Impacts of Public News on Stock Market Prices: Evidence from S&P500[1]

Mihály Ormos
Department of Finance, Budapest University of Technology and Economics
Magyar tudósok krt. 2., 1117 Budapest, Hungary
ormos@finance.bme.hu

Miklós Vázsonyi
Department of Finance, Budapest University of Technology and Economics
Magyar tudósok krt. 2., 1117 Budapest, Hungary
miklos@vazsonyi.com

ABSTRACT

Searching for semantical connections between nouns and adjectives within sentences proves to be useful in investigating the potential impacts of publicly available news on daily returns. We use extraction of composite expressions of adjective-noun pairs from economic and financial news based on a sentence based text processor. Analysis of 10 year period using seven news journals and daily closing stock prices results in approximately 70% precision and 50% recall indicators on S&P500 stocks related to trading days when extreme positive or nagative returns were realized by the majority of stocks. Indicators of general economic news are better then firm specific headline news. News on days T+1 corresponds with returns better than news on day T or T-1. Our method outperforms the analysed bag of words alternatives.

Keywords: Market efficiency; News wire; Semantic analysis; Adjective-noun expressions

1. INTRODCTION

We investigate the information reflection capability of stock prices. The paper summarises a quantitative analysis of daily stock market for S&P500 companies considering the potential impacts of publicly available news announcements on daily returns. We semantically analyse news on all S&P500 companies in a 10 years time period to measure the potential impact of news announcements on daily stock market prices. Days with extreme returns are identified for both positive and negative events in terms of extreme positive or negative impact on the stock market. News announcements are collected from LexisNexis database from seven magazines and journals for those days having extreme absolute return values (positive or negative) during the trading session. We report 70% precision and 50% recall indicators for the selected news with the applied sentence based semantical classifier. We use sentence based text processing and adjective-noun tokenization with expression recognition to classify news into bad and good classes. It turns out that on extreme days the classified news predicts the extreme returns consistently (however, causality is not proven yet). It also turns out that general economic and financial news announcements have better predictive capabilities than company specific headline news (where the name of the company appears in the headline).
Compared to simple bag of word approach, the adjective-noun based expression recognition prooves to be better in indicating extreme returns meaning that it can be useful to look for semantical relationships between words in

[1]We acknowledge Prof. Hercules Vladimirou form University of Cyprus, HERMES European Center of Excellence on Computational Finance and Economics the news wire and financial data and thanks the financial support of the National Eötvös Scholarship Fund, Hungary.

sentences. However, there are two limitations to be considered that motivate further research: (1) the number of days investigated is relatively few to draw a strong conclusion; (2) we analyse only end-of-day closing prices and news with only daily resolution instead of intraday prices and time stamped news so we cannot be sure that causality exists in terms of the news announcement and price change relationship.

2. RIVIEW OF LITERATURE

Early studies of market efficiency (Fama 1965, 1970 and 1991) suggests that at any given time, security prices fully reflect all available information. Thus, no investor has an advantage in predicting a return on a stock price because no one has access to information not already available to everyone else, therefore only unanticipated information can push the prices. By the definition of Fama (1991) the efficiency of a market can be classified by the type of information reflected by the prices. Weak form efficiency states that all past prices, returns and other technical information of a stock are reflected in securities price (i.e. technical analysis cannot be used to predict and beat a market). Semi-strong form efficiency claims that all public information is reflected in securities' current price (i.e. neither fundamental nor technical analysis can be adventegous to gain abnormal return. The efficient market hypothesis does not argue, that abnormal return gains when new information comes, but argues that these informations cann be anticipated, therefore no one can earn higher return than the normal. Financial market prediction methods generally use quantitative values to provide point or interval estimations on returns, volatility or trade volume. Historized timeseries of financial products with previous market prices are commonly used in portfolio optimization. However, quantitative analysis of qualitative economic news given in unstructured textual formats started a couple of years ago when Internet news media became the dominant information source of investors. Recent studies identify systematic relationships between trading volume and measures of communication activity (Tetlock, 2007).

Most of the reseach show that there is a positive correlation between news cardinality and trade volume plus volatility. Liang (2006) defined the web stock news volumes (WSNV) indicator and measured the impact of it on financial market behavior. He classified news according to direct and indirect aspects: direct news comes from a dedicated site containing the current news of a certain company; indirect news: company is mentioned on an other news site. His main conclusion is that significant increases of web stock news volumes are linked with the significant changes of stock prices. These findings are in line with the efficient market hypothesis (EMH). On the other hand the research focuses on trading profiles with wich abnormal returns can be gained using publicly available information. Tetlock (2007) reports significant correlation between written financial media content and aggregate financial market performance as an evidence that news content can predict movements in stock market activity. Principal component analysis is used to create a simple measure of media pessimism from Wall Street Journal news of 16 years then the intertemporal links between this measure of media pessimism and the stock market is estimated using basic vector autoregressions (VARs). 77 predefined word categories of Harvard Psychosocial Dictionary is used to find the category with the highest variance which is called the pessimism factor. This method follows a bag-of-words approach to handle textual documents. One of the conclusions is that high levels of media pessimism robustly predict downward pressure on market prices on the next trading day, followed by a reversion to fundamental value within one week. Unusually high or low values of media pessimism forecasts high market trading volume.

Tetlock, Saar-Tsechansky and Macskassy (2008) describe a news-based automated trading strategy based on relative occurrence of negative words in firm specific financial news in an effort to predict firms' accounting earnings and stock returns. A simplified bag of words representation was used to interpret textual data according to the relative frequency of negative words defined by the Harvard psychosocial dictionary. Key findings of Tetlock et al. (2008) are that the fraction of negative words in firm-specific news forecasts low firm earnings and stock market prices respond to the information embedded in negative words with a small, one-day delay. They find that firms' stock prices briefly underreact to the information embedded in negative words. Surprise analysis also focuses on news announcement and market impact measurement. Zeynep Önder and Can Simga-Mugan (2006) analyse the impact of political and macroeconomic news in emerging markets to investigate the origin of high returns. They examine the effects of macroeconomic and political news items on the volatility of returns and total trading volume between 1995 and 1997. The main difference of Örder et al (2006) works from the previously descibed ones that they concentrate on systemtic information istead of firm specific ones. They conclude that political and world economic news increase the volatility emerging markets prices and there is a significant positive correlation between world economic news items and volume, and a significant positive relationship can be measured between domestic and world economic news and market volume. Several studies recognize that political information significantly affects the stock market prices (e.g., Gartner and Wellershoff, 1995; Hensel and Ziemba, 1995; Herbst and Slinkman, 1984; Huang 1995; Lobo 1999; Riley and Luksetich 1980). Most of these studies examine the effect of presidential and midterm elections, and the result of elections, on returns in U.S. markets, finding noticeable relations.

