World Journal of Social Sciences
Vol. 2. No. 4. July 2012. Pp. 218 – 234

Feedforward Neural Networks for Sentiment Detection in Financial News

Caslav Bozic* and Detlef Seese*

With a rise of algorithmic trading volume in recent years, the need for automatic analysis of financial news emerged. We propose a system for quantifying text sentiment based on a Neural Networks predictor. Using methodology from empirical finance we prove a statistically significant relation between the text sentiment of published news and future daily returns.

JEL Codes: C45, D83, and G17

1. Introduction

News stories make a very important information source for traders. News feeds reach huge number of people, and they can initiate massive market movements, like panic selling, or massive buying, but they can also lead to more subtle market movements. Until recently it was mainly the task of human analysts to determine how positive or negative a news story is for a subject company. In general, we call such a positivity or negativity measure 'text sentiment'.

With the rise of algorithmic trading volume in recent years, the need for quantifying qualitative information in textual news and incorporating that additional information in new trading algorithms emerged. This task has to be done on a vast amount of data and in millisecond frequency range, so these requirements render human analysts less useful and machines have to take over the task of quantifying text sentiment.

In the past decade about a dozen of systems and methods trying to solve this task appeared in the literature. They use text mining of publicly accessible financial texts in order to predict market movements. To do this they employ different machine learning approaches, define important features in text in a different way, and use different and often incomparable criteria for performance measurement. In his work (Tetlock 2007) used a fairly simple text sentiment measurement – the number of words classified as 'negative' in the Harvard IV-4 dictionary. He evaluated the results by building the regression between this text sentiment measure and future stock prices of the subject company. This proved statistically significant predicting power of the text sentiment measure. We will use this basic idea of the regression as a performance assessment tool.

* Institute of Applied Informatics and Formal Description Methods, Karlsruhe Institute of Technology (KIT), Germany, e-mail: {bozic, detlef.seese}@kit.edu.

Financial support from the Graduate School 895 "Information Management and Market Engineering" funded by Deutsche Forschungsgemeinschaft (DFG) is gratefully acknowledged.

2. Literature Review

The methodologies used to explore how published news influence market reactions range from a fairly simple content analysis using classical statistical tools, to complex machine-learning methods. The approaches vary from an engineering approach which focuses on implementation and proving economic relevance, to chiefly theoretical approaches whose goal is to describe underlying economic phenomena.

(Lavrenko et al. 2000) use Naïve Bayes classifier to classify news articles from Yahoo!Finance into five groups, according to the influence on particular U.S. stocks. The features were determined automatically and the forecast horizon was from five to ten hours. (Gidófalvi & Elkan 2003) use again naïve Bayes classifier with three categories to recognize articles which have bigger positive or negative influence on constituents of Dow Jones index. With features defined using mutual information measure they work on ten minutes aggregated intraday data. (Fung et al. 2003) partially use commercially available text mining systems to predict a price trend for intraday market movements of some of the stocks listed on the Hong Kong Stock Exchange. For classification purposes they use support vector machines. Finally, Mittermayer and Knolmayer (2006) propose a high frequency forecast system that classifies press releases of publicly traded companies in the U.S. using a dictionary that combines automatically selected features and a hand-crafted thesaurus. For classification the authors use the polynomial version of SVM.

Another group of publications not included in the survey by Mittermayer and Knolmayer (2006b) contains works that do not primary attempt to prove economical relevance of published text by evaluating specifically tailored trading strategies, but rather to find statistically relevant relations between financial indicators and sentiment extracted from the text.

Antweiler and Frank (2004) use Naïve Bayes and SVM classifiers to classify messages posted to Yahoo!Finance and Raging Bull and determine their sentiment. They do not find statistically significant correlation with stock prices, but they find sentiment and volume of messages significantly correlated to trade volumes and volatility. In their methodological paper Das and Chen (2007) offer a variety of classifiers, as well as composed sentiment measure as a result of voting among classifiers. In the illustrative example they analyze Yahoo stock boards and stock prices of 8 technology companies, but they do not find clear evidence that the sentiment index can be predictive for stock prices.

