SYM-OP-IS 2011
XXXVIII Symposium on Operational Research
Faculty of Economics
University of Belgrade

NEWS ANALYTICS AND TEXT SENTIMENT VISUALIZATION IN FINANCE

ČASLAV BOŽIĆ
Institute AIFB, Karlsruhe institute of Technology (KIT), Karlsruhe, Germany, bozic@kit.edu

DETLEF SEESE
Institute AIFB, Karlsruhe institute of Technology (KIT), Karlsruhe, Germany, detlef.seese@kit.edu

Abstract: *One of the tasks of news analytics in finance is sentiment extraction and determining whether the mention of a company in a news article has rather positive or negative tone. This task is increasingly important because previous works show relation between such news sentiment and market movements, especially pressure on equity prices. We explore the benefits of using sentiment reversal indicator instead of raw sentiment score that is produced by most of the systems. We propose the visualization methodology that makes use of sentiment reversals and describe the visualization system based on this methodology.*

Keywords: *News Analytics, Sentiment Reversal, Visualization*

1. INTRODUCTION

News analytics is established as a discipline of using machines to analyze and summarize news. It can provide timely quantification of up to now mainly qualitative data, like sentiment, importance, or novelty of the news article. This approach is becoming increasingly important, especially in finance where more and more trades are executed by algorithms.

We will focus on sentiment extraction as one of the most complex disciplines of news analytics, and we will propose the system design that can support traders in decision making by visualizing news sentiment data. Chapter 2 contains a definition and an explanation of sentiment in financial setting. Chapter 3 explores benefits of using sentiment reversal indicator instead of raw sentiment score. In Chapter 4 we propose the methodology that makes use of sentiment reversals and describe the visualization system based on this methodology. Chapter 5 concludes this paper.

2. TEXT SENTIMENT IN FINANCE

Since 1997 and the system described in [1] which employs methodology of probabilistic datalog rules, the academic literature describes more than a dozen of different systems that employ text mining approach in financial area. Although some of them do not explicitly use the term 'text sentiment', if we observe text mining methodologies as transformations that assign a numerical value to every textual string, we can refer to that numerical value as a sentiment score. All these publications have at least implicit statements about the predictive power of the specific sentiment score on e.g. returns or volatility. There is also another group of publications, starting with pivotal article [2], which is more concerned with detailed exploration and theoretical grounding of the relationship between sentiment scores and financial indicators. This second group usually employs simple vocabulary-based sentiment scores, or uses the output from commercially available proprietary systems. Some of these systems are established as industry standards, but the used methodology is very rarely described in the academic literature.

The overview of such systems up to the year 2006 is given in survey [3], which compares eight text-mining systems, including the one created by the authors themselves. Since more technical performance criteria are often missing, it is not possible to draw clear conclusions about relative performance. [2] observes Wall Street Journal's column "Abreast of the Market", uses content analysis software General Inquirer together with Principal Component Analysis approach, and finds that high pessimism in published media predicts downward pressure on equity prices.

3. SENTIMENT REVERSALS

Since the scientific exploration of text mining and sentiment extraction in finance lasts for over a decade, as noted in [4] "...simple 'buy on the good news' and 'sell on the bad news' strategies won't likely generate significant alpha as news analytics become more widely adopted." That is one of the reasons for the great number of publications in recent years which try to find novel approaches and exploit sentiment scores provided by existing systems.

One of the recent publications, with an approach similar to ours, is [5]. The author uses sentiment reversals as buy signals, and proves the performance of his approach by simulating set of portfolios over the timeframe 2000-2008. The source of sentiment data is Dow Jones News

Analytics historical database. Each story is assigned a positive or negative sentiment indicator as +1 and -1, respectively. Each story is also related to the company mentioned in the text. This approach builds one timeseries of sentiment indicators per company. The author uses average value of the sentiment timeseries over the certain time window in range from 1 to 24 months. Sentiment reversal is defined as an end of the period in which this average sentiment is constantly positive or negative. Graphically, this corresponds to all points where the graphical representation of the running window average sentiment intersects x axis. There are some additional constraints introduced, like minimal duration of reversal period of 30 days, and minimally 30 published stories in this period.

