NATIONAL RESEARCH UNIVERSITY
HIGHER SCHOOL OF ECONOMICS

Alexander Porshnev, Ilya Redkin, Alexey Shevchenko

IMPROVING PREDICTION OF STOCK MARKET INDICES BY ANALYZING THE PSYCHOLOGICAL STATES OF TWITTER USERS

BASIC RESEARCH PROGRAM

WORKING PAPERS

SERIES: FINANCIAL ECONOMICS
WP BRP 22/FE/2013

This Working Paper is an output of a research project implemented
at the National Research University Higher School of Economics (HSE). Any opinions or claims contained
in this Working Paper do not necessarily reflect the views of HSE.

Alexander Porshnev[1], *Ilya Redkin*[2], *Alexey Shevchenko*[2]

IMPROVING PREDICTION OF STOCK MARKET INDICES BY ANALYZING THE PSYCHOLOGICAL STATES OF TWITTER USERS

In our paper, we analyze the possibility of improving the prediction of stock market indicators by conducting a sentiment analysis of Twitter posts. We use a dictionary-based approach for sentiment analysis, which allows us to distinguish eight basic emotions in the tweets of users. We compare the results of applying the Support Vector Machine algorithm trained on three sets of data: historical data, historical and "Worry", "Fear", "Hope" words count data, historical data and data on the present eight categories of emotions. Our results suggest that the Twitter sentiment analysis data provides additional information and improves prediction as compared to a model based solely on information on previous shifts in stock indicators.

JEL Classification: G17, G02

Keywords: stock market; forecast; Twitter; mood; psychological states; Support Vectors Machine; machine learning,

[1] National Research University Higher School of Economics, Social Science department, N.Novgorod, Russia, aporshnev@hse.ru
[2] National Research University Higher School of Economics, Business Informatics faculty, N.Novgorod, Russia, ilya-redkin@yandex.ru, shevchenko.alexen@gmail.com

Introduction

Predicting financial markets is an interesting task from both practical and theoretical perspectives. By offering users a wide range of opportunities to express themselves, new information technologies make publicly available a huge amount of data about the emotions, moods, and psychological states of Internet citizens. In the U.S.A., which has a profound influence on the global economy, the Internet penetration rate is 78.3%, and active Internet users are also active financially. We think, therefore, that Twitter is a major additional resource of information that may help us improve forecasts of financial market. Although this idea was formulated several years ago, there is still no coherent opinion as to how this could be done (Bollen, Mao, and Zeng, 2011).

Over the past few years, significant progress has been achieved in using Twitter as an additional source of information (O'Connor, Balasubramanyan, Routledge, and Smith, 2010; Paul, and Dredze, 2011; Ruiz, Hristidis, Castillo, Gionis, and Jaimes, 2012). Bollen et al. (2011) reported that analyzing the text content of daily Twitter feeds increased the accuracy of DJIA predictions up to 87.6%. Zhang, Fuehres, and Gloor (2001) analyzed Twitter posts to predict stock market indicators such as the DJIA, S&P500, NASDAQ, and VIX, and found a high negative correlation (0.726, significant at level $p<0.01$) between the Dow Jones index and the presence of the words "hope", "fear", and "worry" in tweets (Zhang, Fuehres, and Gloor, 2011).

Chen and Lazer demonstrated that, using the approach proposed by Bollen, Mao, and Zeng, it is possible to create a more profitable trading strategy, but in their paper they did not provide information about the accuracy of prediction (Chen, Ray, Lazer, and Marius, 2013).

Regarding the application of sentiment analysis as a money generator, we did not find examples of successful projects. Derwent Capital Markets, a hedge fund, was the first to try applying sentiment analysis data, but their attempt proved inefficient (Malakian, 2013). Later the fund was rebranded into DCM Capital and presented a sentiment-based trading platform to retail investors (Malakian, 2013). Since a second attempt delivered no better results, DCM Capital's CEO, Paul Hawtin, put the sentiment-based platform up for auction. Starting at $7.9 million, the auction closed with a winning bid of $186,000 (Malakian, 2013). Yet, in his article, Malakian admits that there is no evidence to conclude that the failure of Derwent Capital Markets was due to poor technology (Malakian, 2013). The question of the applicability of sentiment analysis for real business has yet to be investigated.