In the corpus of research on the influence of news on market reactions, only a humble fraction employs artificial neural networks. The survey (Mittermayer & Knolmayer 2006b) includes only one article that uses neural networks for classification - (Wüthrich et al. 1998). In this paper its authors propose a system that classifies news articles published on web portals during night. Up, down, and steady are three categories that are defined depending on the influence news have on five equity indices: Dow Jones, Nikkei, FTSE, Hang Seng, and Straits Times. The goal was to forecast the trend of this index value one day ahead. The underlying dictionary is hand-crafted, and classifiers used are naïve

Bayes, nearest neighbour, and a neural network with 423 input nodes, 211 hidden nodes, and three output nodes. The results reported for the neural network classifier are somewhat worse than those reported for nearest neighbour classification.

Two articles of interest that tackle market response to news, but were not included in the survey, are (Liang 2005) and (Liang & Chen 2005). The first paper uses only the volume of posted internet stock news to train a neural network and to predict changes in stock prices, so we can not consider the system proposed there as a real text-mining system. As an extension, the second work (Liang & Chen 2005) employs natural language processing techniques and a hand-crafted dictionary to predict stock returns. The authors use a feedforward neural network with five neurons in the input layer, 27 in the hidden layer, and one output neuron. Since only 500 news items were used for the analysis, no statistical significance of the results could be found.

A kind of explanation for such a small number of systems employing neural networks in financial text mining we might find in viewpoints similar to this one:

> *Both our own pre-tests (not shown here) and comparative empirical studies provide evidence that the classification performance of SVM is superior to both parametric data mining techniques, e.g. Naïve Bayes, and non-parametric data mining techniques, e.g. k-Nearest Neighbour or Neural Networks. Moreover, as already stated above, SVM "is usually less vulnerable to the over-fitting problem [and] the solution of SVM is always unique and globally optimal". That is the reason why we decided SVM to be the method of choice in this paper.*
>
> (Groth & Muntermann 2010)

The authors quote (Joachims 1998) and (Yang & Liu 1999) that compared different approaches to text classification, but forget that in recent years we witnessed a huge progress in neural network methodology, like fast training algorithms for deep multilayer neural networks developed by (Hinton & Salakhutdinov 2006). The ability of neural networks to capture very complex patterns, and new learning algorithms that enable training in acceptable time, call for a reconsideration of previous statements.

3. Methodology

The proposed system is based on a Neural Network predictor. The Neural Network has to be trained first. After its training the system is ready to take a text of the news story about a particular company as an input, and it produces a numerical text sentiment measure as an output. Our hypothesis is that the text sentiment produced using this kind of predictor corresponds with future returns of the company's stock.

As a source of financial news we use the archive of all news items published via Reuters NewsScope in year 2003. This same dataset is already analysed in the literature, for example in (Hellinger 2008), so using this dataset gives us a possibility to compare results and complement existing findings in the area of sentiment detection in financial setting.

Besides the news text 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.

