Indicators for measuring the online activity of investors

Lecturer Mihail Dumitru Sacală, PhD
Department of Statistics and Econometrics,
Academy of Economic Studies, Bucharest
sacalamihai@yahoo.com

Abstract

The values of endogenous variables: price and trading volume for shares listed on a stock exchange can be determined by modeling the exogenous variables from the online activity of investors. This online activity refers to the number and content of messages posted by investors on shares analysis dedicated web sites.

The evolution of the comments made by investors upon the articles which have as subject the capital market is a key element in analyzing the investor's online activity.

The determined indicators examine the frequency of these messages and votes obtained by each of them from other participants in online discussion.

Keywords: Asset Pricing, Behavioral Finance, Qualitative Variables, Investors.

JEL Classification Codes: G - Financial Economics, G12 - Asset Pricing; Trading volume; Bond Interest Rates

Acknowledgments

This paper was co-financed by European Social Fund, the Human Resources Development Operational Program 2007-2013, the number POSDRU/89/1.5/S/59184 "Performance and excellence in postdoctoral research in economic sciences in Romania."

1. Introduction

The motivation for our research is pointed out by Daniel Dennett Professor of Philosophy at Tufus University, who said in June 2011 in Bucharest that:

"The individual is responsible for the way he uses the "moral competence": the ability to respond to reason and the ability to recognize and counteract manipulation by other individuals."

This is not entirely true because it is not a question of psychology but one of sociology. A person seeks membership in a social group and people are connected in a global network.

The second reason is that the IQ is different form one person to another and also the experience is obtained by doing wrong in the first place.

Our main goal is indeed to quantify and to exclude the manipulation effect from the stock transactions.

The main approaches in order to quantify the impact of behavioral finance on capital markets are:

- Conducting a survey among the stocks' investors
- Analysis of investors expectations influenced by the trading price and volume
- Investors' behavior by analyzing their online activities (search engines, published articles, forum activity etc.)

To realize a survey is still a challenge because as Devi (2008) said, in the world only 26% of investors are women, although the returns obtained by women traders are beyond that of men (2010) Christiansen.

The problem encountered in attempts to achieve this survey is that men tend to be too self confident, due to overconfidence theories based on biased self-attribution, Glaser et al., (2009)

Regarding the second approach: Analysis of investors' expectations influenced by the trading price and volume, our research started studying some values for which investors are paying extra attention. This method is presented in our paper *"Technical analysis and econometric prediction using wave refraction method", Sacala (2011).* The greater the distance between last time the stock achieved the integer value and now is, the lower the probability to break resistance will be.

Sornette mentioned in 2002 that investors are all connected in a big network.

The result obtained using this approach is a technical analysis method based on wave refraction:

- This method is based on the "Elliot Wave Principle" which says that collective investor psychology or psychology of the masses change between optimism and pessimism in a natural way.
- These changes of mood create patterns evident in prices movement.

Often the future evolution of prices and trading volume for a share are highlighted through the online activity of investors. This activity can be captured in one of two situations:

a) Analysis of listed companies search through search engines or social networking sites (Google, Twitter, Facebook, etc.). Thus, in 2010, Bollen et al., determined a forecasting model for U.S. capital market indexes based on twitter users activity.

Tobias Preis (November 15, 2010) analyzed the searches on Google for S & P 500 component companies along with their price and volume. He failed to determine the weekly price fluctuations but found a strong correlation between the number of Internet searches (Google Trends) and the trading volume for the company's shares.

An academic group access in Google databases is required in order to predict shares price trend due to investor perceptions. Stock market crash can be also prior determined when we have information on the number of searches in each hour for each stock.

b) The modifications of the comments made by investors upon capital market news and papers. The evolution of the comments made by investors upon the articles which have as subject the capital market is a key element in analyzing the investor's online activity.

The determined indicators examine the frequency of these messages and votes obtained by each of them from other participants in online discussion.

This paper studies the second situation. The analyzed messages are representative samples for the opinions of investors for the specific company. Active people on websites and other electronic media are those that invest for short and medium term, have a higher average number of transactions than other investors and thus they affect trading volume and price.

2. Message classification

Regarding the forum messages we are using the Interactive Investor platform (www.iii.co.uk). This is one of the largest online traders in UK. We want to obtain the investors' sentiment out of the online messages and to determine the manipulation effect.