Koppel and Shtrimberg (2006) argue that models based on lexical features can distinguish good news from bad news with accuracy of about 70%. They introduce a simple method for generating labeled examples for sentiment analysis: news stories about publicly traded companies are labeled positive or negative according to price changes of the company stock. The show that there are many lexical markers for bad news but none for good news. The novel idea is the automatic unsupervised clusterization of large amount of news. The use of price movements correlated with the appearance of news items is a promising method for automatically generating a labeled corpus without directly invoking individual human judgments (though, stock movements themselves are a product of collective human judgment). In their work, there are no assumptions by making judgments regarding a story itself, only the reaction of the market to the story is important. Goonatilake and Herath (2007) focus on the effect of news that surfaces throughout the day in the stock market. News stories about publicly traded companies were labeled positive, negative or neutral according to price changes in the company stock. Takahashi, Takahashi, Takahashi and Tsuda (2007) use Naive Bayes classifier for text labeling.They measure stock price change before and after publishing news. They performe a morphological analysis and pattern matching to extract keywords then good, bad and neutral news clusters are created. The text categorization system perfom about 80% of classification accuracy. Liang (2006) use web stock news volumns analysis along with good versus bad classification of news item. His main finding is that in bear market, bad information influences the market more severely than good information. Namely, the asymmetry phenomenon should be addressed separately for the bear market.

Niederhoffer (1971) explore the effect of headline news appearing in the New York Times and the Los Angeles Times on the stock market from 1950 to 1966. He group headline news into various classes, and as good or bad. There are also a couple of commercially available software applications and services focusing on news processing with text analysis methods. Reuters NewsScope Sentiment Engine processes a stream of Reuters news items, producing sentiment data for a list of customer determined target companies[2]. MediaSentiment Pro includes the advanced search engine that generates Buy/Sell signals based on sentiment analysis of earnings releases[3]. YellowBrix Sentiment Analytics provides an aggregated view of media sentiment and news analytics from global newspapers, news wires, trade journals and business blogs, providing information for investors[4]. The Dow Jones Economic Sentiment Indicator aims to predict the health of the U.S. economy by analyzing the coverage of 15 major daily newspapers in the U.S. It uses a numerical scale from 0 to 100 to express the balance of sentiment in articles about the economy[5]. RightTrade gives investors insight into the media trends. The service provides an interactive and visual representation of real-time media sentiment including the leading global newspapers, business blogs, trade journals, and news wires – across indices, companies, multiple tickers, and customized watch lists[6]. The primary motivation of our research is that trhough the review of recent literature we cannot find any which applies adjective and noun based sentence tokenization in text processing with expression recognition. We compile an empirical research to investigate whether the idea to classify the intraday news based on adjective-noun expressions is able to predict daily price changes or at least there is significant relationship between these two variables.

INPUT DATA

We use three types of input data during the investigated period: (1) company specific headline news from Associated Press financial wire between 17 December 2005 and 1 July 2009 (2) general economic and financial news from seven popular daily journals between 26 October 1997 and 9 April 2004 and (3) daily closing stock prices for S&P500 companies between 30 June 1989 and 1 July 2009. We use daily closing price data for all S&P500 companies resulting in a time series which conatins 5219 trading days and multiple news for those trading days having extreme positive or negative returns. Our investigation is built onhistorical daily closing stock prices

[2] http://www.bobsguide.com/guide/prod/5-12205_Reuters_NewsScope_Sentiment_Engine.html
[3]http://www.mediasentiment.com/demo/
[4]http://www.yellowbrix.com/index.nsp?sid=bp&pid=3
[5]http://solutions.dowjones.com/economicsentimentindicator/
[6]http://www.righttrade.com/RT/spi/demo

time series from Thomson's DataStream system. Price data was converted into return values for every S&P500 stocks based on the following formula:

$$r_t = \frac{p_t - p_{t-1}}{p_{t-1}}$$

where r_t indicates the return percentage in time t, and p_t and p_{t-1} stand for the prices in the t and $t-1$ respectively. We collect the general economic and financial news items from seven news journals from the LexisNexis database. The following magazines and journals are used to download filtered economic and financial news from (excluding news on sports, culture, entertainment and other non-business topics):

Table 1: News sources

Journal	Number of news
Daily News (New York)	4.646
International Herald Tribune	1.797
The New York Times	13.604
USA TODAY	2.839
WALL STREET JOURNAL	7.307
The Washington Post	13.013
The Washington Times	4.004
Total	**47.210**

Number of initial economic and financial news from seven popular daily journals.

In order to compare general economic and financial news with company specific headline news (where the name of the company apperars in the news headline) we use headline news as well from Associated Press Financial Wire having overall a 90,495 headline news.

3. METHODOLOGY

In the first step we define the days of interest (DOI), which are the days where extreme price movements, extremly high positive (DOI+) or negative (DOI-) returns occur for the majority of stocks (the exact specification of DOI is given later). At the same time the days of uninterest (DOUI) are identified, these are the days where nothing interesting is happened (the majority of the stocks performed below a certain level of return in absoulte terms). As a second step we collect the news announcements from the above mentioned seven journals (and from the AP news wire in an other scenario) for the days of interest and uninterest. After collecting all the textual data we apply a semantic analysis on the news using sentence based tokenization and noun-adjective expression recognition. Based on the adjective-noun semantical analysis we classify the news into good and bad groups according to their noun-adjective expression characteristics. We do not use all of the news because the lack of predefined adjective-noun expressions. We ignore articles that do not contain adjective-noun expression in sentences, meaning that these news can not help us in the prediction procedure. We perform precision and recall calculations[7] on the successfully classified news to analyse the return predictive capabilities on a daily level.

3.1. Days of interest and uninterest

We define days of interest (DOI) as the trading days during the investigated 20 year period when extreme positive or negative returns are gained by the majority of the S&P500 stocks. It means that we considerextreme days when the majority of the stocks moved to the same direction with a return above market expectations, which are defined from an investor's perspective who would like to realise a certain percentage of net gain in a predifined period (short selling is allowed, i.e. the extreme negative price changes included).

[7]See information retrieval studies for further details, e.g. Buckland and Gey,1994

3.1.1. Days of interest

We are searching the days when the majority of the S&P500 daily stock market is affected and all the affected stocks have an extreme positive or negative return. We handle extreme positive and extreme negative returns separated in order to be able to investigate the potential reasons behind the different events separately. The Table 2 and Table 3 illustrate the number of DOIs in the case of extreme negative and extreme positive returns respectively. A cell in the table shows the number of days when X percent of the stocks produced at least Y percent negative return; where X represents the percent value on the horizontal axis (row) and Y represents the percent value on the vertical axis (column header). Similarly, we identify positive days summarized in Table 3. A cell in the table shows the number of days when the X percent of the stocks produced at least Y percent positive return where X represents the percent value on the horizontal axis (row) and Y represents the percent value on the vertical axis (column header).

3.1.2. Days of uninterest

We analyze those days having low average market activity and low returns and considered these days as days of uninterest. On these days the majority of the S&P500 daily stock market had less than 2% returns. We identify these days to avoid bias in our final conclusions so we analyse the news on these days on the same way to detect false predictions as well. Table 4 illustrates the number of"days of uninterest in different configurations.