There are two signs suggesting that this story is unfinished. The first is that Dow Jones and NYSE Technologies became partners aiming to improve prediction accuracy (Malakian, 2013). Second, according to Seth McGuire, director of Asset Management and Financial Technology, a few funds are now purchasing Twitter analyses, and social media aggregator Gnip will be one of the first to catch shifts in sentiment and capitalize on the market's wild swings (Or, 2011).

This leads us to the main hypothesis of our research, which is that analyzing tweets increases the accuracy of predicting stock market indicators. It is worth noting that analyzing Twitter-usage reliability is no easy task, as analysis algorithms are proprietary and their direct evaluation is impossible. To test our main hypothesis, we had to accomplish the following tasks:

1. Download a representative amount of raw data from Twitter.
2. Develop a sentiment analysis algorithm based on a psychological classification of emotions.
3. Analyze the prediction accuracy for machine-learning algorithms using data obtained from market and sentiment analyses.

Methodology

In our research, we faced two challenges: Twitter sentiment analysis and a prediction of the stock market based on sentiment analysis information.

Twitter sentiment analysis

A sentiment analysis of tweets could be conducted by training machine-learning algorithms on human-developed gold standards (Pang, Lee, and Vaithyanathan, 2002), or by calculating word frequencies from specially compiled dictionaries, which can include a group of n-grams words. In its simplest form, the lexicon approach was used by Zhang, Fuehres, and Gloor, who measured the quantity of tweets with the words "hope", "worry", and "fear" (Zhang et al., 2011). As we implemented the lexicon approach in our study, we found its application to be much faster than applying machine-learning algorithms to sentiment analysis. It allowed us to analyze a huge amount of tweets within reasonable time: it took us less than two days to analyze 288 million Tweets.

First, we used a Brief Mood Introspection Scale, with 8 scales and 2 adjectives representing each mood state as the starting point in creating dictionaries (Mayer and Gaschke, 1988). We also added all the synonyms of the selected adjectives from the WordNet dictionary (Miller, 1995). For example, we measured the presence of an energetic state in tweets by the occurrence of the following words: *animate, animated, athletic, brisk, chipper, emphatic, enterprising, exuberant, fresh, lusty, passionate, robust, sprightly, spry, strenuous, strong, tireless, trenchant, warming party, honor, vote.*

In total, our dictionaries of emotional words consist of 217 words, allowing us to recognize eight psychological states. To recognize derived words like "happyyy" or “happpppppyyyyyyy”, we use not just an exact word form, but regular expressions. For example, the expression [hap*y*] is used for the word “happy”. The entire content of tweets was entirely transferred to the lower case before analysis.

Second, we tested the quality of the sentiment analysis by applying the developed algorithms to the gold standard of 270 tweets (created by a professional translator with a specialist degree in the English language). For the quality test we used the standard measures of recall, precision, and F-measure (Jurafsky and Martin, 2008).

$$Recall_{\text{energetic}} = \frac{A}{A+C}, \tag{1}$$

where A is the amount of tweets correctly recognized as falling within the “energetic” class, and C is the amount of tweets unrecognized by our algorithm, but marked as inherent to this class in the gold standard.

$$Precision_{\text{energetic}} = \frac{A}{A+B}, \tag{2}$$

where A is the amount of tweets correctly recognized as falling within the “energetic” class, and B is the amount of tweets recognized by our algorithm, but marked as not inherent to this class in the gold standard.

$$F - measure_{cal} = \frac{2}{\frac{1}{Precision_{calm}} + \frac{1}{Recall_{calm}}} \tag{3}$$

Tab. 1. Measurement of performance for sentiment analysis

	happy	loving	calm	energetic	fearful	angry	tired	sad
Recall	93%	87%	57%	63%	70%	77%	73%	80%
Precision	90%	84%	71%	63%	70%	79%	79%	89%
F-measure	92%	85%	63%	63%	70%	78%	76%	84%

The results demonstrated a sufficient level of accuracy, and the F-measure for all categories was higher than 63% (Chen, Ray, Lazer, and Marius, 2013). The F-measure varied between 63% for two categories (“calm” and “energetic”) and 92% (“happy”). The achieved level of classification accuracy helped us obtain a fast and reliable algorithm for sentiment analysis.