There is some literature written on this topic but not much.

Depken et al. (2008) found out that:

- authors with high reputation scores are less likely to voluntarily offer a buy-hold-sell sentiment in a particular message
- authors with no reputation at stake tend to be more bearish with their sentiment but, after controlling for selection, authors with more reputation at stake tend to be bullish in their sentiment.
- high-reputation authors tend to offer more accurate sentiments.

Felton et al. (2002) found out for Enron company that there were repeated online warnings for investors to get out while they can.

Dellarocas (2004) generated accurate predictions of a movie's total revenues from statistics of user reviews posted on Yahoo! Movies during the first week of a new movie's release.

Another pointed idea is that the overnight message posting volume is found to predict changes in next day stock trading volume and returns Wysocki (1998)

Investors interested in a particular stock, wanting to obtain as much information about that company, read articles on the Internet using both stock exchanges exclusive websites and electronic publications.

Many of these websites allow users to post messages and comments for articles that refer to a certain company.

These messages fall into two categories:

- Comments related to an article;
- Responses to another post.

This paper presents a methodology for calculating the index and its interpretation for each of the two situations described above.

Analyzing the variables identified by studying these kinds of messages we can determine an indicator that will give us signals about how the price and the trading volume for a specific company's shares will be modified in the immediately following period.

The intensity of online activity is closely linked to investor's interest for the specific instrument. A lot of decisions to buy or sell shares are taken from the interpretation of these messages. This causes both a change in the volume traded and the trading price.

After their contents, these messages fall into two categories:

- messages which show an expected effect of short-term yields - these messages come from one of the following situations: 1. the investor present his view on the share's trend, 2. hopes that this movement will be achieve, 3. thinks that this message can influence the market in its favor. These messages are posted by people who want to trade this action in the near future.
- neutral messages - these messages are not related to that company or they do not present a point of view for the discussed share price movement.

According to Das et al. (2001), the classifier achieves an accuracy of 62% which is higher compared with a random classification accuracy of 33% and it gets closer to the 72%, the human agreement on message classification. They pointed out 5 algorithms to classify the messages:

- NC – Naive Classifier;
- VDC – Vector Distance Classifier;
- DBC – Discriminant Based Classifier;
- AAPC – Adjective Adverb Phrase Classifier;
- BC – Bayesian Classifier.

Some authors use different toolkit for classifying the messages. One of them is Bow - a toolkit for statistical language modeling, text retrieval, classification and clustering. The library and its front-ends were designed and written by Andrew McCallum, with some contributions from several graduate and undergraduate students. *Bow* (or *libbow*) is a library of C code useful for writing statistical text analysis, language modeling and information retrieval programs. The current distribution includes the library, as well as front-ends for document classification (*rainbow*), document retrieval (*arrow*) and document clustering (*crossbow*).

The program has the following features:

- Recursively descending directories, finding text files.
- Finding document' boundaries when there are multiple documents per file.
- Tokenizing a text file, according to several different methods.
- Including N-grams among the tokens.
- Mapping strings to integers and back again, very efficiently.
- Building a sparse matrix of document/token counts.
- Pruning vocabulary by word counts or by information gain.
- Building and manipulating word vectors.
- Setting word vector weights according to Naive Bayes, TFIDF, and several other methods.
- Smoothing word probabilities according to Laplace (Dirichlet uniform), M-estimates, Witten-Bell, and Good-Turning.
- Scoring queries for retrieval or classification.

- Writing all data structures to disk in a compact format.
- Reading the document/token matrix from disk in an efficient, sparse fashion.
- Performing test/train splits, and automatic classification tests.
- Operating in server mode, receiving and answering queries over a socket.

3. Identified variables

The variables we may identify vary depending on how the blog, web page etc is built. Theoretical, the variables which may be analyzed are:

- The message date - is the date when the message was posted as a comment to an article or on a dedicated page. This variable shows the distance in days between the time when the article was posted and the time when this message was posted. The greater the distance is the greater impact of the article will be and therefore the indicator value for that article online activity will be higher.

- The number of responses received from investors by every posted message. If the number of responses is high, the message has an increased impact on the perceptions of investors and therefore it will lead to an increase of the online activity indicator for the studied paper.

- The number of votes received by each users posted message. A large number of votes involve a huge interest in the message.