The author's findings suggest that the portfolios built using sentiment reversal based signals for buying and short selling the companies' stocks can outperform the market both in bull and bear market conditions. We explore a similar approach, although using a different dataset and basing our performance assessment on a more general and theoretical approach described in [6] instead of portfolio simulation.

Sentiment Reversal and Returns

The question we want to answer in this chapter is whether sentiment data transformed by extracting sentiment reversals can be more useful for predicting future returns for a company than the raw sentiment score data. We work with sentiment data provided by Thomson Reuters, and evaluate it according the framework described in [6].

The historical database of Thomson Reuters News Scope Sentiment Engine is a source of sentiment scores for all English language news published via Reuters News Scope system since 2003. For each company mentioned in the news, this database contains probability that the author write about the particular company in rather positive or rather negative context. The sentiment score *S(n,c)* for news article *n* and company *c* is built as a difference between the probability that the mention is positive $\Pr_{pos}$ and the probability that the mention is negative $\Pr_{neg}$. Since we work with daily data, the sentiment score is averaged within each trading day, thus producing daily sentiment score for a company *S(t,c).* With *T* we denote a set of all messages published during the trading day *t*.

$$S(n,c) = \Pr_{pos}(n,c) - \Pr_{neg}(n,c)$$

$$S(t,c) = \sum_{n \in T} S(n,c)$$

Besides sentiment scores this dataset offers additional metadata. Most important for us are the publication timestamp and the identifiers of all the companies mentioned in the news. We form a subset of all news available in the archive by choosing only those news items related to companies that are constituents of the Russell 3000 index. The Russell 3000 Index consists of the largest 3000 U.S. companies representing approximately 98% of the investable U.S. equity market.

The sentiment reversal measure is defined in the following way: for each company a running average sentiment score $S_{avg}(t,c)$ is calculated for each day *t*. In the moment of a news article publication, instead of raw sentiment score as defined in the previous paragraph, we assign a new value that measures sentiment reversal, or sentiment deviation from the average value S_r. Moreover, to allow only for strong reversal signals, this value is clamped to zero if the result is between -0.5 and 0.5. The number of total trading days before the current day *t* is denoted as N_t.

$$S_{avg}(t,c) = \frac{1}{N_t} \sum_{t_l < t} S(t_l, c)$$

$$S_r(t,c) = S(t,c) - S_{avg}(t,c)$$

As a source of trading data we used Thomson Reuters Tick History database. We extract opening prices for all trading days in 2003 for each company from the Russell 3000 index. The opening prices are adjusted for dividends and then transformed into log-returns. In this way we get open-to-open R_{OO} returns for each trading day in 2003 and each Russell 3000 company. The respective equation is given below, where P_O represents opening stock price and *t* represents the current trading day.

$$R_{OO}(t) = \ln\left(\frac{P_o(t)}{P_o(t-1)}\right)$$

We next align daily sentiment data (both raw data and reversal measure) with daily returns, and build the regression to explore the relation between these variables. If the observed sentiment measure actually correlates with the future stock returns, and if we represent the current day's return as a regression of previous sentiments (as in Equation 1, then the coefficients in front of the text sentiment measures should be significantly different from zero. We estimate regression parameters for linear regression with open-to-open return R_{OO} as a dependent variable using ordinary least squares method. As independent variables we use the contemporaneous daily sentiment score *S(t)*, the daily sentiment score from day before *S(t-1)*, two days before *S(t-2)*, and three, four, and five days before *S(t-3)*, *S(t-4)*, and *S(t-5)*. This is done with respect to the subject company *c*, which is represented as an additional parameter in the equation, besides time *t*. We order all the companies in our dataset according to their market capitalization (total market value of all shares of the company), and divide them into 10 equally sized groups. In this way we get the values for ten additional dummy variables dd_1 to dd_{10} (being 1 if the subject company falls into the respective group and 0 otherwise). We include them into the regression to account for the variations of returns as a result of company's size.

$$R_{OO}(t,c) = \sum_{i=1}^{5} \alpha_i S(t-i,c) + \sum_{i=1}^{9} \beta_i dd_i + \gamma \qquad (1)$$

We repeat the estimation of the regression parameters using sentiment reversal measure $S_r(t,c)$ instead of raw

sentiment score *S(t,c)*. The results are presented in Table 1 which offers a possibility for a comparison of these two approaches.