Machine learning algorithms for stock market prediction

To test our main hypothesis, we used the Support Vector Machine algorithm, which helped us to classify days according to shifts in stock market indices and use the created model for prediction. The Support Vector Machine was chosen as it demonstrated the best performance in our preliminary research.

In order to understand whether a sentiment analysis of tweets provides any additional information, we used the SVM algorithm on three datasets. The first dataset characterized the stock market over the three previous days and was termed the basic set (Basic). The second set was created by adding to the basic set a normalized number of tweets with the words "Worry", "Hope", and "Fear" (Basic&WHF). The third set was formed by adding a normalized number of tweets from each of eight categories of the following emotions: "happy", "loving", "calm", "energetic", "fearful", "angry", "tired", and "sad" (Basic&8EMO). We expect that the comparison between prediction accuracies based on our three learning sets will be different. According to our hypothesis about additional information available in Twitter, we expect the first set to provide the lowest accuracy level, the second set to provide a somewhat higher accuracy, while the highest prediction accuracy will be achieved by using the Basic&8EMO dataset.

In their work, Bollen and his co-authors found better predictions based on data from the four previous days, and adding data from extra days led to an overtraining model (Bollen et al., 2011). To test these findings, we trained the Support Vector Machine on datasets, including different periods from the previous days (from one to seven days).

Data description

By making use of Twitter API, we managed to download more than 700 million tweets from the period of 13/02/2013 to 29/09/2013 (we downloaded an average of 3,483,642 tweets per day). All the tweets were sorted by day, analyzed automatically according to data counts of the words "Worry", "Hope", and "Fear" (WHF dataset), and assigned by the developed sentiment analyzer counting tweets in the following categories: "happy", "loving", "calm", "energetic", "fearful", "angry", "tired", and "sad" (8EMO dataset).

For the stock market data (S&P500, DJIA) we used the Yahoo finance website (http://finance.yahoo.com), which provides opening and closing historical prices as well as the volume for any given trading day. To apply the Support Vector Machine algorithm, we divided the days into two groups by adding a variable growth (0.1): 1 when the opening price was lower than the price at close, and 0 when the opening price was higher than or equal to the price at close. As a result, the Basic dataset consisted of 16 columns.

The first column provided information about index shift (1 or 0), then presented the opening price, closing price, maximum price, minimum price, and volume for three previous days. The Basic&WHF dataset, created by adding columns for the frequencies of the words “worry”, “hope”, and “fear” for the previous day (one day – three columns) or for several days (3 x number of days). For example, the Basic&WHF dataset for the previous 7 days consisted of 37 columns (16+3x7). While the Basic&8EMO dataset is formed in the same way, the three columns with word frequencies are replaced by 8 columns with frequencies of the words from the developed dictionary of emotional states. For example, the Basic&8EMO set with data from the sentiment analysis of the previous 7 days is composed of 72 columns (16+8x7).

The whole period from 13/02/2013 to 29/09/2013 was divided into sets with data from 95 days. The period of 95 days was chosen to enable the use of 80 days for training and 15 days for prediction. Within that period we ran a minimum of 5 experiments with the dataset containing information from the previous 7 days (75 predictions) and a maximum of 40 experiments, with information just of the previous day (600 predictions). The 75 days given for prediction is a much longer period than the 19 days that Bollen and his colleagues had. The increased number of experiments helped us enhance the validity of findings.