- The grade for each message (the difference between the number of positive votes and the number of negative votes). A low grade in terms of a large number of votes indicates an increased activity with big standard deviation also.

- The sign of planned returns for specified share, subtracted from the message.

- The number of times each message was open/read. This is a very important variable because a message is not open by accident and it is open just once by a single investor.

- The length of each message. This is important because according with Nathaniel Bulkley and Marshall W. Van Alstyne (2010), the short messages containing a forecast have a bigger impact on investors psihology. The length can be measured in words or characters.

4. Research plan

In order to determine the desired output we must complete the following steps:

a) obtaining the data for price and trading volume for each analyzed stock in the same period as the messages are. We must get the data for Reuters, Data Stream, Bloomberg etc. The values will be on a daily frequency. It is a must to obtain the volatility and other indicators to consider them as exogenous variables when modeling with the messages variables.

b) obtaining the values for each variables, from the ones mentioned in the last chapter, for messages. We must also build the

data base for each stock and each year in order to calculate the indicators in a more facile way.

c) getting the forecast sign of the future return for a specific stock in each message. This means to put each message in buy, hold or sell class. One of most simple and efficient algorithms is Naive Bayes Classifier. In spite of their naive design and apparently over-simplified assumptions, naive Bayes classifiers have worked quite well in many complex real-world situations. Analysis of the Bayesian classification problem has shown that there are some theoretical reasons for the apparently unreasonable efficacy of naive Bayes classifiers, Zhang (2004). Still, a comprehensive comparison with other classification methods in Niculescu (2006) showed that Bayes classification is outperformed by more current approaches, such as boosted trees or random forests.

An advantage of the naive Bayes classifier is that it only requires a small amount of training data to estimate the parameters (means and variances of the variables) necessary for classification. Because independent variables are assumed, only the variances of the variables for each class need to be determined and not the entire covariance matrix.

d) calculating the proposed in chapter 5 indicators or some derivate ones.

e) determining the influence of the calculated indicators over the exogenous variables.

The methods we are using (some of them are used in the literature for related researches) are:

- Different types of regression
  - linear panel regressions,
  - negative binomial panel regressions,
  - Tobit panel regressions, Glaser et al. (2010)
- Nonparametric correlation
- Neural Networks (using some test samples)

We expect to have all the messages from Interactive Investor at the end of September 2011 and to make the largest analysis on UK market using the www.iii.co.uk - a single online trading platform in the UK which have, in the last decade, for the FTSE100 shares components, over 10.5 million messages.

The expected results are:

- To determine if the messages have an influence over the stock trend
- To examine how investors sentiment (as measured by the messages) reacts to changes in market trends – e.g. does sentiment become more positive following recent price rises etc
- To see if the correlation is as strong (weak) for upper trends as for decreasing trends
- To compare the results with other forms of manipulation (advice) to see if anonymous tips have a greater impact than tips from known persons, advisers etc

5. Neighborhoods of extreme points for a stock price

5.1 Commented papers

The papers which discuss a particular stock appear in various electronic publications without having continuity. In this particular situation we can not build an indicator to characterize continuously the online activity for this company. The only method of analysis is to build an indicator, the values of which will classify the papers and therefore to identify changes that will occur in the price and trading volume for shares of the company.

For example we used www.hotnews.ro, the biggest Romanian online news portal. Dates of occurrence (picked-up in accordance with important macroeconomic changes in order to compare the results in a more facile way) of the analyzed four items are:

1. December 19, 2007
2. July 10, 2008
3. April 27, 2009
4. March 2, 2011

Appendix 1 provides an example of message analysis. The message appeared as a comment on one of the analyzed four items.

$I_{articol}$ is the indicator which characterizes the activity of investors in one online paper. Based on the values of this indicator we may determine the following changes for the trading volume and the share price.

Proposed formula for this indicator is:

$$I_{articol} = \frac{\sigma_{sens}^2 * \sum_{i=1}^{n} [d_{ci} - (d_a - 1)](n_{raspi} + 1) * k_i * (\frac{nvot_i + 1}{nota_i + 1})}{\sum_{i=1}^{n} (nvot_i + 1)}; \quad (1)$$

Where:

1. n- the total number of messages recorded for the current paper;
2. σ_{sens}^2 dispersion of the binary variable: "forecasted sign of the future return".

$$\sigma_{sens}^2 = \frac{n_+ * n_-}{n_{sens}^2}; \quad (2)$$

Where: n_+ - the number of messages that predict an increase of the price of the stock;

n_- - the number of messages that predict an decrease of the price of the stock;

n_{sens} - the number of messages that predict the sign for future returns of the specified stock.