3.2. News cardinality analysis in different configurations

We collect publicly available news items from LexisNexis database for those days when the majority of the S&P500 daily stock market was affected with extreme positive or extreme negative returns. News items are coming from the previously defined major US newspapers without industry filtering. In order to compare the news cardinalities with other days with minimal activity in terms of extreme returns and to obtain training news items for days of uninterest as well we analyzed the news cardinalities on days with low returns across the S&P500 daily stock market. We also gathered news items for the previous days when major market events happened. Table 5 and Table 6 illustrate the news cardinalities on the "days of interest" and the cardinality of the subset of these news that have adjective-noun based expressions respectively.

3.3. Text processing

We use simple text processing methods to make the news ready to analyse with sentence based expression recognition algorithms: (1) basic stemming to reduce the different forms of the same words to their common stem; (2) sentence tokenization to handle sentences independently; (3) adjective-noun pair based expression recognition to extract basic semantics from sentences.

3.3.1. Basic stemming

Stemming of words were performed with a simple stemmer algorithm consisting of the following replacement operations. Since we focused only on adjectives and nouns the proper stemming of the rest of the words was irrelevant to us.

3.3.2. Sentence tokenization

We handle sentences of news independently. Our simple approach recognises sentence terminator characters (e.g. period, question mark) and separates the sentences from each other. This approach is limited in recognizing semantic connections between succeeding sentences. However, it proves to be enough to identify basic semantics inside sentences.

3.3.3. Adjective-noun expression recognition

We define four categories of words according to the following list:positive nouns;negative nouns;positive adjectives;negative adjectives. The motivation to define these categories of words is that adjective-noun expression can accurately describe the news article's author's messages. There are four possible composite expressions consisting of one adjective and one noun as illustrated below.

4. Results - precision and recall calculations

There are two commonly used indicators in information retrieval to characterize textual documents: precision and recall. We used these two indicators to measure the predictive capabilities of news on the „days of interest". The formal definitions of the precision (Pr) and recall (Re) indicators are the following.

$$\mathrm{Pr} = \frac{TP}{TP + FP}; \quad \mathrm{Re} = \frac{TP}{TP + FN}$$

where TP, FP, FN and TN can be defined according to Table 11. Precision expresses the percentage of positively classified news on a given day having positive returns compared to the total number of classified news on that given day. Recall shows the percentage of positively classified news on a given day having positive returns compared to the number of all positively classified news on that particular day. We compare the precision and recall indicators for general economic news and firm specific news as well. We note here that the journals we use in the analysis can contain previously edited articles published on a certain day, which means that the content of some news may refer to an event in the past. However, news published on a certain date can affect investor sentiment and may have an impact on the investments on that day according to our hypothesis.We run the analyis on different absolute return and market impact values. We repeat this parametrized analysis for days T-1 and T+1 as well in order to investigate the potential impacts of news before and after a major price change on the S&P500 stock market.

In order to show the corresponding news in a particular research scenario we indicate the number of days of interest and uninterest and we also include the number of days of uninterest when the S&P500 stocks were calm in a sense that basically none of them realized extreme absolute returns.Our integrated analysis also contains anovel adjective-noun based expression recognition classifying and the bag of words method as well. In the bag of words approach we do not use sentence tokenization and adjective-noun expressions. Instead we count only the positive and negative adjectives to classify the news into bad or good classes. The results can be found in Table 12, Table 13 and Table 14 covering different parameter ranges for days T, T-1 and T+1. Absolute return parameter varies from 0,1% to 1,2% while market impact parameter varies from 45% to 65%. After calculating the precision and recall indicators for T day news (news published on the same day when we analyse the returns, we repeat the same calculations for T-1 and T+1 news (news published on the previous and the next day of the return day respectively).

We find that our adjective-noun based expression classifier outperforms the results of the simple bag of words algorithm in the majority of research scenarios. We state that the reason behind is that semantic analysis of sentences can detect the real intent of the news content more accuraately than the pure bag of words method. Comparing general economic news with firm specific ones shows in our case (having a pool of seven journals only) that general news announcements have higher precision and recall indicator values than firm specific news.Analysis of days T, T-1 and T+1 show that the precision and recall indicators of news have the highest values for day T+1. This finding can show that the stock market price changes preceed the corresponding news announcements. However, we also have to note here that this statement refers to the seven journals involved in our analysis. Tables 12, 13 and 14 show that the higher the market impact, the fewer of days of interest we find. Increasing the absolute return parameter in a given market impact scenario has the same effect: the higher the absolute return value, the fewer of days of interest we can observe.The most important output is the promising signs of semantically analysis of news in the research after casualty between news announcements and large price changes.

CONCLUDING REMARKS

Our main contribution is that we show an adjective-noun based sentence level expression recognition method to semantically classify news into bad and good categories based on its content. On top of this semantic news classifier we report an average of 70% predictive capability for days of interest when there are large price changes on small scales for tha majority of the S&P500 stocks. We compare our adjective-noun expression approach with the classical bag of word approach with the same adjectives and find that the novel method outperforms the bag of word approach. Also, we recognize that T+1 news have better predictive power in general weakening the hypothesis of causality but motivating further research to exactly prove the hypothesis or exclude its possibility.

The limitation of the conclusion is that causality cannot be shown because of the resolution of the data.Our longer term motivation is to improve the resolution of our analysis by acquiring intraday price data and timestamped news. This research isa preparation work for further analysis where we plan further improvements in order to make our claims stronger by stock price resolution improvements (intraday stock price data is required to be able to perform accurate event analysis capable of indicating causality); news resolution improvements (intraday timestamped news items are required to have the accurate time of the news announcement); semantical text processing improvement (improving the morphological and semantical capabilities of our text mining algorithm).