Analysis

Stock market growth prediction

Applying the Support Vector Machine algorithm trained only on the Basic DJIA data provided an accuracy of 65.17%, which became the baseline for our analysis. We also tried to train the SVM on data with information from more than one day (from two to seven days), but it resulted in a less accurate forecast (Table 2). Prediction accuracy for the algorithm trained on the Basic dataset with one day’s information was used as a baseline in further analysis.

Tab. 2. DJIA prediction. Accuracy of the Support Vector Machines algorithm trained on the Basic dataset.

Number of previous days included in dataset	**1 day**	**2 days**	**3 days**	**4 days**	**5 days**	**6 days**	**7 days**
Basic	**65.17%**	59.84%	60.00%	60.78%	57.93%	49.30%	48.68%
Number of days for prediction	600	645	630	645	675	645	645

The results presented in Table 3 demonstrate that using a more complex approach to extract emotional states does not provide more information than the basic method of relying on the appearance of the three words “worry”, “hope”, and “fear”. Although the usage of the WHF dataset provided a better forecast, this improvement was not significant ($\chi^2(1)= 1.099$, $p=0.294$).

Tab. 3. DJIA prediction. Accuracy of the Support Vector Machines algorithm versus the training dataset.

	Number of previous days included in dataset						
Dataset	**1 day**	**2 days**	**3 days**	**4 days**	**5 days**	**6 days**	**7 days**
Basic&WHF	63.00%	61.01%	61.73%	66.35%	69.02%	**70.00%**	73.33%
Basic&8EMO	62.50%	60.20%	59.26%	65.40%	68.63%	**70.00%**	73.33%
Number of experiments	40	33	27	21	17	12	5
Number of days for prediction	600	495	405	315	255	180	75

Training the SVM algorithm on the Basic S&P500 dataset provided a baseline accuracy of 57.00%. Similarly, it is evident from Table 4 that applying the basic algorithm to sentiment analysis provided better information to improve forecasting (insignificant differences). The SVM algorithm trained on technical data and the number of instances of “fear”, “hope”, and “worry” in the previous five days demonstrated a higher accuracy of DJIA prediction (68.10%), and it was 10% more accurate than the baseline approach ($\chi^2(1)= 5.027$, $p<0.05$).

Table 4. S&P500 prediction. Accuracy of the Support Vector Machine algorithm versus the training dataset.

	Number of previous days included in dataset						
Dataset	**1 day**	**2 days**	**3 days**	**4 days**	**5 days**	**6 days**	**7 days**
Basic&WHF	59.33%	54.75%	54.07%	63.49%	**68.63%**	63.33%	60.00%
Basic&8EMO	56.67%	53.94%	58.77%	62.86%	67.84%	66.11%	56.00%
Number of experiments	40	33	27	21	17	12	5
Number of days for prediction	600	495	405	315	255	180	75

The baseline accuracy for the NASDAQ index was 50.67% (training on the Basic dataset). An analysis of the S&P500 prediction accuracy (Table 5) showed that both algorithms performed better if trained on datasets including information about the emotional states of Twitter users, but the accuracy differences were insignificant ($\chi^2(1)= 1.171$, p=0.279).

Tab. 5. NASDAQ prediction. Accuracy of the Support Vector Machine algorithm versus the training dataset.

	Information from previous						
Dataset	**1 day**	**2 days**	**3 days**	**4 days**	**5 days**	**6 days**	**7 days**
Basic&WHF	52.17%	50.30%	46.91%	47.62%	49.80%	45.00%	33.33%
Basic&8EMO	52.00%	49.90%	52.35%	52.70%	**56.08%**	52.78%	44.00%
Number of experiments	40	33	27	21	17	12	5
Number of days for prediction	600	495	405	315	255	180	75

Discussion

The application of Twitter data for stock market prediction looks like an attempt to use a magic crystal ball or unrelated data. However, it may not be as far-fetched as it appears at first sight. Impressed by the work of Bollen and his colleagues, we wanted to replicate and expand their results. Since 2008, when Bollen and his colleagues conducted their research, Twitter has changed dramatically. In 2008, the number of tweets from February 28 to December 19 was significant, amounting to 9,853,498 (Bollen, Mao, and Zeng, 2011). We now have to download much more. In the period from13/02/2013 to 29/09/2013 we downloaded 755,000,101 tweets – 76 times more than Bollen and his colleagues did, but within a shorter period of time.