3. d_{ci}- the date when the "i" message was posted;
4. d_a- the date when the discussed paper appeared;
5. n_{raspi} - the number of responses for the "i" message;
6. k- the factor that highlights the importance of forecasting the direction of the future return:

   k=2 if the message predicts the sign for the future performance;
   k=1 otherwise;

7. $nvot_i$ - the number of votes obtained by the "i" message;

8. $nota_i$ - the grade obtained by the "i" message.

Given the methodology of defining the indicator $I_{articol} \in [0; +\infty)$, and its values may fall an article in the following categories:

1. $I_{articol} \in [0; 0,3]$, the online activity is low, the trading volume will be low, while the sign of return can not be established due to high volatility.

2. $I_{articol} \in (0,3;1]$, the online activity is moderate, the trading volume will fluctuate around the average and the price will follow the short-term trend.

3. $I_{articol} \in (1; +\infty)$, the online activity is high, we are in the neighborhood of an extreme point for the price trend and this indicates that the sign for the return on short term will be opposite to the present average return. We expect a high trading volume.

Appendix 2 shows the recorded values for the identified variables for each of the four articles. The values of indicators in the four cases are:

- 1. 0,99
- 2. 1,33
- 3. 0,79
- 4. 1,6

These values show us that in March 2011 the price for the analyzed stock changed the trend from a down trend to an up trend. It was a moment of high online activity after months of low online activity. We may now sustain this affirmation because the price for that stock started to grow in March.

5.2 Specialized online forums

In online, dedicated to stock exchanges, forums (eg www.agf.ro Blogs General Assembly) messages related to certain companies appear every day. In this situation it is necessary to build an indicator that can be calculated daily and therefore online activity can be characterized at all times.

Appendix 3 contains an example of message analysis. The message appeared as a comment on the forum page dedicated to FP stock.

Proposed indicators have the following formula:

$$a)\ I_t = \frac{n_t}{\frac{\sum_{i=t-19}^{t} n_i}{20}} * \sigma^2_{senst}; \qquad (3)$$

Where:

1. n_t- the analyzed number of messages recorded for the stock on the day t;

2. σ^2_{senst} - dispersion of the binary variable: "forecasted sign of the future return" calculated for day t.

$$\sigma^2_{senst} = \frac{n_{t+} * n_{t-}}{n^2_{tsens}};$$

Where: n_{t+} - the number of messages that predict an increase of the price of the stock on day t;

n_{t-} - the number of messages that predict an decrease of the price of the stock on day t;

n_{tsens} - the number of messages that predict the sign for future returns of the specified stock during day t.

3. $\frac{\sum_{i=t-19}^{t} n_i}{20}$ (4)

- the average number of messages in the last twenty days prior to the analyzed day "t".

Given the methodology of determining the indicator $I_t \in [0; 5]$, and its values are interpreted as follows:

1. $I_t \in [0; 0,1]$, the online activity is low, trading volume will be low, while the return sign can not be established due to high volatility.

2. $I_t \in (0,1;0,25]$, the online activity is moderate, trading volume will be fluctuating around the average and the price will follow the short-term trend.

3. $I_t \in (0,25;5]$, the online activity is high, we are in the neighborhood of an extreme point for the price trend and this indicates that the sign for the return on short term will be opposite to the present average return. We expect a high trading volume.

b) If the listed shares webpage contain all the above variables, the indicator has the following proposed formula:

$$I_t = \frac{n_t}{\frac{\sum_{i=t-19}^{t} n_i}{20}} * \sigma^2_{senst} * \overline{n_{votest}} * \overline{n_{responsest}} * d_{MEt}; \quad (5)$$

Where:

1. n_t- the analyzed number of messages recorded for the stock on the day t;

2. σ^2_{senst} - dispersion of the binary variable: "forecasted sign of the future return" calculated for day t.

$$\sigma^2_{senst} = \frac{n_{t+} * n_{t-}}{n^2_{tsens}};$$

Where: n_{t+} - the number of messages that predict an increase of the price of the stock on day t;

n_{t-} - the number of messages that predict an decrease of the price of the stock on day t;

n_{tsens} - the number of messages that predict the sign for future returns of the specified stock during day t.

