Machine learning in prediction of stock market indicators based on historical data and data from Twitter sentiment analysis.

Alexander Porshnev, Ilya Redkin, Alexey Shevchenko
National Research University Higher School of Economics
Nizhniy Novgorod, Russia
aporshnev@hse.ru, ilya-redkin@yandex.ru, shevchenko.alexen@gmail.com

Development of linguistic technologies and penetration of social media provide powerful possibilities to investigate users' moods and psychological states of people. In this paper we discussed possibility to improve accuracy of stock market indicators predictions by using data about psychological states of Twitter users. For analysis of psychological states we used lexicon-based approach, which allow us to evaluate presence of eight basic emotions in more than 755 million tweets. The application of Support Vectors Machine and Neural Networks algorithms to predict DJIA and S&P500 indicators are discussed.

Prediction; stock market indicator; Twitter; mood; psychological states; Support Vectors Machine; Neural Networks

I. INTRODUCTION

Machine learning algorithms are used in the stock market forecasting for a long time [1], [2]. The most common methods are Neural Networks and Support Vector Machine [1], [3]. Usually machine learning algorithms trained on technical data about stock movements, for example moving averages. Although, technical data is important for stock prediction contemporary traders need more advanced strategies to outperform market. According to behavioral economics it be could useful to add information about emotions, moods and psychological states of people [4].

In the last years significant progress was demonstrated in using Twitter as additional source of information [5], [6]. Bollen et al. (2011) reported that the analysis of the text content of daily Twitter feeds increased accuracy of DJIA predictions up to 87.6%. It worth to mention, that in spite of wide time range of available data the accuracy of prediction were measured only for 19 days. Bollen and his colleagues wrote: "February 28, 2008 to November 28, 2008 is chosen as the longest possible training period while Dec 1 to Dec 19, 2008 was chosen as the test" [1, p.5]

Zhang, Fuehres, and Gloor (2001) analyzed Twitter posts to predict stock market indicators such as DJIA, S&P500, NASDAQ, VIX and found a high negative correlation (0.726, significant at level $p<0.01$) between Dow Jones index and presence of words "hope", "fear", "worry" in tweets [7].

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 do not provide information about the accuracy of prediction [2].

Although, in April 23th hackers attack on Associated Press Twitter account showed that analysis of news is widely used in trading [8], we could not found such strong evidence for sentiment analysis techniques. The first attempt to apply sentiment analysis data was made by a hedge fund named Derwent Capital Markets, but their results did not show any efficiency [9]. Later the fund was rebranded into DCM Capital and presented to the retail investors sentiment-based trading platform [9]. However, a second attempt was not more successful and DCM Capital CEO Paul Hawtin put the sentiment-based platform up for sale in an auction. The asking price was $7.9 million, but the auction closed with winning bid $186,000 [9]. However, in his article Malakian admits that there is no evidence to conclude that failure of Derwent Capital Markets happened because of poor technology [9]. The question about the applicability of sentiment analysis in real business is yet to be investigated.

We observe two signs that this story is not over. First, Dow Jones and NYSE Technologies became partners in order to increase accuracy of prediction [9]. Second, Seth McGuire, Director of Asset Management and Financial Technology said that several funds buy analyses of Twitter and other social media from Gnip to be the first who can catch shifts in sentiment as the key to capitalizing on the market's wild swings [10].

This lead us to the main hypothesis of our research, that analyses of tweets increase the accuracy of predication for stock market indicators. In spite of the significant process on application of Twitter data it is not become easier, first because amount of tweets growth rapidly and second - used algorithms of analysis are proprietary and should be developed. This adds two more tasks to our project: to download representative amount of raw data from Twitter and to develop an algorithm for sentiment analysis, based on psychological classification of emotions.

II. METHODOLOGY

In our research we met with two major tasks: Twitter sentiment analysis and prediction of stock market based on sentiment analysis information.

