lower than 50% it was 0.The Basic Dataset consisted of 16 columns (variables: shift (information about index shift 1 or 0), $Open_{-1\ day}$, $Close_{-1\ day}$, $Min_{-1\ day}$, $Max_{-1\ day}$, $Volume_{-1\ day}$, $Open_{-2\ day}$, $Close_{-2\ day}$, $Min_{-2\ day}$, $Max_{-2\ day}$, $Volume_{-2\ day;}$ $Open_{-3\ day}$, $Close_{-3\ day}$, $Min_{-3\ day}$, $Max_{-3\ day}$, $Volume_{-3\ day}$) and was employed to establish the baseline accuracy.

The Emoticons Dataset was created by adding columns about normalized frequencies of emoticons for the previous day (one day – 12 columns). To calculate normalized frequencies for each day, we divided the number of tweets with selected emoticons by the total number of tweets downloaded on this day.

The whole period from 13/02/2013 to 19/05/2014 contained 277 working days with available stock market information. This period was used for the first validation.

For the second validation the whole dataset was divided into sets with data of 90 days. The period of 90 days was chosen to enable the use of 89 days for training and 1 day for prediction. Our period of time allowed us to perform at least 189 prediction experiments for each time lag (from one to seven days), depending on the availability of stock market data.

The most frequent emoticons in 1.6 billion tweets were ":)" and ":(" – the same as in the study of the Twitter emoticon dictionary [18].

---

[1] http://www.nyse.com
[2] http://www.nasdaq.com
[3] http://www.finance.yachoo.com

4. Analysis

4.1 DJIA stock market prediction.

First, to find the baseline accuracy we trained Neural Networks, Support Vector Machine and Random Forest on the Basic DJIA data with one-day time lags (Table 1). The best accuracy was demonstrated by the Support Vector Machine technique with neural kernel (52.20%). We ran parameter optimization for this technique, which helped us increase prediction accuracy to 53.20% (kappa= 0.018, RMSE=0.500).

That level of performance became the baseline for our analysis. As the Support Vector Machine technique provided better performance, we used it in further analyses.

Table 1.“Shift” value prediction for DJIA. Accuracy of the Support Vector Machines, Random Forest and Neural Networks trained on the Basic Dataset

Model	Accuracy	Kappa	RMSE	Calculation time
SVM (dot)	48.80%	-0.038	0.54	6 sec
SVM (radial)	48.60%	-0.038	0.522	6 sec
SVM (polynomial)	48.80%	-0.038	0.543	6 sec
SVM (neural)	52.20%	0.043	0.61	7 sec
W-Random Forest	51.30%	0.017	0.54	32 sec
Neural Net	47.80%	-0.049	0.553	6 min

It is worth mentioning that Random Forest demonstrated compatible performance and the calculation time was reasonable (in comparison with SVM). In our next study we plan to focus more on applying Random Forest, as it allows the multiclass classification.

Next we extended prediction datasets with Twitter information and the train selected machine learning model – Support Vector Machine. Prediction accuracy for the SVM model trained on the Emoticons Dataset with different time lags is presented in Table 2.

Table 2."Shift" value prediction for DJIA. Accuracy of the Support Vector Machine trained on the Emoticons Dataset with different time lags

Lag in days	Accuracy	Kappa	RMSE
1	52.38%	0.049	0.642
2	**57.59%**	**0.154**	**0.511**
3	47.62%	-0.052	0.506
4	52.33%	0.046	0.524
5	41.75%	-0.165	0.532
6	48.68%	-0.024	0.533
7	**41.80%**	**-0.164**	**0.529**

Although the model showed better performance, the additional Tweeter information failed to significantly increase accuracy (Chi-square=1.085, p=0.297).

4.2. S&P500 index prediction

To establish the baseline accuracy, we ran the Support Vector Machine (neural kernel) on historical data with parameter optimization. The results showed that the Support Vector Machine provided a baseline accuracy of 50.70%.