Table 1: Estimated regression parameters, open-to-open return as a dependent variable

Time Lag	Raw Sentiment Score	Sentiment Reversal
t	104.159***	100.850***
t-1	22.052**	34.596***
t-2	-9.126**	-5.034
t-3	-8.236**	0.173
t-4	-3.569	-2.887
t-5	-2.863	0.910

From Table 1 it is visible that both sentiment measures have a positive and significant relation to contemporaneous return as well as to the return one day ahead. With increasing the lag between sentiment measure and return, this relation becomes less significant for both sentiment measures. Since contemporaneous sentiment does not introduce any incremental forecasting ability, we are more interested in lagged sentiments. Relation between today's sentiment reversal and tomorrow's return is stronger and noticeably more significant than the relation between today's raw sentiment score and tomorrow's return.

The conclusion from the results presented in this chapter, as well from the previous works is that sentiment reversal measure can be very useful dimension in news analytics, and that any system aiming in representing the results of the sentiment extraction should not omit representing this parameter. The next chapter describes the news analytics visualization system build upon this premise.

4. VISUALIZATION IN TEXT ANALYTICS

Visualization of data extracted from great amounts of text is a task being usually solved as a part of the wider project, and often without extensive exploration of previous work in this area. All existing works cluster around two main focuses: topic detection and sentiment extraction from user opinions. For the extensive overview of existing prototypes and approaches, the reader is advised to consult prominent surveys [7, 8].

FINDS System

Based on our findings presented in the previous chapters, and on the previous works in related areas of text sentiment visualization, we developed a visualization component as a part of the project FINDS (Financial News and Data Services, see [9]). The project FINDS is being developed within Information Management and Market Engineering Graduate School of Karlsruhe Institute of Technology (KIT). The goal of the project is to offer services to professional traders and researchers, and help them in managing huge amounts of financial data produced by news agencies, marketplaces, or companies themselves.

The FINDS system itself consists of three main components: the first component is a natural language processing and text classification engine which is used to extract sentiment from text documents and news articles. By now a number of engines was built each employing different systems for text classification and machine learning (e.g. for a neural network classifier see [10]). The extracted sentiment score is delivered to the second component which is used for performance assessment. Employing the methodology of empirical finance, and aligning sentiment data with quantitative data on an extensive dataset, this benchmark can determine relative quality of different classifying engines, and can help in choosing the appropriate one for the specific task [6].

FINDS Visualization Component

The produced sentiment scores are numerical values that summarize the tone of the great amount of news articles for number of different companies. To render this numerical summarization useful, we developed FINDS visualization component as a third main component of the FINDS system.

The visualization component is loosely coupled with the rest of the system via a messaging interface. This allows it to be connected and receive inputs from various systems, either other classifier components of FINDS system, or directly from commercially available providers of sentiment scores, like Thomson Reuters.

When a news article arrives it is analyzed by one of the FINDS classifiers or Reuters News Score Sentiment Engine, and the sentiment score for each of the companies mentioned in the article is calculated as a real number between -1 and 1. The values near -1 indicate news that are rather bad for the affected company, and implicitly signalize negative pressure on company's equity price and returns. As we learned from the previous chapters, sudden change of the sentiment are even better signals of this negative pressure, so when negative news arrives for the company that had in average positive sentiment, that is a strong signal that the downward pressure on prices will be strong in short future period. This is why our visualization system is designed in the way to make such and similar cases clearly noticeable.

Picture 1: Screenshot of FINDS visualization component

The typical screen of FINDS visualization component is shown in Picture 1. The users have a possibility to set a list of all companies they are interested in, so only mention of these companies will be filtered out. Each

company is represented by the colored circle. We can vary four dimensions for each company: x axis position, y axis position, size and color.