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

$$R_{OC} = \ln \frac{P_O(t)}{P_C(t-1)}$$

$$R_{OO} = \ln \frac{P_O(t)}{P_O(t-1)}$$

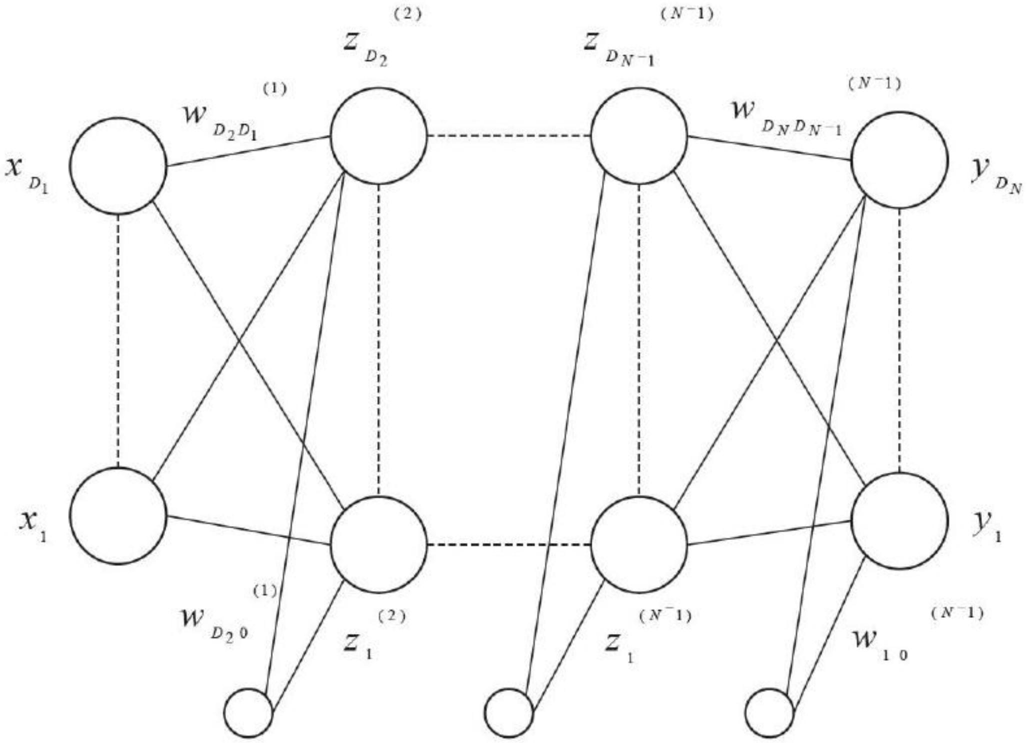
$$R_{CO} = \ln \frac{P_C(t)}{P_O(t)}$$

$$R_{CC} = \ln \frac{P_C(t)}{P_C(t-1)}$$

For training we singled out news about four companies: Apple Computer Inc., International Business Machines Corp., Microsoft Corp., and Oracle Corp. As news items can be rather long, we kept only the paragraphs where the subject company is mentioned and four surrounding paragraphs. The words with only one or two characters are discarded. All other words are stemmed, and their absolute frequencies in the text are calculated. Stemming is a process of mapping from particular word to its root or stem by stripping off the ending of a given word. Each of the 11781 distinct words in our training set represents one dimension of the training vector. Each news item represents one training vector. The target value is determined according to the next day’s open-to-open return of the subject company. To decrease the overlapping between time range of news publication and time range of returns, all the news items published after the closing time of the market (3:30 pm, local time) are considered to belong already to the next date.

The system uses fairly simple feedforward Neural Network with an input layer, two hidden layers, and an output layer. The information in feedforward Neural Network flow only in one direction and their graph representation doesn't have any cycles. The size of the input layer depends on the properties of the input text and it is defined by the number of distinct words in the training dataset. In our case the number of neurons in the input layer is 11781. The two hidden layers consist of 16 and 8 neurons, while the output layer has one or two neurons, depending on a version of neural network, as explained below.

Figure 1: Neural network structure



The general structure of feed forward neural network is shown in Figure 1. The neural network function can be described using the following equations:

$$z_{j_1}^{(1)} = x_{j_1}$$

$$z_{j_{k+1}}^{(k+1)} = h\left(\sum_{i=1}^{D_k} w_{j_{k+1}i}^{(k)} z_i^{(k)} + w_{j_{k+1}0}^{(k)}\right)$$

$$y_{jN} = z_{jN}^{(N)}$$

$$k = 1, \ldots, (N-1) \quad N \in \mathbb{N}$$

$$j_p = 1, \ldots, D_p$$

$$D_q \in \mathbb{N} \quad q = 1, \ldots, N$$

Input and output vectors are denoted by **x** and **y**, respectively. The parameters ω_{ji} are weights, while ω_{j0} is denoted as *bias*. To determine the value of output, each neuron transforms its sum of inputs using a function h() which is called *activation function*. In our case the activation function is a logarithmic function.

We compared three versions of neural network. They differ in structure and training procedure.

- Version 1 has one neuron in the output layer, and the output value of that neuron is the value of future return.
- Version 2 has two neurons in the output layer. During the training, the neurons can take only one of two values --- zero or one. The first neuron is set to one if the future return is positive; the second neuron is set to one if the future return is negative.
- Version 3 has also two neurons in the output layer. If the future return is positive, first neuron is set to the value of that return, while the second is set to zero; if the future return is negative, the second neuron is set to the absolute value of that return, while the first is set to zero.

4. Findings

Each of 107266 news items published in the year 2003 that mentions any of the Russell 3000 companies is classified. The paragraphs which mention the subject company and the four surrounding paragraphs are singled out, the words with only one or two letters are discarded, and the other words stemmed. This word vector is fed to the Neural Network predictor, and as an output we get the text sentiment.