Bollen, Mao, and Zeng reported that they had downloaded data for a period from February 28 to December 19, 2008, with the bulk of the data used for training and testing the algorithm (February 28 to November 28, 2008) and only 19 days spent for prediction (December 1 to 19, 2008). In our study the minimum number of days for prediction was 75 and the maximum number was 675, which enabled us to formulate more statistically valid statements.

The addition of sentiment analysis data to the training dataset for the SVM algorithm resulted in a 70% accuracy for the stock market predictions of the DJIA (previous 6 days included), 68.63% for the S&P500 (previous 5 days included), and 56.08% for NASDAQ (previous 5 days included). While these results do not outperform the accuracy achieved by Bollen and his colleagues, the high prediction rate they demonstrated could have been achieved by chance, as they had a short testing period of 19 days. We asked the authors to send us their dataset, as it is no longer available on the web (http://terramood.informatics.indiana.edu/data), but have not yet received an answer.

However, we have a higher prediction accuracy for the S&P500 indicator, than that achieved by Ding et al (51.88%) (Ding, Fang, and Zuo, 2013) and the 68% reported by Mao et al (Mao, Wei, Wang, and Liu, 2012).

We found the Basic dataset to provide less information than the Basic&WHF. It supports the findings by Zhang, Fuehres, and Gloor, who maintain that calculation alone of the three words "Worry", "Hope", and "Fear" in tweets can provide additional information that increases prediction accuracy. We found it interesting that almost all sets of experiments with different amounts of information, included in the SVM dataset trained on Basic&WHF, on average performed better than SVM trained on Basic&8EMO. There are two possible explanations. First, Basic&8EMO provided more data that led to model overtraining. The second explanation is connected with the need to further improve sentiment analysis of tweets. The enhanced accuracy of sentiment analysis may deliver better results. Potential areas of improvement may be adding a weight to the words. For example, the word "*fear*" should weigh heavier in tweet analysis than, say, "*coward"*.

While collecting data, we downloaded about 1% of tweets published by users in historical order. That could not guarantee, however, that messages came from US citizens, even though 55% of them were in English. In order to improve prediction accuracy, we have to limit downloaded tweets only by analyzing user location, and we plan to continue in this manner in our future research.

Another point we found was a different baseline level for different indices. For example, we observed a maximum baseline accuracy of 65.17% for the DJIA, 57% for the S&P500, and only 50.67% for NASDAQ.

Analysis showed that even simple Twitter sentiment analysis data could significantly improve forecasting, which confirms our hypothesis (17 experiments within 15 days for predications, 255 days in total). It should be mentioned that we predicted the direction of change for stock market indices, but not the level of change, which may seriously limit the application of our findings to real-world trading strategy.

Conclusion

Our research sought to test the hypothesis that sentiment analysis of Twitter data may provide additional information, which may improve the accuracy of stock market prediction.

First, we created a server application to download and store tweets. Over the period from 13/02/2013 to 29/09/2013, we downloaded 755,000,101 tweets, with the daily average being 3,483,642. To analyze this huge amount of data, we needed a fast and reliable algorithm for sentiment analysis. To solve this task we used the lexicon-based approach that showed satisfactory performance.

Our results suggest that our hypothesis can be confirmed, at least for predicting the S&P500 index, for which we significantly improved forecasting accuracy. An accuracy of 57.00% provided by SVM, trained only on historical data, was increased to 68.63% through the use of Twitter sentiment analysis data.

In our further research, we plan to enhance accuracy by improving dictionaries and introducing word weights, to thoroughly check the reliability and validity of our dictionaries' algorithm, to continue data collection, and implement machine learning algorithms to predict the amount of change for chosen stock market indices.