REFERENCES

1. Buckland, Michael K. and Gey, Fredric. 1994, The Relationship Between Recall and Precision. Journal of the American Society for Information Science, Vol. 45., No. 1., pp. 12-19.
2. Fama, Eugene F, 1965, The Behavior of Stock-Market Prices. The Journal of Business, Vol. 38., No. 1., pp. 34-105.
3. Fama, Eugene F, 1970, Efficient Capital Markets: A Review of Theory and Empirical Work, Journal of Finance, American Finance Association, Vol. 25., No. 2., pp.383-417
4. Fama, Eugene F, 1991, Efficient Capital Markets: II. Journal of Finance, Vol. 46., No. 5., pp. 1575-1617
5. Gartner, M., and K.W. Wellershoff. 1995, Is There an Election Cycle in American Stock Returns?" International Review of Economics and Finance, Vol. 4, No. 4., pp 387–410.
6. Hensel, C.R., and W.T. Ziemba. 1995, United States Investment Returns During Democratic and Republican Administrations, 1928–1993. Financial Analysts Journal, Vol. 51., No. 2., pp. 61–69.
7. Herbst, A.F., and C.W. Slinkman. 1984, Political–Economic Cycles in the U.S. Stock Market. Financial Analysts Journal, Vol. 40., No. 2., pp. 38–44.
8. Huang, R.D. 1995, Common Stock Returns and Presidential Elections. Financial Analysts Journal,Vol. 41., No. 2., pp. 58–65.
9. Koppel, Moshe; Shtrimberg, Itai, 2006, Good News or Bad News? Let the Market Decide. The Information Retrieval Series, Vol 20, pp. 297-301
10. Lobo, B.J. 1999, Jump Risk in the U.S. Stock Market: Evidence Using Political Information. Review of Financial Economics, Vol. 8.,No. 2., pp 149–163.
11. Önder, Zeynep; Şimga-Muoğan, Can, 2006, How Do Political and Economic News Affect Emerging Markets? Evidence from Argentina and Turkey. Emerging Markets Finance & Trade, Vol. 42., Issue 4., pp. 50-77
12. Riley, W.B., and W.A. Luksetich. 1980, The Market Prefers Republicans: Myth or Reality. Journal of Financial and Quantitative Analysis, Vol. 15., No. 3., pp. 541–559.
13. Rohitha Goonatilake; Susantha Herath, 2007, The Volatility of the Stock Market and News. International Research Journal of Finance and Economics, Issue 11., pp. 53-65.
14. Satoru Takahashi; Masakazu Takahashi; Hiroshi Takahashi; Kazuhiko Tsuda , 2007, Analysis of the Relation Between Stock Price Returns and Headline News Using Text Categorization. Lecture Notes in Computer Science, Vol. 4693., pp. 1339-1345.
15. Tetlock, Paul C., 2007, Giving Content to Investor Sentiment: The Role of Media in the Stock Market. Journal of Finance, Vol. 62.,No.3., pp. 1139-1168.
16. Tetlock, Paul C.; Saar-Tsechansky, Maytal; Macskassy, Sofus, 2008, More Than Words: Quantifying Language to Measure Firms' Fundamentals. Journal of Finance, Vol. 63.,No.3., pp. 1437-1467.
17. X. Liang., 2006, Mining Associations Between Web Stock News Volumes and Stock Prices, International Journal of Systems Science, Vol. 37., pp. 919-930.

Table 2: Days of Interest with positive returns

	Market impact (%)									
Absolute return (%)	**50**	**55**	**60**	**65**	**70**	**75**	**80**	**85**	**90**	**95**
0,1	727	508	353	233	155	84	36	9	0	0
0,2	624	415	297	183	127	62	27	5	0	0
0,3	521	353	244	155	94	52	18	2	0	0
0,4	437	297	200	134	72	37	11	0	0	0
0,5	369	246	166	104	59	29	8	0	0	0
0,6	308	210	130	88	43	22	4	0	0	0
0,7	256	173	112	63	32	19	3	0	0	0
0,8	212	143	86	52	28	14	3	0	0	0
0,9	181	117	68	40	22	7	3	0	0	0
1	156	93	53	35	20	4	3	0	0	0
1,1	126	74	48	27	14	4	3	0	0	0
1,2	100	61	39	22	13	3	2	0	0	0

Number of DOIs on the S&P500 daily stock market regarding extreme negative returns with high market impact. Columns indicate the daily extreme negative return values, rows indicate the market impact (100% means that all of the S&P500 stocks were affected at the same day). The element in the ith row and jth column indicates the number of days between 30 June 1989 and 1 July 2009 when the S&P500 market was affected by at least the ith market impact percentage and the affected stocks had at least the jth negative return. Five configurations were analyzed manually during the study, these configurations are indicated with thick rectangles.

Table 3: Days of Interest with negative returns

	Market impact (%)									
Absolute return (%)	**50**	**55**	**60**	**65**	**70**	**75**	**80**	**85**	**90**	**95**
0,1	805	580	411	289	190	110	43	7	0	0
0,2	719	508	353	246	161	89	30	5	0	0
0,3	623	433	300	207	123	65	19	5	0	0
0,4	531	376	260	165	100	50	14	4	0	0
0,5	458	311	217	130	77	35	11	3	0	0
0,6	393	272	176	108	59	22	6	1	0	0
0,7	323	225	150	95	45	19	5	1	0	0
0,8	275	187	116	68	35	16	3	1	0	0
0,9	238	155	95	56	28	11	2	0	0	0
1	198	132	72	42	22	10	1	0	0	0
1,1	170	99	63	37	14	8	1	0	0	0
1,2	138	81	51	25	12	6	1	0	0	0

Number of "days of interest" on the S&P500 daily stock market regarding extreme positive returns with high market impact. Columns indicate the daily extreme positive return values, rows indicate the market impact (100% means that all of the S&P500 stocks were affected at the same day). The element in the ith row and jth column indicates the number of days between 30 June 1989 and 1 July 2009 when the S&P500 market was affected by at least the ith market impact percentage and the affected stocks had at least the jth positive return. Five configurations were analyzed manually during the study, these configurations are indicated with thick rectangles.

Table 4: Days of uninterest with positive returns

	Market impact (%)									
Absolute return (%)	**50**	**55**	**60**	**65**	**70**	**75**	**80**	**85**	**90**	**95**
0,1	370	209	169	152	145	145	145	145	145	145
0,2	414	222	171	152	145	145	145	145	145	145
0,3	581	297	185	158	147	145	145	145	145	145
0,4	915	470	248	173	152	146	145	145	145	145
0,5	1387	759	372	212	159	148	145	145	145	145
0,6	1851	1197	575	294	184	151	145	145	145	145
0,7	2229	1674	936	441	227	163	146	145	145	145
0,8	2564	2070	1397	664	316	190	151	145	145	145
0,9	2854	2432	1842	1063	459	235	162	145	145	145
1	3085	2722	2245	1544	735	319	193	150	145	145
1,1	3272	2940	2549	1975	1160	473	236	162	145	145
1,2	3464	3135	2806	2312	1576	727	289	177	146	145

Number of "days of uninterest" on the S&P500 daily stock market regarding low returns of the majority of the stock. Columns indicate the daily low return values, rows indicate the market impact (100% means that all of the S&P500 stocks were affected at the same day). The element in the ith row and jth column indicates the number of days between 30 June 1989 and 1 July 2009 when the S&P500 market was affected by at least the ith market impact percentage and the affected stocks had at most the jth low return. One configuration were analyzed manually during the study, this configuration is indicated with thick rectangle.