All the text sentiment results for one company and one day (in this case the next day starts already with closing the market – 3:30 pm local time) are averaged, and aligned with the corresponding return for the same company and the same date.

$$R_{OO}(t,c) = \alpha_0 S(t,c) + \alpha_1 S(t-1,c) + \alpha_2 S(t-2,c) + \alpha_3 S(t-3,c) + \sum_{i=2}^{10} \beta_i dd_i(c) + \gamma \quad (1)$$

At this point we need a way to determine the predicting power of the text sentiment measure. As suggested in the literature (see Rachev et al. 2010), the appropriate tool for the analysis of how two entities behave together and for describing their joint distribution is correlation. That is why we built correlation coefficients using text sentiment values lagged up to three days S(t), S(t-1), S(t-2), S(t-3), open-to-open return R_{OO}, and variables representing companies' market capitalization decile dd_1 to dd_{10}. The results can be seen in Tables 1 – 6.

The most important relation for us, the correlation between today's open-to-open return R_{OO} and yesterday's sentiment value S(t-1), is positive in all versions, except for version 3 run b. That is a strong sign that the sentiment value of one day is correlated with a future returns, with lag equals to one day. It is also visible that both version 1 and version 2 have negative correlation coefficients between today's open-to-open return R_{OO} and sentiment value of two days ago S(t-2). This is in accordance with effective markets hypothesis, so these changes in returns predicted by sentiment value are only temporary shocks and they are reversed within a few days.

To confirm our findings further, we applied the multivariate linear regression method. If the observed text sentiment measure actually correlates with the future stock returns, as our hypothesis states, 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 estimated regression parameters for linear regression with open-to-open return R_{OO} as a dependent variable using ordinary least squares method. Contemporaneous text sentiment value S(t), text sentiment value from the day before S(t-1), two days before S(t-2), and three days before S(t-3) were used as independent variables. This has been done with respect to the subject company c, which is represented as an additional parameter in the equation, besides time t. All the companies in our dataset were ordered according to their market capitalization (total market value of all shares of the company), and divided into 10 equally sized groups. In this way the values for ten additional “dummy” variables dd_1 to dd_{10} were created (being 1 if the subject company was assigned to the respective group, and 0 otherwise). These “dummy” variables were included into the regression to account for the variations of returns as a result of company's size.

Table 1: Correlation coefficients of lagged sentiment values sent, sent(-1), sent(-2), sent(-3), open-to-open return oo, and variables representing companies' market capitalization decile dd2 to dd10 in the Verison 1a of the neural network

	sent	sent(-1)	sent(-2)	sent(-3)	oo	dd2	dd3	dd4	dd5	dd6	dd7	dd8	dd9	dd10
sent	1.00000	0.07274	0.03779	0.03842	0.00019	-0.00235	-0.00088	0.00151	0.00037	-0.00150	0.00078	-0.00178	-0.00423	0.00935
sent(-1)	0.07274	1.00000	0.07274	0.03779	0.00131	-0.00235	-0.00088	0.00151	0.00037	-0.00150	0.00078	-0.00178	-0.00423	0.00935
sent(-2)	0.03779	0.07274	1.00000	0.07274	-0.00054	-0.00235	-0.00088	0.00151	0.00037	-0.00150	0.00078	-0.00178	-0.00423	0.00935
sent(-3)	0.03842	0.03779	0.07274	1.00000	-0.00101	-0.00235	-0.00088	0.00151	0.00037	-0.00150	0.00078	-0.00178	-0.00423	0.00935
oo	0.00019	0.00131	-0.00054	-0.00101	1.00000	0.00459	0.00158	0.00201	-0.00051	0.00198	-0.00143	-0.00286	-0.00360	-0.00435
dd2	-0.00235	-0.00235	-0.00235	-0.00235	0.00459	1.00000	-0.10820	-0.10755	-0.10940	-0.11085	-0.10993	-0.11113	-0.11254	-0.11059
dd3	-0.00088	-0.00088	-0.00088	-0.00088	0.00158	-0.10820	1.00000	-0.10797	-0.10982	-0.11128	-0.11035	-0.11156	-0.11298	-0.11102
dd4	0.00151	0.00151	0.00151	0.00151	0.00201	-0.10755	-0.10797	1.00000	-0.10917	-0.11062	-0.10970	-0.11090	-0.11230	-0.11036
dd5	0.00037	0.00037	0.00037	0.00037	-0.00051	-0.10940	-0.10982	-0.10917	1.00000	-0.11252	-0.11158	-0.11280	-0.11423	-0.11225
dd6	-0.00150	-0.00150	-0.00150	-0.00150	0.00198	-0.11085	-0.11128	-0.11062	-0.11252	1.00000	-0.11306	-0.11430	-0.11575	-0.11374
dd7	0.00078	0.00078	0.00078	0.00078	-0.00143	-0.10993	-0.11035	-0.10970	-0.11158	-0.11306	1.00000	-0.11335	-0.11478	-0.11279
dd8	-0.00178	-0.00178	-0.00178	-0.00178	-0.00286	-0.11113	-0.11156	-0.11090	-0.11280	-0.11430	-0.11335	1.00000	-0.11604	-0.11403
dd9	-0.00423	-0.00423	-0.00423	-0.00423	-0.00360	-0.11254	-0.11298	-0.11230	-0.11423	-0.11575	-0.11478	-0.11604	1.00000	-0.11547
dd10	0.00935	0.00935	0.00935	0.00935	-0.00435	-0.11059	-0.11102	-0.11036	-0.11225	-0.11374	-0.11279	-0.11403	-0.11547	1.00000