Addition of Twitter information significantly improved our prediction accuracy(Chi-square=5.189, p<0.05). The best performance was achieved using the Emoticons Dataset with a two-day lag.

Table 3."Shift" value prediction for S&P500. Accuracy of the Support Vector Machine trained on the Emoticons Dataset with different time lags

Lag in days	Accuracy	Kappa	RMSE
1	52.91%	0.044	0.517
2	**59.69%**	0.192	0.504
3	47.62%	-0.054	0.527
4	49.22%	-0.023	0.529
5	50.00%	-0.002	0.518
6	59.26%	0.186	0.507
7	52.38%	0.051	0.051

5. Discussion

In our previous research we found that Twitter sentiment analysis data could significantly improve forecasting for the S&P500 index, and the new results with emoticons support our findings. Addition of Twitter information allowed us to increase accuracy from 50.70% (baseline) to 59.69% (SVM trained Emoticons data). It should be mentioned that a more complex approach we took in our previous research allowed forecast improvement of up to 68.63%.

Compared to other studies which used Twitter data analysis, we obtained lower accuracy. Bollen and his colleagues obtained an 86.7% accuracy for determining stock market movement [10]. Aanalyzing prediction of stock prices movements for the Apple company,Vu, Shu, Thuy and Nigel demonstrated 82.93%[19]. It should be mentioned, however, that these results were obtained on a relatively small number of testing days (21 days in the study by Bollen et al., and 41 days in that of Vu et al.) In our study the test sample was 189 days.

Comparison with other results in the stock market prediction field showed that what we demonstrated was almost the average performance. For example, usage analysis of Financial news gained 57% of directional accuracy [20]. Mahajan et al. taking the same approach obtained an accuracy of 60% [21]. While analyzing ad hoc announcements, Groth and Muntermann reached an accuracy of 75% [22].

Also worthy of mention is that we did not simulate any trading strategy based on our results, but expected that it would ultimately deliver more than 2.06% (demonstrated by simulation in a study by Schumaker and Chen who obtained an accuracy of 57.1% [20]).

Therefore, the emoticons approach may be used alone if only we need small improvement, but it is not suitable for more complex calculations.

SVM techniques exhibit not only the best performance for our classification task, but also enable the best calculation speed.

It should be also mentioned that analysis of correlations between normalized frequencies of different emoticons showed that they were highly related. For example, the correlation between the most frequent emoticons ":)" and ":(" is 0.965 (Fig.1). Such a high correlation between the appearance of sad and happy emoticons remains yet to be accounted for. We can only suppose that it may be connected with emotionality, and a rise in public emotionality will lead to increased emotions, whether good or bad.