A. Twitter sentiment analysis

Research in natural language processing provides several directions for sentiment analysis, first is classification based on human developed gold standard [11]. All categories of sentiments should be presented in gold standard, so it could be used to train Naïve Bayes or other machine learning algorithms for the analysis of other tweets [12]. Creation of a

978-0-7695-5109-8/13 $31.00 © 2013 IEEE
DOI 10.1109/ICDMW.2013.111

IEEE computer society

gold standard is usually associated with a lot of efforts and work of a team of linguistics (e.g. Lyashevskaya et al. [13]).

The second approach is based on dictionaries. This approach was used by Bollen and his colleagues, who have received the best results to this moment, and we decided to follow them in choosing a dictionary approach for sentiment analysis [4]. In its simplest form, this approach was used by Zhang, Fuehres and Gloor by measuring the quantity of tweets with the words "hope", "worry" and "fear" [7].

In our study we realize two versions of lexicon based approach. First, we simply calculate frequencies of words "hope", "worry" and "fear" in tweets. Second, we create more complex dictionaries for each of eight basic emotions and analyze the presence of these words. To analyze the efficiency of recognitions of emotions we ask experts in linguistics to create gold-standard for emotions in tweets. To check quality of emotions recognition we used standard measures recall, precision and F-measure [12].

B. Machine learning algorithms for stock market prediction

To test our main hypothesis we used two machine learning algorithms which allow us to classify days by appearance of events and use created model for prediction. They are Neural Networks and Support Vector Machine.

In order to answer the question: "Do sentiment analysis of tweets provide additional information?", we use learning algorithms on three sets of data. The first set of data were the characteristics of stock market in previous days, we call it basic set (Basic). The second set was created by adding a normalized number of tweets with words "Worry", "Hope", "Fear" to the basic set (Basic&WHF). The third set was created by adding a normalized number of tweets from each of 8 categories of the following emotions: "happy", "loving", "calm", "energetic", "fearful", "angry", "tired", "sad" (Basic&8EMO). We expect that the comparison between accuracy of predictions based on our three learning sets will be different. According to our hypothesis about the existence of additional information in Twitter, we expect that the first set will provide lowest accuracy level, second provides somewhat higher accuracy and the highest level of prediction accuracy will be received based on the usage data set Basic&8EMO.

In work of Bollen and his co-authors, they found better predictions based on data that occur during 3 to 4 earlier shift in the DJIA [4]. To test these findings, data from Twitter were used to train Neural Networks and Support Vector Machine algorithms with the time lags from one to seven days.

C. Data description

To download the tweets we used Twitter API which allows us to download approximately 145 000 tweets in one hour and in period from 13/02/2013 till 29/09/2013 we downloaded 755'000 101 messages (on average we downloaded 3483642 tweets per day). All tweets were sorted by day and analyzed automatically according to data counts of the words "Worry", "Hope", "Fear" (data set WHF) and assigned by a developed sentiment analyzer counting tweets to the following categories: "happy", "loving", "calm", "energetic", "fearful", "angry", "tired", "sad" (data set 8EMO).

For the stock market data 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.

The period from 13/02/2013 till 29/04/2013 was divided on 61 days periods. First 60 days were used to train machine learning algorithms, and then trained algorithm makes prediction for last 61th day. We can use only data from business days, and after division we received 80 periods (every period consists from 61 day).

For lagged analysis we have to shift data, that is why amount of experiments vary a little from 76 to 80 in dependence with time lag.

III. Analysis

A. Sentiment analysis

For sentiment analysis we decided to use the dictionary approach, firstly because it can provide reliable information, and secondly because it requires fewer resources to run and can be much faster than widely used Naïve Bayes algorithm. We used a Brief Mood Introspection Scale with 8 scales and 2 adjectives representing each mood state for starting point in creation of dictionaries [14]. We also added all synonyms of selected adjectives from the WordNet dictionary [15].

To test the quality of the sentiment analysis of our algorithm we manually created a gold standard from 240 tweets, 30 per sentiment category. Each from 240 tweet was analyzed by professional translator with specialist degree in English language, and distributed to one or several emotions categories (it also could happen that tweets have no emotional information, meaning that a tweet had a score of 0 on all 8 scales). The first version of our dictionaries provided a good result on the test data, but the analysis of mistakes does not allow us to improve our algorithms by adding new adjectives, nor to recognize derivative words like "happyyy" or "happpppppyyyyyyy". The second version of the questionnaire consists of 217 words and provides better results for all parameters of efficiency of sentiment analysis (see Table 1.)