Table 2: Correlation coefficients of lagged sentiment values sent, sent(-1), sent(-2), sent(-3), open-to-open return oo, and variables representing companies' market capitalization decile dd2 to dd10 in the Verison 1b of the neural network

	sent	sent(-1)	sent(-2)	sent(-3)	oo	dd2	dd3	dd4	dd5	dd6	dd7	dd8	dd9	dd10
sent	1.00000	0.08315	0.04060	0.03638	-0.01001	-0.01620	-0.01484	-0.00995	-0.00813	-0.00499	-0.00060	0.00806	0.01070	0.05392
sent(-1)	0.08315	1.00000	0.08315	0.04060	0.00236	-0.01620	-0.01484	-0.00995	-0.00813	-0.00499	-0.00060	0.00806	0.01070	0.05392
sent(-2)	0.04060	0.08315	1.00000	0.08315	-0.00195	-0.01620	-0.01484	-0.00995	-0.00813	-0.00499	-0.00060	0.00806	0.01070	0.05392
sent(-3)	0.03638	0.04060	0.08315	1.00000	0.00211	-0.01620	-0.01484	-0.00995	-0.00813	-0.00499	-0.00060	0.00806	0.01070	0.05392
oo	-0.01001	0.00236	-0.00195	0.00211	1.00000	0.00459	0.00158	0.00201	-0.00051	0.00198	-0.00143	-0.00286	-0.00360	-0.00435
dd2	-0.01620	-0.01620	-0.01620	-0.01620	0.00459	1.00000	-0.10820	-0.10755	-0.10940	-0.11085	-0.10993	-0.11113	-0.11254	-0.11059
dd3	-0.01484	-0.01484	-0.01484	-0.01484	0.00158	-0.10820	1.00000	-0.10797	-0.10982	-0.11128	-0.11035	-0.11156	-0.11298	-0.11102
dd4	-0.00995	-0.00995	-0.00995	-0.00995	0.00201	-0.10755	-0.10797	1.00000	-0.10917	-0.11062	-0.10970	-0.11090	-0.11230	-0.11036
dd5	-0.00813	-0.00813	-0.00813	-0.00813	-0.00051	-0.10940	-0.10982	-0.10917	1.00000	-0.11252	-0.11158	-0.11280	-0.11423	-0.11225
dd6	-0.00499	-0.00499	-0.00499	-0.00499	0.00198	-0.11085	-0.11128	-0.11062	-0.11252	1.00000	-0.11306	-0.11430	-0.11575	-0.11374
dd7	-0.00060	-0.00060	-0.00060	-0.00060	-0.00143	-0.10993	-0.11035	-0.10970	-0.11158	-0.11306	1.00000	-0.11335	-0.11478	-0.11279
dd8	0.00806	0.00806	0.00806	0.00806	-0.00286	-0.11113	-0.11156	-0.11090	-0.11280	-0.11430	-0.11335	1.00000	-0.11604	-0.11403
dd9	0.01070	0.01070	0.01070	0.01070	-0.00360	-0.11254	-0.11298	-0.11230	-0.11423	-0.11575	-0.11478	-0.11604	1.00000	-0.11547
dd10	0.05392	0.05392	0.05392	0.05392	-0.00435	-0.11059	-0.11102	-0.11036	-0.11225	-0.11374	-0.11279	-0.11403	-0.11547	1.00000