Table 5: Number of news on days of interest with high positive and negative returns

Date (YYYY.MM.DD)	Daily News (New York)	International Herald Tribune	The New York Times	USA TODAY	Wall Street Journal	The Washington Post	The Washington Times
DOI+							
2002.07.05	28	20	122	50	138	94	39
2002.07.24	74	16	162	56	164	132	40
2002.07.29	52	18	86	44	108	90	34
2002.10.11	49	16	165	47	137	172	49
2002.10.15	58	19	133	38	119	106	43
2003.01.02	31	14	101	25	107	235	21
2003.03.17	45	49	88	43	99	107	32
DOI-							
1997.10.27	47	12	171	49	138	92	63
1998.08.27	56	17	138	43	139	172	48
1998.08.31	43	10	81	56	136	81	61
2000.04.14	78	19	137	62	166	106	77
2002.07.10	73	18	164	46	122	117	45
2002.07.19	59	17	145	50	158	142	60
2002.08.05	46	8	61	38	104	92	30
2002.09.03	45	26	110	44	123	96	31
2003.01.24	64	67	182	27	120	146	54
2003.03.10	61	57	98	36	96	101	31
2003.03.24	56	70	105	51	114	124	31

Cardinality of news items from seven newspapers (Daily News, International Herald Tribune, The New York Times, USA Today, Wall Street Journal Abstracts, The Washington Post, The Washington Times) on "days of interest" in a certain market impact vs extreme return configuration (where more than 75% of S&P500 stocks showed not less than 0.9% return). The number of analyzed days is 18.

Table 6: Number of news having adjective-noun based expressions on days of interest

Date (YYYYMMDD)	The New York Times	USA TODAY	Wall Street Journal	The Washington Post	The Washington Times	International Herald Tribune	Daily News (New York)
DOI+							
2002.07.05	4	3	9	2	0	0	4
2002.07.24	6	2	3	38	4	2	6
2002.07.29	1	3	2	13	3	0	1
2002.10.11	7	4	7	3	1	0	7
2002.10.15	5	1	12	6	0	0	5
2003.01.02	4	1	4	26	0	0	4
2003.03.17	1	3	11	3	2	0	1
DOI-							
1997.10.27	4	1	19	3	4	0	4
1998.08.27	7	18	14	0	0	0	7
1998.08.31	1	1	5	14	3	1	1
2000.04.14	2	8	3	14	3	0	2
2002.07.10	10	3	12	6	2	0	10
2002.07.19	1	7	3	24	3	0	1
2002.08.05	1	2	1	11	2	1	1
2002.09.03	3	1	2	12	1	1	3
2003.01.24	4	3	1	19	3	0	4
2003.03.10	2	2	1	8	3	1	2
2003.03.24	3	3	8	5	2	0	3

Cardinality of news items from seven newspapers (Daily News, International Herald Tribune, The New York Times, USA Today, Wall Street Journal Abstracts, The Washington Post, The Washington Times) on "days of interest" in a certain market impact vs extreme return configuration (where more than 75% of S&P500 stocks showed not less than 0.9% return) where adjective-noun expressions can be found in the given news. The number of analyzed days is 18.

Table 7: Simple stemming in text processing

Removed tailing characters	Example
s	earns
est	strongest
ly	strongly
ing	earning
er	lower
ened	strenghtened

Basic stemming steps used in our analysis to convert the different forms of the same words into a common form.

Table 8: Possible combinations of adjective-noun pairs

		Noun	
		Positive	**Negative**
Adjective	**Positive**	e.g. high profit	e.g. high losses
	Negative	e.g. low profit	e.g. minimal losses

The four possible combinations of positive and negative andjectives and nouns. For proper sentence recognition examples see Appendix 2, for mistaken sentence recognition examples see Appendix 3.

Table 9: Number of relevant articles in general news

Journal	Number of news with relevant expressions
WALL STREET JOURNAL	819
The Washington Post	656
The New York Times	534
Daily News (New York)	139
The Washington Times	119
USA TODAY	116
International Herald Tribune	6
Total	**2389**

The four possible combinations of positive and negative andjectives and nouns. For proper sentence recognition examples see Appendix 2, for mistaken sentence recognition examples see Appendix 3.

Table 10: Number of relevant articles in company specific headline news

Journal	Number of news with relevant expressions
Associated Press Financial Wire	3433
Total	**3433**

The four possible combinations of positive and negative andjectives and nouns. For proper sentence recognition examples see Appendix 2, for mistaken sentence recognition examples (as a motivation for further research) see Appendix 3.

Table 11: Number of relevant articles in company specific headline news

	Positively classified news	Negatively classified news
Extreme positive return	TP (true positive)	FP (false positive)
Extreme negative return	FN (false negative)	TN (true negative)

The four possible combinations of events regarding stock market activity and news classification on the same day.

