CASA RIREA EDITRICE

Italian Journal of Accounting and Economia Aziendale

International Area

Fondata nel 1901

NICOLA D'URSO dis. CROCE 15 - ROMA

Year - 2014 - 1-2-3

Could we make better prediction of stock market indicators through Twitter sentiment analysis?

ALEXANDER PORSHNEV - ILYA REDKIN - ALEXEY SHEVCHENKO*

ABSTRACT: *In our paper, we analyzed the possibility of improving prediction of stock market indicators by conducting a sentiment analysis of Twitter posts. We used a dictionary-based approach for the sentiment analysis which allowed us to distinguish eight basic emotions in users' tweets. We conducted a correlation analysis with different time lags to find the relation between quantity of tweets from different emotional categories and movement of stock market indicators. Quantities of tweets from different emotional categories were also used to train a Naïve Bayes machine learning algorithm to predict DJIA, NASDAQ and S&P500 indicators. Our results indicate that the constructed model provides additional information and increases the level of prediction in comparison to a model based solely on information about previous shifts in stock indicators.*

1. Introduction

Predicting financial markets is an interesting task from both practical and theoretical perspectives. New information technologies provide users with a wide range of possibilities to express themselves and this leads to enormous amounts of data available about emotions, moods and psychological states of Internet citizens. In the USA, which has a huge influence on global economy, the Internet penetration rate is 78.3%, and active Internet users are also active financially. We therefore think that Twitter is great resource of additional information that may help us to make better forecasts for financial markets. Although this idea was formulated several years ago by there is no coherent opinion about the possibility to improve financial market forecasts (Bollen, Mao, & Zeng, 2011).

In the last years significant progress was demonstrated in using Twitter as additional source of information (O'Connor, Balasubramanyan, Routledge, & Smith, 2010; Paul & Dredze, 2011). 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%. 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 (Zhang, Fuehres, & 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 it their paper they do not provide information about the accuracy of prediction (Chen, Ray & Lazer, Marius, n.d.).

Regarding the application of sentiment analysis as a money generator we did not find examples of successful projects. 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 (Malakian, 2013). Later the fund was rebranded into DCM Capital and presented to the retail investors sentiment-based trading platform (Malakian, 2013).

---

* Articolo ad invito.

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 (Malakian, 2013). However, in his article Malakian admits that there is no evidence to conclude that failure of Derwent Capital Markets happened because of poor technology (Malakian, 2013). 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 (Malakian, 2013).

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 (Or, 2011).

This lead us to the main hypothesis of our research, that analyses of tweets increase the accuracy of predication for stock market indicators. It is worth mentioning that the reliability of using Twitter is not easy task, as algorithms of analysis are proprietary and their direct evaluation is impossible. To test our main hypothesis we have to accomplish following tasks:

1. Download representative amount of raw data from Twitter.
2. Develop an algorithm for sentiment analysis, based on psychological classification of emotions.
3. Analyze accuracy of prediction of machine learning algorithms using data received from market and sentiment analysis.

2. Methodology

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

Twitter sentiment analysis

Research in natural language processing provides several directions for sentiment analysis, first is classification based on human developed gold standard (Pang, Lee, & Vaithyanathan, 2002). 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 (Jurafsky & James, 2000). Creation of a gold standard is usually associated with a lot effort and work of a team of linguistics (e.g. Lyashevskaya et al. , 2010).

The second approach is based on dictionaries. 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" (Zhang et al., 2011). In our study we realize more complex algorithms based on eight dictionaries we created.

For the analysis of efficiency of algorithms we used standard measures recall, precision and F-measure (Jurafsky & James, 2000).

$Recall_{calm} = A/(A+C)$, where A is the amount of tweets correctly recognized as belonging to class "calm" and C amount of tweets not recognized by our algorithm, but marked by an expert as inherent to this class.

$Precision_{calm}=A/(A+B)$, where A is the amount of tweets correctly recognized as belonging to class "calm" and B is the amount of tweets recognized by our algorithm, but marked by an expert as not inherent to this class.

$F\text{-}measure_{calm}=2/(1/(Precision_{calm})+1/(Recall_{calm}))$

The dictionary 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 (Bollen et al., 2011).

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 Naïve Bayes and Support Vector Machines.