Table 3: Correlation coefficients of lagged sentiment values sent, sent(-1), sent(-2), sent(-3), open-to-open return oo, and variables representing companies' market capitalization decile dd2 to dd10 in the Verison 2a of the neural network

	sent	sent(-1)	sent(-2)	sent(-3)	oo	dd2	dd3	dd4	dd5	dd6	dd7	dd8	dd9	dd10
sent	1.00000	0.11517	0.06591	0.05697	0.00069	-0.01032	-0.00719	-0.00328	-0.00595	-0.00544	0.00466	-0.00439	-0.00997	0.04682
sent(-1)	0.11517	1.00000	0.11517	0.06591	0.00377	-0.01032	-0.00719	-0.00328	-0.00595	-0.00544	0.00466	-0.00439	-0.00997	0.04682
sent(-2)	0.06591	0.11517	1.00000	0.11517	-0.00155	-0.01032	-0.00719	-0.00328	-0.00595	-0.00544	0.00466	-0.00439	-0.00997	0.04682
sent(-3)	0.05697	0.06591	0.11517	1.00000	0.00030	-0.01032	-0.00719	-0.00328	-0.00595	-0.00544	0.00466	-0.00439	-0.00997	0.04682
oo	0.00069	0.00377	-0.00155	0.00030	1.00000	0.00459	0.00158	0.00201	-0.00051	0.00198	-0.00143	-0.00286	-0.00360	-0.00435
dd2	-0.01032	-0.01032	-0.01032	-0.01032	0.00459	1.00000	-0.10820	-0.10755	-0.10940	-0.11085	-0.10993	-0.11113	-0.11254	-0.11059
dd3	-0.00719	-0.00719	-0.00719	-0.00719	0.00158	-0.10820	1.00000	-0.10797	-0.10982	-0.11128	-0.11035	-0.11156	-0.11298	-0.11102
dd4	-0.00328	-0.00328	-0.00328	-0.00328	0.00201	-0.10755	-0.10797	1.00000	-0.10917	-0.11062	-0.10970	-0.11090	-0.11230	-0.11036
dd5	-0.00595	-0.00595	-0.00595	-0.00595	-0.00051	-0.10940	-0.10982	-0.10917	1.00000	-0.11252	-0.11158	-0.11280	-0.11423	-0.11225
dd6	-0.00544	-0.00544	-0.00544	-0.00544	0.00198	-0.11085	-0.11128	-0.11062	-0.11252	1.00000	-0.11306	-0.11430	-0.11575	-0.11374
dd7	0.00466	0.00466	0.00466	0.00466	-0.00143	-0.10993	-0.11035	-0.10970	-0.11158	-0.11306	1.00000	-0.11335	-0.11478	-0.11279
dd8	-0.00439	-0.00439	-0.00439	-0.00439	-0.00286	-0.11113	-0.11156	-0.11090	-0.11280	-0.11430	-0.11335	1.00000	-0.11604	-0.11403
dd9	-0.00997	-0.00997	-0.00997	-0.00997	-0.00360	-0.11254	-0.11298	-0.11230	-0.11423	-0.11575	-0.11478	-0.11604	1.00000	-0.11547
dd10	0.04682	0.04682	0.04682	0.04682	-0.00435	-0.11059	-0.11102	-0.11036	-0.11225	-0.11374	-0.11279	-0.11403	-0.11547	1.00000