References

Bollen, Johan, Huina Mao, and Xiaojun Zeng. 2011. "Twitter Mood Predicts the Stock Market." *Journal of Computational Science* 2 (1): 1–8. doi:10.1016/j.jocs.2010.12.007.

Chen, Ray, and Lazer, Marius. 2013. "Sentiment Analysis of Twitter Feeds for the Prediction of Stock Market Movement." *Stanford.edu*. http://cs229.stanford.edu/proj2011/ChenLazer-SentimentAnalysisOfTwitterFeedsForThePredictionOfStockMarketMovement.pdf.

Ding, Tina, Vanessa Fang, and Daniel Zuo. 2013. "Stock Market Prediction Based on Time Series Data and Market Sentiment." http://murphy.wot.eecs.northwestern.edu/~pzu918/EECS349/final_dZuo_tDing_vFang.pdf.

Jurafsky, Daniel, and James H. Martin. 2008. *Speech and Language Processing An Introduction to Natural Language Processing, Computational Linguistics, and Speech*. New Jersey: Pearson Prentice Hall.

Malakian, Anthony. 2013. "Was DCM Capital's Failure a Sign That the Industry Is Not Ready for Sentiment Analysis? Or Was It a Blip? Anthony Explores." *WatersTechnology*. February 22. http://www.waterstechnology.com/buy-side-technology/opinion/2250200/sentiment-analysis-still-has-a-long-way-to-go-on-wall-street.

Mao, Yuexin, Wei Wei, Bing Wang, and Benyuan Liu. 2012. “Correlating S&P 500 Stocks with Twitter Data.” In *Proceedings of the First ACM International Workshop on Hot Topics on Interdisciplinary Social Networks Research*, 69–72. http://dl.acm.org/citation.cfm?id=2392634.

O’Connor, Brendan, Ramnath Balasubramanyan, Bryan R. Routledge, and Noah A. Smith. 2010. “From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series.” In *Proceedings of the International AAAI Conference on Weblogs and Social Media*, 122–129. http://www.aaai.org/ocs/index.php/ICWSM/ICWSM10/paper/viewPDFInterstitial/1536/1842.

Or, Amy. 2011. “Now Trending: Turning Tweets Into Trades.” *MarketBeat*. December 12. http://blogs.wsj.com/marketbeat/2011/12/12/now-trending-turning-tweets-into-trades/.

Pang, Bo, Lillian Lee, and Shivakumar Vaithyanathan. 2002. “Thumbs up?: Sentiment Classification Using Machine Learning Techniques.” In *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing-Volume 10*, 79–86. http://dl.acm.org/citation.cfm?id=1118704.

Paul, Michael J., and Mark Dredze. 2011. “You Are What You Tweet: Analyzing Twitter for Public Health.” In *Fifth International AAAI Conference on Weblogs and Social Media (ICWSM 2011)*. http://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/viewPDFInterstitial/2880/3264.

Ruiz, Eduardo J., Vagelis Hristidis, Carlos Castillo, Aristides Gionis, and Alejandro Jaimes. 2012. “Correlating Financial Time Series with Micro-Blogging Activity.” In *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining*, 513–522. http://dl.acm.org/citation.cfm?id=2124358.

Zhang, Xue, Hauke Fuehres, and Peter A. Gloor. 2011. “Predicting Stock Market Indicators Through Twitter ‘I Hope It Is Not as Bad as I Fear.’” *The 2nd Collaborative Innovation Networks Conference - COINs2010* 26 (0): 55–62. doi:10.1016/j.sbspro.2011.10.562.

Contact details and disclaimer:

Alexander V. Porshnev
National Research University Higher School of Economics (N.Novgorod, Russia). Social Science Department, Associate Professor
E-mail: aporshnev@hse.ru, Tel. +7 (920) 2532970

Any opinions or claims contained in this Working Paper do not necessarily reflect the views of HSE.

© Porshnev, 2013