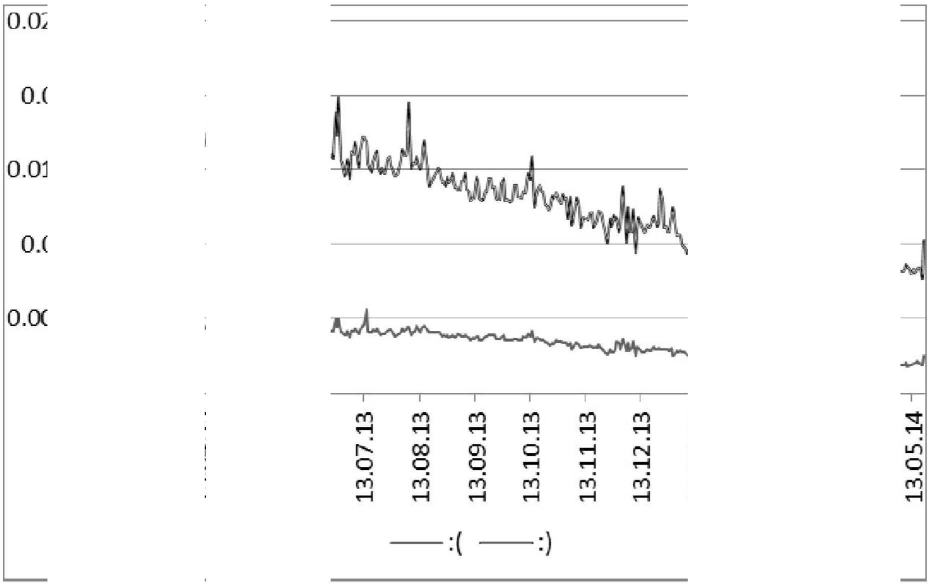

Fig. 1. Changes in normalized frequencies of ":)" and ":(" emoticons.

Interestingly, the best performance was demonstrated by SVM trained on the Emoticons Dataset with a two-day lag for both DJIA and S&P500 indices. In the research undertaken by Bollen, Mao, and Zeng the best results were achieved on a database with a four-day lag [10]. In our previous research, the lag for the best accuracy rate was 5 days. Such instability of results can be explained by changes occurring on a stock market and model mishitting could be a sign of information leakage, emerging policies or other events.

6. Conclusions

In our research we planned to test the hypothesis that even a simple sentiment analysis of Twitter data, such as that of emoticons frequency, can provide additional information capable of enhancing prediction accuracy for DJIA and S&P500. Analysis of 1.6 billion tweets downloaded over a period from 13/02/2013 to 19/05/2014 allowed us to significantly improve forecasts using the SVM technique. The obtained results suggest that our hypothesis can be confirmed. However, we found no significant improvement in accuracy, so our further research will combine both lexical and emoticons techniques in order to receive more information from Twitter. High correlations between sad and happy emoticons also need special attention, as does the instability of time lag in which public emotions are expressed in Twitter influence on the stock market prices. Our future research will also deal with shaping and testing a trading strategy based on Twitter data analysis.