The size of the circle representing company and its sentiment is defined in such a manner that the visibility is kept also in the case of simultaneous representation of a great number of companies. The circle area is linearly dependent on number of messages published about particular company, and the total area of all displayed representations is kept constant. Using this representation approach, the companies that draw more attention and are mentioned in more news stories are represented with larger circle. The diameter D_c of a company representation is calculated following Equation 2, as described in [11]. The value N_c represents the number of published news articles about a company, N_{tot} is the total number of news articles for all companies, and l is the length of the representation window that can be scaled by the user.

$$D_c = \frac{\sqrt{N_c}}{\sqrt{N_{tot}}}\frac{l}{C}, \qquad C = \frac{1}{4\sqrt{\pi}} \tag{2}$$

The color of the company's representation is chosen from the palette that ranges from green, over yellow and orange, to red. The color directly represents the last sentiment score, by using bright red to represent sentiment value of -1 and very bad news to bright green used for the sentiment value of +1 and very positive news. The x position is determined in the same manner, ordering companies from left to right according to increasing recent sentiment score.

The y position of the company's representation is determined by a deviation of the recent sentiment score from the average sentiment score for the particular company, so the sentiment reversals can be easily noticed. The companies with highest sudden increase in sentiment are singled out in upper portion of display, while those companies with highest decrease in sentiment are occupying the lower portion of the display. To make this relation more obvious, we use a transformation described by Equation 3 to calculate y position of the company representation y_c. With $S_r(c)$ we denote deviation of current sentiment score from the average value, as described in previous chapters. The values for constants are chosen empirically, after testing the system on real data.

$$\begin{aligned} y_c &= C_1 \arctan(C_2 S_r(c)) + C_3 \\ C_1 &= \frac{1.056}{\pi}, \quad C_2 = 6, \quad C_3 = 0.5 \end{aligned} \tag{3}$$

The user is offered the possibility to see the details of the selected company, like numerical values of current sentiment, average sentiment, the integral text of news story, and the historical development of stock prices for the selected company. Our goal is to offer an integrated decision support tool for the users interested in analyzing influence of published news articles to market movements. The Visualization component is developed within "Education in Programming Projects" course held at the Institute AIFB, with extensive help of Hagen Buchwald and Maik Landwehr.

5. CONCLUSION

In an attempt to help professional traders and researchers we apply the methods of news analytics and explored the benefits of using sentiment reversal indicator instead of raw sentiment score that is produced by most of the systems. The findings suggest that this approach can improve the performance, so we propose the visualization methodology that makes use of sentiment reversals and describe the visualization system based on this methodology implemented as part of FINDS project.

REFERENCES

[1] Steven Kung Fan Leung, *Automatic stock market predictions from World Wide Web data*, Hong Kong University of Science and Technology, 1997

[2] P. C. Tetlock, "Giving Content to Investor Sentiment: The Role of Media in the Stock Market", *The Journal of Finance*, 62(3), 2007

[3] Mittermayer, M. and Knolmayer, G., *Text mining systems for market response to news: A survey*, 2006

[4] Richard Brown, "Incorporating news into algorithmic trading strategies: Increasing the signal-to-noise ratio", *The Handbook of News Analytics in Finance*, Wiley Finance, 2011

[5] John Kittrell, "Sentiment Reversals as Buy Signals", *The Handbook of News Analytics in Finance*, Wiley Finance, 2011

[6] Caslav Bozic, Ryan Riordan, Detlef Seese, and Christof Weinhardt, "Towards a Benchmarking Framework for Financial Text Mining", *Studies on eOrganisation and Market Engineering*, KIT Scientific Publishing, Karlsruhe, 2010

[7] Artur Silic and Bojana Basic, "Visualization of Text Streams: A Survey", *Knowledge-Based and Intelligent Information and Engineering Systems*, Springer Berlin / Heidelberg, 2010

[8] Bo Pang and Lillian Lee, "Opinion Mining and Sentiment Analysis", *Found. Trends Inf. Retr.*, 2008

[9] Caslav Bozic, "FINDS - Integrative services", *IEEE/ACS International Conference on Computer Systems and Applications*, 2009

[10] Caslav Bozic and Detlef Seese, "Neural Networks for Sentiment Detection in Financial Text", *14th International Business Research Conference*. World Business Institute Australia, 2011

[11] Amira Darmoul, Gregor Hackensellner, and Philipp Rouast, "FINDSVisions", KIT, Karlsruhe, 2011