In order to answer the question: "Do sentiment analysis of tweets provide additional information?", we use learning algorithms on four 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 random data to the basic set of data (Basic&Random). The third set was created by adding a normalized number of tweets with words "Worry", "Hope", "Fear" to the basic set (Basic&WHF). The fourth 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 four learning sets will be different. According to our hypothesis about the existence of additional information in Twitter, we expect that the first and second data sets will provide almost the same accuracy level, somewhat higher accuracy for the data set Basic&WHF 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 (Bollen et al., 2011). To test these findings, data from Twitter could help to improve stock market predictions when we train the Naïve Bayes algorithm with different time lags from one to seven days.

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 05/05/2013 we downloaded 275'207 700 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.

In a period from 14/02/2013 till 25/04/2013 we have 49 business days, which were divided into training (34 days) and test data (15 days).

3. Analysis

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 (Mayer & Gaschke, 1988). We also added all synonyms of selected adjectives from the WordNet dictionary (Miller, 1995).

To test the quality of the sentiment analysis of our algorithm we manually created a gold standard from 270 tweets, 30 per sentiment category. Each from 270 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 results 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 "happppppyyyyyyy". 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 1. Measurement of performance for sentiment analysis with first dictionaries and with second dictionary

First version of dictionnaires								
	happy	loving	calm	energetic	fearful	angry	tired	sad
Precision	87%	77%	63%	57%	61%	70%	69%	85%
Recall	87%	77%	40%	57%	57%	63%	67%	73%
F-measure	87%	77%	49%	57%	59%	67%	68%	79%
Second version of dictionnaires								
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 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 2. Comparison of sentiment analysis efficacy 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.

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&Random, Basic&WHF, Basic&8EMO. Each data set had 7 sub tables for lag in time from one to seven days.

To train Naïve Bayes algorithm 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 used the first 34 days as a training sample and later 15 days as test sample. Application of Naïve Bayes trained just on Basic data set showed low accuracy for all lags from one to seven days. Where accuracy was measured:

Accura[cy]_growth=$(A+\bar{A})/(A+\bar{A}+C+\bar{C})$, where A is the amount of days correctly recognized as belonging to class "growth", $\bar{A}$ is the amount of days correctly recognized as not belonging to class "growth", C is the amount of days not recognized correctly by our algorithm, but marked as days with growth of stock market, $\bar{C}$ - amount of days recognized by our algorithm, but not inhered to class "growth". The sum of $A+\bar{A}+C+\bar{C}$ is the amount of days in a test sample, equal to 15.

The addition of random numbers from the range of the minimal amount of tweets to the maximal amount does not provide better accuracy.

Table 3. Accuracy of Naïve Bayes algorithm in dependence from training data set.

Data set\ Lag	1 day	2 days	3 days	4 days	5 days	6 days	7 days
Basic	33.33%	46.67%	46.67%	40.00%	40.00%	46.67%	46.67%
Basic&Random	46.67%	40.00%	40.00%	40.00%	33.33%	46.67%	46.67%
Basic&WHF	26.67%	46.67%	53.33%	46.67%	53.33%	46.67%	53.33%
Basic&8EMO	26.67%	40.00%	66.67%	60.00%	46.67%	53.33%	53.33%

Application of data on DJIA allow us to test our main hypothesis that sentiment analysis of twits could provide information to make a better prediction of a stock market.

Next we apply the same strategy of Naïve Bayes testing for SP&500 and NASDAQ indexes. Growth of accuracy in increasing lag for S&P500 was also tested for 8 and 9 day lags. We found that the maximum accuracy for predicting the SP&500 index was for lag 8 days. With NASNAQ we observed a different situation from the one-day delay: our data provide the possibility of forecast with 60% accuracy. Manipulating the lag we observe accuracy of 60% on three occasions, but the maximum accuracy was observed with the lag of 8 days, as well as for SP&500 (see Table 4.).

Table 4. Accuracy of Naïve Bayes algorithm in predictions of DJIA, SP&500, NASDAQ.