3. $\frac{\sum_{i=t-19}^{t} n_i}{20}$ - the average number of messages in the last twenty days prior to the analyzed day "t".

4. $\overline{n_{votest}}$ - the average number of votes for the messages from t day.

5. $\overline{n_{responsest}}$ - the average number of responses obtained by a message in t day.

6. d_{MEt} - the median of the distances between messages in day t. It is calculated in hours.

Given the methodology of determining the indicator $I_t \in [0; 5]$, and its values are interpreted as follows:

1. $I_t \in [0; 1]$, the online activity is low, trading volume will be low, while the return sign can not be established due to high volatility.

2. $I_t \in (1;3]$, the online activity is moderate, trading volume will be fluctuating around the average and the price will follow the short-term trend.

3. $I_t \in (3;+\infty]$, the online activity is high, we are in the neighborhood of an extreme point for the price trend and this indicates that the sign for the return on short term will be opposite to the present average return. We expect a high trading volume.

6. Conclusions

This paper studies the impact of behavioral economics over the price and especially over the trading volume for a specific stock. The presented indicators attempt to explain this influence.

The main problem that arises in calculating these indicators is given by finding the proper vocabulary in order to determine the meaning of the messages in terms of expected sign of the future return. For example, for a single online trading platform in the UK in the last decade, the FTSE100 shares components have over 10.5 million messages.

The advantage of using these methods is that these alone may determine the foregoing neighborhood of a local extreme point for the share price trend.

7. References

Anghelache, Gabriela; 2009 „Piaţa de capital în context european” , Editura Economică, Bucureşti

Antweiler, W.; Frank, M., 2001 „Is all that talk just noise? The information content of internet stock message boards” University of British Columbia

Ariely, Dan, 2008 „Predictably Irrational: The Hidden Forces That Shape Our Decisions “, HarperCollins

Das, S.; Chen, M, 2001 „Yahoo! For Amazon: Opinion extraction from small talk on the web” Santa Clara University

Dellarocas, C.; Awad, N.; Zhang, 2004, X. „Using online reviews as a proxy of world-of-mouth for motion picture revenue forecasting” MIT

Felton, J.; Kim, J., 2002 „Warnings from the Enron message board” Central Michigan University

Wysocki, Peter, 1998 „Cheap talk on the web: the determinants of posting on stock message board”, University of Michigan Research

Appendix 1 - Example of message analysis

"Records FP (Wednesday, March 2, 2011, 6:28 p.m.)
Nae [anonymous]

A record for FPR volume, but the price is still low due to bad company contract made by ... "non-owners, obviously. "

Appendix 2 - observed values for the variables identified in an article

Comment Date	Number of answers	Number of votes	Vote result	Foresight	dc-da+1	Number of answers +1	K
27	4	9	7	s	1	5	2
27	0	3	1	s	1	1	2
27	2	1	1	c	1	3	2
27	0	5	-1	s	1	1	2
28	0	0	0	c	2	1	2
27	2	4	2	c	1	3	2
27	0	0	0	n	1	1	1
27	0	3	3	s	1	1	2
28	0	0	0	s	2	1	2
27	0	4	2	s	1	1	2
27	0	7	5	s	1	1	2
27	0	5	3	s	1	1	2
27	1	1	1	c	1	2	2
27	0	1	-1	s	1	1	2
27	3	3	-1	n	1	4	1
27	1	0	0	n	1	2	1
27	0	1	1	n	1	1	1
27	0	2	0	s	1	1	2
27	2	5	-1	s	1	3	2
27	1	0	0	n	1	2	1
27	0	1	1	n	1	1	1
27	1	6	0	c	1	2	2
27	0	2	2	s	1	1	2
27	4	0	0	s	1	5	2
27	3	1	1	c	1	4	2
27	2	0	0	C	1	3	2
27	0	0	0	S	1	1	2
28	0	0	0	S	2	1	2
29	0	0	0	N	3	1	1

Appendix 3 - Example of message analysis, recorded on www.agf.ro portal

24-May-2011 4:23 p.m. - stef_nic: if the stock market next year remains at this size 0,2 RON is likely for fp