Table 12: Precision and Recal calculations on day T

					General financial and economic news				Company specific headline news			
Absolute return	Market impact	DOI+	DOI-	DOUI	Precision (AN)	Recall (AN)	Precision (BOW)	Recall (BOW)	Precision (AN)	Recall (AN)	Precision (BOW)	Recall (BOW)
0,1	45	1034	1103	739	0,80	0,55	0,81	0,53	0,78	0,51	0,59	0,49
0,2	45	906	1000	855	0,83	0,54	0,79	0,51	0,78	0,51	0,59	0,50
0,3	45	780	861	1119	0,84	0,55	0,78	0,52	0,78	0,51	0,59	0,50
0,4	45	641	750	1559	0,86	0,56	0,80	0,49	0,78	0,51	0,59	0,50
0,5	45	535	647	1956	0,87	0,52	0,82	0,43	0,78	0,51	0,59	0,50
0,6	45	458	547	2313	0,89	0,52	0,80	0,40	0,78	0,51	0,59	0,50
0,7	45	397	474	2673	0,88	0,50	0,76	0,36	0,78	0,52	0,59	0,51
0,8	45	328	419	2965	0,88	0,47	0,73	0,31	0,78	0,52	0,59	0,51
0,9	45	276	354	3208	0,85	0,46	0,64	0,25	0,77	0,51	0,59	0,50
1	45	231	294	3429	0,78	0,35	0,57	0,17	0,77	0,53	0,59	0,52
1,1	45	196	250	3614	1,00	0,37	0,60	0,16	0,77	0,53	0,59	0,52
1,2	45	170	215	3770	1,00	0,31	0,60	0,16	0,77	0,50	0,60	0,50
0,1	50	727	805	370	0,83	0,54	0,79	0,51	0,78	0,51	0,59	0,50
0,2	50	624	719	414	0,84	0,55	0,78	0,53	0,78	0,51	0,59	0,50
0,3	50	521	623	581	0,85	0,52	0,78	0,47	0,78	0,51	0,59	0,50
0,4	50	437	531	915	0,86	0,49	0,79	0,40	0,78	0,51	0,59	0,50
0,5	50	369	458	1387	0,88	0,53	0,80	0,41	0,77	0,51	0,59	0,50
0,6	50	308	393	1851	0,91	0,48	0,79	0,34	0,77	0,53	0,59	0,52
0,7	50	256	323	2229	0,88	0,50	0,73	0,31	0,77	0,53	0,59	0,52
0,8	50	212	275	2564	0,78	0,35	0,57	0,17	0,78	0,53	0,60	0,52
0,9	50	181	238	2854	0,88	0,35	0,67	0,18	0,78	0,53	0,60	0,52
1	50	156	198	3085	1,00	0,37	0,60	0,16	0,77	0,50	0,60	0,50
1,1	50	126	170	3272	1,00	0,27	0,50	0,11	0,77	0,50	0,60	0,50
1,2	50	100	138	3464	1,00	0,25	-	-	0,77	0,52	0,60	0,51
0,1	55	508	580	209	0,84	0,54	0,77	0,52	0,78	0,51	0,59	0,50
0,2	55	415	508	222	0,86	0,56	0,80	0,51	0,78	0,52	0,60	0,52
0,3	55	353	433	297	0,86	0,53	0,83	0,45	0,78	0,52	0,60	0,52
0,4	55	297	376	470	0,91	0,49	0,80	0,36	0,78	0,52	0,60	0,52
0,5	55	246	311	759	0,90	0,48	0,79	0,35	0,78	0,52	0,60	0,52
0,6	55	210	272	1197	0,85	0,46	0,64	0,24	0,78	0,52	0,60	0,52
0,7	55	173	225	1674	0,78	0,35	0,57	0,18	0,78	0,53	0,60	0,52
0,8	55	143	187	2070	1,00	0,37	0,60	0,16	0,77	0,52	0,60	0,51
0,9	55	117	155	2432	1,00	0,35	0,50	0,11	0,77	0,52	0,60	0,51
1	55	93	132	2722	1,00	0,15	-	-	0,77	0,52	0,60	0,51
1,1	55	74	99	2940	-	-	-	-	0,78	0,50	0,60	0,50
1,2	55	61	81	3135	-	-	-	-	0,78	0,50	0,60	0,50
0,1	60	353	411	169	0,85	0,52	0,84	0,48	0,78	0,52	0,60	0,52
0,2	60	297	353	171	0,86	0,57	0,83	0,46	0,78	0,52	0,60	0,52
0,3	60	244	300	185	0,89	0,44	0,76	0,31	0,78	0,52	0,60	0,52
0,4	60	200	260	248	0,89	0,50	0,75	0,35	0,78	0,53	0,60	0,52
0,5	60	166	217	372	0,85	0,46	0,64	0,26	0,77	0,52	0,60	0,51
0,6	60	130	176	575	0,88	0,37	0,50	0,16	0,78	0,50	0,60	0,50
0,7	60	112	150	936	1,00	0,39	0,60	0,16	0,78	0,50	0,60	0,50
0,8	60	86	116	1397	1,00	0,27	0,50	0,11	0,78	0,49	0,60	0,48
0,9	60	68	95	1842	-	-	-	-	0,78	0,49	0,60	0,48
1	60	53	72	2245	-	-	-	-	0,78	0,49	0,60	0,48
1,1	60	48	63	2549	-	-	-	-	0,78	0,50	0,60	0,49
1,2	60	39	51	2806	-	-	-	-	0,78	0,51	0,60	0,50
0,1	65	233	289	152	0,82	0,55	0,79	0,44	0,78	0,52	0,60	0,52
0,2	65	183	246	152	0,90	0,49	0,79	0,36	0,78	0,50	0,60	0,50
0,3	65	155	207	158	0,88	0,50	0,73	0,35	0,78	0,50	0,60	0,50
0,4	65	134	165	173	0,92	0,48	0,60	0,26	0,78	0,49	0,60	0,48
0,5	65	104	130	212	1,00	0,39	0,60	0,16	0,78	0,49	0,60	0,48
0,6	65	88	108	294	1,00	0,39	0,60	0,16	0,78	0,49	0,60	0,48
0,7	65	63	95	441	-	-	-	-	0,78	0,49	0,60	0,48
0,8	65	52	68	664	-	-	-	-	0,78	0,50	0,60	0,49
0,9	65	40	56	1063	-	-	-	-	0,78	0,49	0,59	0,48
1	65	35	42	1544	-	-	-	-	0,78	0,52	0,59	0,51
1,1	65	27	37	1975	-	-	-	-	0,78	0,52	0,59	0,51
1,2	65	22	25	2312	-	-	-	-	0,78	0,56	0,59	0,55

Precision and recall calculations for daily returns (%) with different market impacts (%) (meaning the percentage of the S&P500 stocks having not less absolute returns than the corresponding return value in different scenarios (AN: adjective-noun based expression recognition; BOW: bag of words approach) – day T