TABLE I. Comparison of performance in sentiment analysis of first and second vestions of dictionaries.

	First version			**Second version**		
	Precision	***Recall***	***F-measure***	***Precision***	***Recall***	***F-measure***
Happy	87%	87%	87%	90%	93%	92%
Loving	77%	77%	77%	84%	87%	85%
Calm	63%	40%	49%	71%	57%	63%
energetic	57%	57%	57%	63%	63%	63%
Fearful	61%	57%	59%	70%	70%	70%
Angry	70%	63%	67%	79%	77%	78%
Tired	69%	67%	68%	79%	73%	76%
Sad	85%	73%	79%	89%	80%	84%

The comparison with efficacy of Naïve Bayes algorithm trained on subset from 180 tweets and tested on 90 tweets showed that our algorithm worked better (Table 2).

TABLE II. COMPARISON OF SENTIMENT ANALYSIS PERFORMANCE MEASURES FOR NAÏVE BAYES AND DICTIONARIES APPROACH

	Naïve Bayes	Dictionaries
Recall	52%	88%
Precision	68%	79%
F-measure	59%	83%

This allowed us to conclude that we solved a first task, to receive a reliable algorithm for sentiment analysis and could move on to its application for the prediction of a stock data.

B. Prediction of the growth of stock market

We started by generating data sets. First, we filtered tweets only from business days, and wrote a Java-script to generate the data sets Basic, Basic&WHF, Basic&8EMO. Each data set had 7 sub tables for lag in time from one to seven days. To apply Support Vector Machine, and Neural Networks algorithms we divided the days into two groups by adding a variable growth (0,1), (1 when the opening price was lower than price at close, 0 when the opening price was higher than or equal to the price at close).

We divided the analyzed period in to data sets contained 61 days. We used the first 60 days as a training sample and 1 day as test sample. Analyzed period allow us to conduct more than 70 prediction experiments.

Results presented in Table 3 demonstrate that using more complex approach to extract emotional states do not provide more information than basic method rely on appearance of three words "worry", "hope" and "fear". Although, Twitter analysis add some information we could not say that quality of forecast changes significantly. The higher accuracy demonstrated by Basic&Emo data set is equal to 61.10% (time lag 2), for Basic&WHF is equal to 61.84% (time lag =1), difference is not significant ($\chi^2(df=1)= 0.084$, $p= 0.771$).

TABLE III. PREDICTION OF DJIA. AVERAGE ACCURACY OF SUPPORT VECTOR MACHINE ALGORITHM IN DEPENDENCE FROM TRAINING DATA SET

Lag	Basic	Basic&WHF	Basic&8EMO	Number of experiments
1 day	60.53%	61.84%	60.53%	76
2 days	60.26%	61.54%	64.10%	78
3 days	51.32%	53.95%	56.58%	76
4 days	57.14%	57.14%	54.55%	77
5 days	58.75%	60.00%	61.25%	80
6 days	58.23%	55.70%	58.23%	79
7 days	60.76%	59.49%	58.23%	79

In Table 4, it is that and simple approach to sentiment analysis provide better information to improve forecast (59.49%). Although, there is no significant differences in accuracy of predictions (for Basic&EMO time lag 5 days (53.75%) and for Basic&WHF 7 days (59.49%)).

TABLE IV. PREDICTION OF DJIA. AVERAGE ACCURACY OF NEURAL NETWORKS ALGORITHM IN DEPENDENCE FROM TRAINING DATA SET

Lag	Basic	Basic&WHF	Basic&8EMO	Number of experiments
1 day	59.21%	39.47%	52.63%	76
2 days	57.69%	56.41%	47.44%	78
3 days	46.05%	47.37%	44.74%	76
4 days	51.95%	53.25%	53.25%	77
5 days	55.00%	55.00%	53.75%	80
6 days	45.57%	45.57%	48.10%	79
7 days	48.10%	59.49%	49.37%	79

Although, analysis of accuracy for prediction S&P500 indicator showed in Table V the paradoxical situation than addition of data from Twitter decrease accuracy of forecast, the difference in average accuracy of prediction is not significant.