References

1. Johnson, E.J., Tversky, A.: Affect, generalization, and the perception of risk. J. Pers. Soc. Psychol. 45, 20 (1983).
2. Isen, A.M., Patrick, R.: The effect of positive feelings on risk taking: When the chips are down. Organ. Behav. Hum. Perform. 31, 194–202 (1983).
3. Mayer, J.D., Gaschke, Y.N., Braverman, D.L., Evans, T.W.: Mood-congruent judgment is a general effect. J. Pers. Soc. Psychol. 63, 119 (1992).
4. Schwarz, N., Clore, G.L.: Mood, misattribution, and judgments of well-being: Informative and directive functions of affective states. J. Pers. Soc. Psychol. 45, 513 (1983).
5. Isen, A.M., Means, B.: The influence of positive affect on decision-making strategy. Soc. Cogn. 2, 18–31 (1983).
6. Nofsinger, J.R.: Social Mood and Financial Economics. J. Behav. Finance. 6, 144–160 (2005).
7. Bikhchandani, S., Hirshleifer, D., Welch, I.: A theory of fads, fashion, custom, and cultural change as informational cascades. J. Polit. Econ. 992–1026 (1992).
8. Salganik, M.J., Dodds, P.S., Watts, D.J.: Experimental Study of Inequality and Unpredictability in an Artificial Cultural Market. Science. 311, 854–856 (2006).
9. Bikhchandani, S., Hirshleifer, D., Welch, I.: Learning from the behavior of others: Conformity, fads, and informational cascades. J. Econ. Perspect. 12, 151–170 (1998).
10. Bollen, J., Mao, H., Zeng, X.: Twitter mood predicts the stock market. J. Comput. Sci. 2, 1–8 (2011).
11. Selyukh, A.: Hackers send fake market-moving AP tweet on White House explosions | Reuters, http://www.reuters.com/article/2013/04/23/net-us-usa-whitehouse-ap-idUSBRE93M12Y20130423.
12. Market reaction to Tuesday's erroneous tweet, http://pdf.reuters.com/pdfnews/pdfnews.asp?i=43059c3bf0e37541&u=2013_04_23_07_12_0ae1bd28b07544d5a23c965af0b0ac10_PRIMARY.jpg.
13. Sprenger, T.O., Tumasjan, A., Sandner, P.G., Welpe, I.M.: Tweets and Trades: the Information Content of Stock Microblogs. Eur. Financ. Manag. n/a–n/a (2013).
14. Ding, T., Fang, V., Zuo, D.: Stock Market Prediction based on Time Series Data and Market Sentiment. (2013).
15. Porshnev, A., Redkin, I., Shevchenko, A.: Improving Prediction of Stock Market Indices by Analyzing the Psychological States of Twitter Users. Social Science Research Network, Rochester, NY (2013).
16. Boia, M., Faltings, B., Musat, C.-C., Pu, P.: A :) Is Worth a Thousand Words: How People Attach Sentiment to Emoticons and Words in Tweets. (2013).
17. Rüping, S.: SVM kernels for time series analysis. Technical Report, SFB 475: Komplexitätsreduktion in Multivariaten Datenstrukturen, Universität Dortmund (2001).
18. Schnoebelen, T.: Do you smile with your nose? Stylistic variation in Twitter emoticons. Univ. Pa. Work. Pap. Linguist. 18, 14 (2012).
19. Vu, T.-T., Chang, S., Ha, Q.T., Collier, N.: An experiment in integrating sentiment features for tech stock prediction in Twitter. (2012).
20. Schumaker, R.P., Chen, H.: Textual analysis of stock market prediction using breaking financial news: The AZFin text system. ACM Trans. Inf. Syst. 27, 1–19 (2009).
21. Mahajan, A., Dey, L., Haque, S.M.: Mining financial news for major events and their impacts on the market. Web Intelligence and Intelligent Agent Technology, 2008. WI-IAT'08. IEEE/WIC/ACM International Conference on. pp. 423–426. IEEE (2008).

22. Groth, S.S., Muntermann, J.: An intraday market risk management approach based on textual analysis. Decis. Support Syst. 50, 680–691 (2011).

Appendix 1

Table 1. List of emoticons analyzed in the study

Emoticon	M	SD	Included in analysis	M (normalized)	Total amount
:)	10,176.87	36,170.15	+	0.00977	15,625,504
:(	3,643.26	12,539.19	+	0.00338	5,416,930
:))	1,415.28	4,748.58	+	0.00128	2,051,390
:-)	1,066.40	3,380.31	+	0.00091	1,460,296
:'(	720.09	2,197.49	+	0.00059	949,318
:)))	449.71	1560.92	+	0.00042	674,320
:((	382.91	1,417.75	+	0.00038	612,471
:(((	187.83	701.45	+	0.00018	303,027
:-(	142.69	580.67	+	0.00015	250,852
o_o	135.94	360.83	+	9.8383E-05	155,881
:-))	26.24	120.45	+	3.2152E-05	52,036
:-|	18.19	37.01	+	1.0073E-05	15,992
:-o	9.27	24.73		6.6889E-06	10,686
:-&	5.44	9.03		2.4352E-06	3,904
:'-(	3.65	8.86		2.3701E-06	3,828
x-(	3.28	5.28		1.4477E-06	2,285
:-@	2.74	4.63		1.2474E-06	2,002
:-!	2.65	4.23		1.1599E-06	1,829
:o	2.58	2.53		6.7428E-07	1,094
:(::	1.11	1.06		2.802E-07	461
:-t	0.50	0.21		5.7265E-08	94
:-1	0.49	0.16		4.2563E-08	72
:-c	0.30	0.07		1.9534E-08	34
:-o	0.27	0.06		1.6285E-08	28
:-(:::	0.04	0.002		6.4163E-10	1
>:o	0	0		0	0