Table 13: Precision and Recal calculations onday T+1

					General financial and economic news				Company specific headline news			
Absolute return	Market impact	DOI+	DOI-	DOUI	Precision (AN)	Recall (AN)	Precision (BOW)	Recall (BOW)	Precision (AN)	Recall (AN)	Precision (BOW)	Recall (BOW)
0,1	45	1034	1103	739	0,85	0,36	0,70	0,37	0,79	0,47	0,83	0,45
0,2	45	906	1000	855	0,85	0,38	0,69	0,38	0,78	0,45	0,84	0,43
0,3	45	780	861	1119	0,85	0,38	0,70	0,39	0,78	0,43	0,90	0,41
0,4	45	641	750	1559	0,86	0,32	0,70	0,32	0,78	0,41	0,89	0,39
0,5	45	535	647	1956	0,86	0,34	0,69	0,33	0,82	0,41	0,86	0,37
0,6	45	458	547	2313	0,86	0,34	0,69	0,33	0,78	0,30	0,80	0,25
0,7	45	397	474	2673	0,86	0,36	0,69	0,35	0,79	0,27	0,73	0,19
0,8	45	328	419	2965	0,86	0,36	0,69	0,35	0,80	0,22	0,75	0,15
0,9	45	276	354	3208	0,86	0,35	0,66	0,33	0,78	0,20	0,71	0,15
1	45	231	294	3429	0,86	0,41	0,66	0,39	0,88	0,28	0,83	0,19
1,1	45	196	250	3614	0,85	0,41	0,65	0,39	0,88	0,35	0,83	0,24
1,2	45	170	215	3770	0,88	0,34	0,62	0,30	0,75	0,19	1,00	0,20
0,1	50	727	805	370	0,85	0,30	0,68	0,29	0,77	0,43	0,83	0,41
0,2	50	624	719	414	0,86	0,33	0,69	0,32	0,78	0,43	0,90	0,41
0,3	50	521	623	581	0,86	0,33	0,69	0,32	0,80	0,41	0,89	0,39
0,4	50	437	531	915	0,86	0,34	0,69	0,33	0,79	0,37	0,84	0,33
0,5	50	369	458	1387	0,86	0,34	0,66	0,32	0,81	0,30	0,77	0,22
0,6	50	308	393	1851	0,86	0,35	0,66	0,33	0,79	0,27	0,73	0,19
0,7	50	256	323	2229	0,85	0,36	0,65	0,34	0,78	0,20	0,71	0,14
0,8	50	212	275	2564	0,85	0,36	0,62	0,33	0,88	0,21	0,83	0,15
0,9	50	181	238	2854	0,85	0,36	0,62	0,33	0,88	0,27	0,83	0,18
1	50	156	198	3085	0,88	0,34	0,62	0,30	0,88	0,35	0,83	0,24
1,1	50	126	170	3272	0,88	0,35	0,62	0,31	0,75	0,19	1,00	0,20
1,2	50	100	138	3464	0,88	0,40	0,62	0,35	0,75	0,33	1,00	0,36
0,1	55	508	580	209	0,86	0,34	0,69	0,33	0,77	0,41	0,90	0,39
0,2	55	415	508	222	0,86	0,31	0,64	0,29	0,78	0,42	0,89	0,41
0,3	55	353	433	297	0,86	0,32	0,64	0,30	0,78	0,33	0,80	0,29
0,4	55	297	376	470	0,85	0,31	0,62	0,28	0,81	0,30	0,77	0,23
0,5	55	246	311	759	0,85	0,31	0,62	0,28	0,83	0,25	0,70	0,17
0,6	55	210	272	1197	0,85	0,35	0,62	0,32	0,78	0,20	0,71	0,14
0,7	55	173	225	1674	0,85	0,36	0,62	0,33	0,88	0,21	0,83	0,15
0,8	55	143	187	2070	0,88	0,36	0,62	0,32	0,88	0,35	0,83	0,24
0,9	55	117	155	2432	0,88	0,38	0,62	0,33	0,83	0,28	1,00	0,20
1	55	93	132	2722	0,88	0,40	0,62	0,35	0,75	0,19	1,00	0,20
1,1	55	74	99	2940	0,85	0,30	0,66	0,28	0,75	0,50	1,00	0,57
1,2	55	61	81	3135	0,85	0,30	0,66	0,28	1,00	0,50	1,00	0,50
0,1	60	353	411	169	0,85	0,31	0,62	0,28	0,76	0,36	0,86	0,35
0,2	60	297	353	171	0,85	0,31	0,62	0,28	0,79	0,37	0,80	0,31
0,3	60	244	300	185	0,85	0,31	0,62	0,28	0,75	0,23	0,78	0,17
0,4	60	200	260	248	0,85	0,36	0,62	0,33	0,78	0,19	0,71	0,13
0,5	60	166	217	372	0,88	0,36	0,62	0,32	0,78	0,20	0,71	0,14
0,6	60	130	176	575	0,87	0,32	0,58	0,27	0,88	0,29	0,83	0,21
0,7	60	112	150	936	0,87	0,34	0,58	0,28	0,88	0,35	0,83	0,24
0,8	60	86	116	1397	0,84	0,25	0,60	0,22	0,75	0,19	1,00	0,20
0,9	60	68	95	1842	0,84	0,25	0,60	0,22	0,75	0,33	1,00	0,33
1	60	53	72	2245	0,84	0,25	0,60	0,22	0,75	0,50	1,00	0,57
1,1	60	48	63	2549	0,84	0,27	0,60	0,23	1,00	0,50	1,00	0,50
1,2	60	39	51	2806	0,84	0,29	0,60	0,25	1,00	1,00	1,00	1,00
0,1	65	233	289	152	0,85	0,29	0,59	0,26	0,81	0,35	0,78	0,29
0,2	65	183	246	152	0,87	0,32	0,58	0,27	0,79	0,27	0,73	0,19
0,3	65	155	207	158	0,87	0,32	0,58	0,27	0,78	0,19	0,71	0,13
0,4	65	134	165	173	0,84	0,22	0,60	0,19	0,78	0,23	0,71	0,16
0,5	65	104	130	212	0,84	0,24	0,60	0,20	0,88	0,33	0,83	0,24
0,6	65	88	108	294	0,84	0,25	0,60	0,22	0,86	0,32	0,80	0,20
0,7	65	63	95	441	0,84	0,25	0,60	0,22	0,75	0,19	1,00	0,21
0,8	65	52	68	664	0,84	0,28	0,60	0,23	0,75	0,33	1,00	0,36
0,9	65	40	56	1063	0,84	0,28	0,60	0,23	1,00	0,50	1,00	0,50
1	65	35	42	1544	0,84	0,33	0,60	0,29	1,00	1,00	1,00	1,00
1,1	65	27	37	1975	0,84	0,33	0,60	0,29	1,00	1,00	1,00	1,00
1,2	65	22	25	2312	0,77	0,23	0,61	0,24	1,00	1,00	1,00	1,00

Precision and recall calculations for daily returns (%) with different market impacts (%) (meaning the percentage of the S&P500 stocks having not less absolute returns than the corresponding return value in different scenarios (AN: adjective-noun based expression recognition; BOW: bag of words approach) – day T+1