TABLE V. PREDICTION OF S&P500. AVERAGE ACCURACY OF SUPPORT VECTOR MACHINE ALGORITHM IN DEPENDENCE FROM TRAINING DATA SET

Lag	Basic	Basic&WHF	Basic&8EMO	Number of experiments
1 day	52.63%	55.26%	57.89%	76
2 days	56.41%	56.41%	55.13%	78
3 days	52.63%	53.95%	60.53%	76
4 days	55.84%	57.14%	57.14%	77
5 days	56.25%	58.75%	58.75%	80
6 days	59.49%	58.23%	58.23%	79
7 days	62.03%	58.23%	55.70%	79

Comparing results obtained from Support Vector Machine and Neural Networks algorithms we could see that for our data sets first approach provide a little bit more accurate prediction for DJIA 64.10%, instead of 59.21% (difference is not significant), for S&P500 62.03% instead of 59.74% (difference is not significant).

TABLE VI. PREDICTION OF S&P500. AVERAGE ACCURACY OF NEURAL NETWORKS ALGORITHM IN DEPENDENCE FROM TRAINING DATA SET

Lag	Basic	Basic&WHF	Basic&8EMO	Number of experiments
1 day	55.26%	44.74%	42.11%	76
2 days	48.72%	50.00%	43.59%	78
3 days	46.05%	51.32%	51.32%	76
4 days	50.65%	59.74%	57.14%	77
5 days	52.50%	53.75%	53.75%	80
6 days	49.37%	54.43%	48.10%	79
7 days	54.43%	53.16%	45.57%	79

IV. 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. Based on work by Bollen and his colleagues we wanted to replicate and expand their results in a wide time frame. Application of sentiment analysis data for machine learning algorithms allows us to receive maximum accuracy of stock market predictions for DJIA – 64.10%. For DJIA our accuracy is below 87.6% that reported by Bollen and co-authors. This could lead to a conclusion that probably higher prediction rate demonstrated by Bollen and co-athours was related to a small test period (only 19 days).

These results could also be explained by other factors. First, it could be that information about application of Twitter for DJIA become available to trading society in 2010 and now this analysis technique could not consistently beat the market as some of traders already used it. Partially this could confirm efficient market hypothesis. Second, probably we need to extend training period from 60 days to several month like Bollen and his colleagues did. Third, we could not compare performance directly because proprietary nature their algorithm and further improvement of our sentiment analyzer needed.

However, we found out that Support Vector Machine provide a little better prediction accuracy of S&P500 indicator (62.03%), than 51.88% demonstrated by Ding et Al. [3].

We found that our Twitter analyzer could provide significantly higher accuracy of prediction and could not confirm our hypothesis, as we found no significant differences in average accuracy of predictions based on all three data sets.

Our research provides a new argument about potential possibility to improve predication of stock market indicators using human sentiments analysis. Although, we think that at this point it is too early to suggest that Twitter sentiment analysis could not to improve forecasts and more experiments are needed. Also it can be seen that further experiments will require more effort as it Twitter is growing rapidly: in 2008, 9,853,498 tweets could represent the period from February 28 to December 19th, 2008, and in 2013 for representing period from 13 February until 29 September 2013 we have to download 755'000 101 tweets. Taking into account different length eleven months in research of Bollen et al. and eight months in ours, we could estimate that to make a whole year analysis we have to download and analyze more than one billion of tweets.

V. Conclusion

In our research we tested hypothesis that sentiment analysis of Twitter data could provide additional information and this could increase accuracy of stock market prediction.

We created server application and in the period from 13/02/2013 till 29/09/2013 downloaded 755'000 101 tweets. Next step was creation of fast and reliable algorithm for sentiment analysis. To accomplish it we used a lexicon based approach and the second version of dictionaries showed satisfactory performance.

Our preliminary results indicate that addition of information from Twitter do not allow us to significantly increase accuracy. The best average accuracy rate 64.10% was achieved using Support Vector Machine algorithm to predict DJIA indicator.