Index\ Lag	1 day	2 days	3 days	4 days	5 days	6 days	7 days	8 days	9 days
DJIA	26.6%	40.0%	66.6%	60.0%	46.6%	53.3%	53.3%	33.3%	33.3%
SP&500	33.3%	40.0%	33.3%	26.6%	40.0%	46.6%	53.3%	64.2%	53.8%
NASDAQ	60.0%	53.3%	53.3%	60.0%	33.3%	40.0%	60.0%	73.3%	71.4%

4. 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. Application of sentiment analysis data for training Naïve Bayes algorithms

allow us to receive accuracy of stock market predictions for DJIA - 66.6%, SP&500 - 64.2%, and NASDAQ - 73.3%. This accuracy is below that reported by Bollen and co-authors, which was 87,6%, but we could not run our algorithm on the same data as Bollen and his colleagues, as they do not publish their algorithm or and the raw Twitter data is unavailable (http://terramood.informatics.indiana.edu/data) .

Although, we report interesting and probably promising results they should be interpreted with caution. As we used 34 days for training and 15 days for prediction, an error in one day leads to a 6.6% measurement error, so in best case we could think that our accuracy is not fixed to 66.6% for DJIA, but lies in between 60% and 73.6%. The measurement error of about 6.6% could be an explanation for the fluctuations of the accuracy (see results presented in Tables 3,4).

We found that a Basic data set provides less information than Basic&WHF.

Basic&WHF provides less information than our Basic&8EMO dataset. In the work of Zhang, Fuehres, Gloor the authors suggest that even sentiment analysis with recognition of three words "Worry", "Hope", and "Fear" could provide additional information and our results support their findings. This provides a new argument and if future attempts to investigate this issue show the same trend it can be proposed that human sentiments can be used to predict the stock market.

We think that at this point it is too early to suggest that we confirmed our hypothesis 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 14 February until 25 April 2013 we have to download 275'207 700 tweets (approximately in 30 times more).

5. Conclusion

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

First we created server application to download tweets and store them. In the period from 14/02/2013 till 25/04/2013 we downloaded 275'207 700 tweets (on average we downloaded 3483642 per day). To analyze this huge amount of data we needed fast and reliable algorithm of sentiment analysis. To solve this task we used an approach based on dictionaries and the second version of dictionaries showed satisfactory performance.

Our preliminary results indicate that our hypothesis could be confirmed. In our further research we plan to analyze the correlation between results of sentiment analysis and events, to check the reliability and validity of our dictionaries algorithm, to continue data collection and to make a scrupulous analysis of confidence intervals for accuracy of prediction in a larger time frame, and to look for possible solutions to transfer our results in to a profitable trading strategy.

ALEXANDER PORSHNEV
Associate Professor, PhD in Psychology

National Research University Higher School of Economics
Social Science Department - Nizhni Novgorod (Ru)

ILYA REDKIN
MSc. in business informatics

National Research University Higher School of Economics
Business Informatics Faculty and Applied Mathematics
Nizhni Novgorod (Ru)

ALEXEY SHEVCHENKO
MSc. in computational linguistics

National Research University Higher School of Economics
Business Informatics Faculty and Applied Mathematics
Nizhni Novgorod (Ru)

References

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

CHEN, RAY, & LAZER, MARIUS. (n.d.). Sentiment Analysis of Twitter Feeds for the Prediction of Stock Market Movement. stanford.edu. Retrieved January 25, 2013, from http://www.google.ru/url?sa=t&rct=j&q=sentiment%20analysis%20of%20twitter%20feeds%20for%20the%20prediction%20of%20stock%20market%20movement&source=web&cd=1&cad=rja&ved=0CDsQFjAA&url=http%3A%2F%2Fcs229.stanford.edu%2Fproj2011%2FChenLazer-SentimentAnalysisOfTwitterFeedsForThePredictionOfStockMarketMovement.pdf&ei=UT4CUbvcKeGs4ASImYCICQ&usg=AFQjCNFSwgoqVn8OV0s7xpsax1Hbg6eJbw&bvm=bv.41524429,d.bGE

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

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

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.

MILLER, G. A. (1995). WordNet: A Lexical Database for English. Communications of the ACM, 38(11), 39–41.

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

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/

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

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

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

LYASHEVSKAYA, O., ASTAFIEVA, I., BONCH-OSMOLOVASKAYA, A., GAREISHINA, U. (2010) "Dialog 2010) Conference proceedings: repost 49 Retrieved June 28, 2013, from http://www.dialog-21.ru/digests/dialog2010/materials/html/49.htm