Table 14: Precision and Recal calculations on day T-1

					General financial and economic news				Company specific headline news			
Absolute return	Market impact	DOI+	DOI-	DOUI	Precision (AN)	Recall (AN)	Precision (BOW)	Recall (BOW)	Precision (AN)	Recall (AN)	Precision (BOW)	Recall (BOW)
0,1	45	1034	1103	739	0,80	0,54	0,60	0,53	0,74	0,46	0,69	0,45
0,2	45	906	1000	855	0,80	0,54	0,60	0,52	0,72	0,46	0,71	0,47
0,3	45	780	861	1119	0,81	0,54	0,61	0,53	0,75	0,45	0,67	0,43
0,4	45	641	750	1559	0,80	0,57	0,61	0,56	0,75	0,43	0,61	0,39
0,5	45	535	647	1956	0,80	0,55	0,62	0,54	0,76	0,48	0,62	0,42
0,6	45	458	547	2313	0,80	0,55	0,61	0,54	0,76	0,48	0,61	0,41
0,7	45	397	474	2673	0,80	0,56	0,61	0,55	0,79	0,44	0,54	0,33
0,8	45	328	419	2965	0,80	0,56	0,61	0,55	0,86	0,41	0,70	0,35
0,9	45	276	354	3208	0,80	0,56	0,61	0,55	0,83	0,37	0,88	0,37
1	45	231	294	3429	0,80	0,56	0,61	0,55	0,86	0,35	0,85	0,34
1,1	45	196	250	3614	0,80	0,56	0,62	0,55	0,75	0,29	0,86	0,29
1,2	45	170	215	3770	0,80	0,56	0,62	0,55	0,67	0,22	0,83	0,29
0,1	50	727	805	370	0,81	0,58	0,60	0,57	0,71	0,46	0,71	0,47
0,2	50	624	719	414	0,80	0,56	0,61	0,55	0,75	0,46	0,67	0,45
0,3	50	521	623	581	0,80	0,57	0,61	0,56	0,77	0,46	0,64	0,42
0,4	50	437	531	915	0,80	0,57	0,62	0,56	0,76	0,47	0,61	0,41
0,5	50	369	458	1387	0,80	0,56	0,61	0,55	0,81	0,48	0,65	0,42
0,6	50	308	393	1851	0,80	0,56	0,61	0,55	0,86	0,40	0,70	0,33
0,7	50	256	323	2229	0,80	0,56	0,61	0,55	0,86	0,41	0,70	0,35
0,8	50	212	275	2564	0,80	0,57	0,62	0,56	0,86	0,35	0,85	0,34
0,9	50	181	238	2854	0,80	0,57	0,62	0,56	0,80	0,33	0,89	0,35
1	50	156	198	3085	0,80	0,59	0,62	0,59	0,75	0,29	0,86	0,29
1,1	50	126	170	3272	0,80	0,59	0,62	0,58	0,60	0,18	0,80	0,25
1,2	50	100	138	3464	0,80	0,59	0,62	0,58	0,75	0,23	0,75	0,25
0,1	55	508	580	209	0,80	0,57	0,61	0,56	0,75	0,46	0,67	0,45
0,2	55	415	508	222	0,80	0,57	0,62	0,56	0,75	0,45	0,61	0,42
0,3	55	353	433	297	0,80	0,57	0,61	0,56	0,76	0,48	0,62	0,42
0,4	55	297	376	470	0,80	0,57	0,61	0,56	0,88	0,43	0,75	0,38
0,5	55	246	311	759	0,80	0,57	0,62	0,56	0,86	0,40	0,70	0,33
0,6	55	210	272	1197	0,80	0,57	0,62	0,57	0,85	0,39	0,78	0,35
0,7	55	173	225	1674	0,80	0,59	0,62	0,58	0,86	0,41	0,85	0,41
0,8	55	143	187	2070	0,80	0,59	0,62	0,59	0,75	0,29	0,86	0,29
0,9	55	117	155	2432	0,80	0,59	0,62	0,58	0,71	0,26	0,83	0,29
1	55	93	132	2722	0,80	0,58	0,62	0,57	0,75	0,18	0,75	0,20
1,1	55	74	99	2940	0,81	0,55	0,62	0,55	1,00	0,33	0,67	0,22
1,2	55	61	81	3135	0,82	0,54	0,63	0,54	1,00	0,50	0,67	0,33
0,1	60	353	411	169	0,80	0,58	0,61	0,57	0,75	0,47	0,61	0,43
0,2	60	297	353	171	0,80	0,59	0,61	0,59	0,77	0,48	0,62	0,43
0,3	60	244	300	185	0,80	0,58	0,62	0,58	0,87	0,40	0,71	0,34
0,4	60	200	260	248	0,80	0,59	0,62	0,58	0,86	0,40	0,70	0,34
0,5	60	166	217	372	0,81	0,58	0,62	0,58	0,83	0,43	0,88	0,42
0,6	60	130	176	575	0,81	0,58	0,62	0,57	0,83	0,38	0,82	0,36
0,7	60	112	150	936	0,81	0,58	0,62	0,57	0,75	0,30	0,86	0,33
0,8	60	86	116	1397	0,81	0,55	0,62	0,55	0,67	0,22	0,83	0,29
0,9	60	68	95	1842	0,82	0,55	0,63	0,54	1,00	0,21	0,67	0,17
1	60	53	72	2245	0,82	0,54	0,63	0,54	1,00	0,33	0,67	0,22
1,1	60	48	63	2549	0,82	0,54	0,63	0,54	1,00	0,50	0,67	0,33
1,2	60	39	51	2806	0,81	0,54	0,62	0,53	1,00	0,50	0,67	0,33
0,1	65	233	289	152	0,81	0,57	0,62	0,56	0,82	0,45	0,61	0,38
0,2	65	183	246	152	0,81	0,57	0,62	0,57	0,86	0,37	0,70	0,31
0,3	65	155	207	158	0,81	0,58	0,62	0,57	0,86	0,45	0,70	0,39
0,4	65	134	165	173	0,81	0,56	0,62	0,56	0,81	0,42	0,86	0,39
0,5	65	104	130	212	0,81	0,56	0,62	0,56	0,75	0,27	0,86	0,30
0,6	65	88	108	294	0,82	0,55	0,63	0,54	0,75	0,30	0,86	0,33
0,7	65	63	95	441	0,82	0,55	0,63	0,54	1,00	0,22	0,75	0,20
0,8	65	52	68	664	0,82	0,54	0,63	0,54	1,00	0,27	0,67	0,22
0,9	65	40	56	1063	0,83	0,53	0,64	0,53	1,00	0,50	0,67	0,33
1	65	35	42	1544	0,82	0,53	0,63	0,53	1,00	0,50	0,67	0,33
1,1	65	27	37	1975	0,82	0,53	0,63	0,53	1,00	0,50	0,67	0,33
1,2	65	22	25	2312	0,82	0,56	0,63	0,56	-	-	-	-

Precision and recall calculations for daily returns (%) with different market impacts (%) (meaning the percentage of the S&P500 stocks having not less absolute returns than the corresponding return value in different scenarios (AN: adjective-noun based expression recognition; BOW: bag of words approach) – day T-1

Table 15 Precision and Recal calculations

	General financial and economic news				Company specific headline news			
	Precision (AN)	Recall (AN)	Precision (BOW)	Recall (BOW)	Precision (AN)	Recall (AN)	Precision (BOW)	Recall (BOW)
T	0,78	0,51	0,59	0,50	0,71	0,36	0,54	0,25
T-1	0,81	0,56	0,62	0,56	0,82	0,38	0,73	0,34
T+1	0,85	0,33	0,63	0,30	0,82	0,37	0,86	0,33

Average precision and recall values of Table 12, Table 13 and Table 14 for days T, T+1 and T-1 respectively in table form.

Appendix 1: Precision and Recal calculations

Positive noun	Negative noun	Positive adjective	Negative adjective
earn	loss	rose	fell
share	debt	gain	fall
revenue	damage	yield	weak
sale	failure	growth	low
income	crisis	rise	decreas
stock	expenditure	high	decreased
profit	drawback	strong	bear
market	disadvantage	lot	bad
trade	low	unexpected	fallen
economy		good	loose
value		big	lost
return		increas	dropped
investment		increased	tumbled
growth		heavy	deep
success		bull	cut
advantage		large	disappoint
confidence		great	
earning		strength	
high		happy	
		successful	
		unexpected	
		better	
		advance	

Appendix 2: Examples of expected classifications

Expression type	Noun	Adjective	Example sentences (fractions)
PN-NA	stock	fall	... small cap stock fall ...
PN-NA	earn	fell	... wf financial inc report third quart earn fell to ...
PN-NA	share	fell	stock fell yesterday a investor concern that asia economic crisi will continue to erode the profit of american companie depressed share price ...
PN-NA	profit	fell	... abstract credit suisse first boston report year to date profit fell ... to ... million as of aug from ... million a of june ...
PN-PA	income	rose	... dryp corp report third quart net income rose to ...
PN-PA	sale	rose	... abstract canon japan report first half sale rose ...
PN-PA	income	rose	... abstract general mill inc report fiscal third quart net income rose ...
PN-PA	earn	rose	... abstract owen corn report first quart earn rose to million a revenue rose to ...

Appendix 3: Examples of bad classifications

Expression type	Noun	Adjective	Example sentences (fractions)
PN-PA	stock	gain	... bond market rack up furth gain a u stock lose additional ground and investor continue to fear additional turbulence in asian financial market ...
PN-PA	market	high	... abstract investor are beginn to question wheth japanese government bond market can continue to hit new high ...