In further research we plan to increase training period, and improve our sentiment analysis algorithms.

References

[1] Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. Journal of Computational Science, 2(1), 1–8. doi:10.1016/j.jocs.2010.12.007

[2] Chen, Ray and Lazer, Marius, "Sentiment Analysis of Twitter Feeds for the Prediction of Stock Market Movement," stanford.edu, 2013. [Online]. Available: http://cs229.stanford.edu/proj2011/ChenLazer-SentimentAnalysisOfTwitterFeedsForThePredictionOfStockMarketMovement.pdf. [Accessed: 25-Jan-2013].

[3] Ding, T., Fang, V., & Zuo, D. (2013). Stock Market Prediction based on Time Series Data and Market Sentiment. Retrieved from http://murphy.wot.eecs.northwestern.edu/~pzu918/EECS349/final_dZuo_tDing_vFang.pdf

[4] Jurafsky, D., & James, H. (2000). Speech and Language Processing An Introduction to Natural Language Processing, Computational Linguistics, and Speech. Retrieved from http://repository.unikom.ac.id/repo/sector/buku/view/3/key/12110/Speech-and-Language-Processing-An-Introduction-to-Natural-Language-Processing-Computational-Linguistics-and-Speech-Recognition.html

[5] Lawrence, R. (1997). Using neural networks to forecast stock market prices. University of Manitoba. Retrieved from http://sharesforcast.googlecode.com/svn-history/r77/nn.pdf

[6] Malakian, A. (2013, February 22). Was DCM Capital's failure a sign that the industry is not ready for sentiment analysis? Or was it a blip? Anthony explores. WatersTechnology. Retrieved June 29, 2013, from http://www.waterstechnology.com/buy-side-technology/opinion/2250200/sentiment-analysis-still-has-a-long-way-to-go-on-wall-street

[7] Mayer, J. D., & Gaschke, Y. N. (1988). The Experience and Meta-Experience of Mood 4. Journal of personality and social psychology, 55(1), 102–111.

[8] A. Selyukh, "Hackers send fake market-moving AP tweet on White House explosions | Reuters." [Online]. Available: http://www.reuters.com/article/2013/04/23/net-us-usa-whitehouse-ap-idUSBRE93M12Y20130423. [Accessed: 17-Sep-2013].

[9] O'Connor, B., Balasubramanyan, R., Routledge, B. R., & Smith, N. A. (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 (pp. 122–129). Retrieved from http://www.aaai.org/ocs/index.php/ICWSM/ICWSM10/paper/viewPDFInterstitial/1536/1842

[10] Or, A. (2011, December 12). Now Trending: Turning Tweets Into Trades. MarketBeat. Retrieved June 28, 2013, from http://blogs.wsj.com/marketbeat/2011/12/12/now-trending-turning-tweets-into-trades/

[11] Pang, B., Lee, L., & Vaithyanathan, S. (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 (pp. 79–86). Retrieved from http://dl.acm.org/citation.cfm?id=1118704

[12] Paul, M., & Dredze, M. (2011). You are what you tweet: Analyzing Twitter for public health. In Fifth International AAAI Conference on Weblogs and Social Media (ICWSM 2011). Retrieved from http://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/viewPDFInterstitial/2880/3264

[13] Shah, V. H. (2007). Machine learning techniques for stock prediction. Courant Institute of Mathematical Science, New York University. Retrieved from http://stocktrendresearch.googlecode.com/svn/trunk/Paper/SVM_ANN/MachineLearningTechniquesforStockPrediction.pdf
[14] Zhang, X., Fuehres, H., & Gloor, P. A. (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
[15] Lyashevskaya, O., Astafieva, I., Bonch-Osmolovaskaya, A., Gareishina, U. (Ляшевская, О., Астафьева, И., Бонч-Осмоловская, А., & Гарейшина, Ю.) (2010) "Dialog 2010) Conference proceedings: repost 49 Retrieved June 28, 2013, from http://www.dialog-21.ru/digests/dialog2010/materials/html/49.htm .