Table 4: Correlation coefficients of lagged sentiment values sent, sent(-1), sent(-2), sent(-3), open-to-open return oo, and variables representing companies' market capitalization decile dd2 to dd10 in the Verison 2b of the neural network

	sent	sent(-1)	sent(-2)	sent(-3)	oo	dd2	dd3	dd4	dd5	dd6	dd7	dd8	dd9	dd10
sent	1.00000	0.12618	0.07623	0.06629	-0.00105	-0.01486	-0.01117	-0.00712	-0.00886	-0.00704	0.00550	-0.00273	-0.00933	0.06467
sent(-1)	0.12618	1.00000	0.12618	0.07623	0.00368	-0.01486	-0.01117	-0.00712	-0.00886	-0.00704	0.00550	-0.00273	-0.00933	0.06467
sent(-2)	0.07623	0.12618	1.00000	0.12618	-0.00195	-0.01486	-0.01117	-0.00712	-0.00886	-0.00704	0.00550	-0.00273	-0.00933	0.06467
sent(-3)	0.06629	0.07623	0.12618	1.00000	-0.00073	-0.01486	-0.01117	-0.00712	-0.00886	-0.00704	0.00550	-0.00273	-0.00933	0.06467
oo	-0.00105	0.00368	-0.00195	-0.00073	1.00000	0.00459	0.00158	0.00201	-0.00051	0.00198	-0.00143	-0.00286	-0.00360	-0.00435
dd2	-0.01486	-0.01486	-0.01486	-0.01486	0.00459	1.00000	-0.10820	-0.10755	-0.10940	-0.11085	-0.10993	-0.11113	-0.11254	-0.11059
dd3	-0.01117	-0.01117	-0.01117	-0.01117	0.00158	-0.10820	1.00000	-0.10797	-0.10982	-0.11128	-0.11035	-0.11156	-0.11298	-0.11102
dd4	-0.00712	-0.00712	-0.00712	-0.00712	0.00201	-0.10755	-0.10797	1.00000	-0.10917	-0.11062	-0.10970	-0.11090	-0.11230	-0.11036
dd5	-0.00886	-0.00886	-0.00886	-0.00886	-0.00051	-0.10940	-0.10982	-0.10917	1.00000	-0.11252	-0.11158	-0.11280	-0.11423	-0.11225
dd6	-0.00704	-0.00704	-0.00704	-0.00704	0.00198	-0.11085	-0.11128	-0.11062	-0.11252	1.00000	-0.11306	-0.11430	-0.11575	-0.11374
dd7	0.00550	0.00550	0.00550	0.00550	-0.00143	-0.10993	-0.11035	-0.10970	-0.11158	-0.11306	1.00000	-0.11335	-0.11478	-0.11279
dd8	-0.00273	-0.00273	-0.00273	-0.00273	-0.00286	-0.11113	-0.11156	-0.11090	-0.11280	-0.11430	-0.11335	1.00000	-0.11604	-0.11403
dd9	-0.00933	-0.00933	-0.00933	-0.00933	-0.00360	-0.11254	-0.11298	-0.11230	-0.11423	-0.11575	-0.11478	-0.11604	1.00000	-0.11547
dd10	0.06467	0.06467	0.06467	0.06467	-0.00435	-0.11059	-0.11102	-0.11036	-0.11225	-0.11374	-0.11279	-0.11403	-0.11547	1.00000

Table 5: Correlation coefficients of lagged sentiment values sent, sent(-1), sent(-2), sent(-3), open-to-open return oo, and variables representing companies' market capitalization decile dd2 to dd10 in the Verison 3a of the neural network

	sent	sent(-1)	sent(-2)	sent(-3)	oo	dd2	dd3	dd4	dd5	dd6	dd7	dd8	dd9	dd10
sent	1.00000	0.05813	0.02753	0.02256	0.01070	0.01817	0.01591	0.01516	0.00749	0.00133	-0.00276	-0.01496	-0.02971	-0.03082
sent(-1)	0.05813	1.00000	0.05813	0.02753	0.00154	0.01817	0.01591	0.01516	0.00749	0.00133	-0.00276	-0.01496	-0.02971	-0.03082
sent(-2)	0.02753	0.05813	1.00000	0.05813	0.00134	0.01817	0.01591	0.01516	0.00749	0.00133	-0.00276	-0.01496	-0.02971	-0.03082
sent(-3)	0.02256	0.02753	0.05813	1.00000	0.00095	0.01817	0.01591	0.01516	0.00749	0.00133	-0.00276	-0.01496	-0.02971	-0.03082
oo	0.01070	0.00154	0.00134	0.00095	1.00000	0.00459	0.00158	0.00201	-0.00051	0.00198	-0.00143	-0.00286	-0.00360	-0.00435
dd2	0.01817	0.01817	0.01817	0.01817	0.00459	1.00000	-0.10820	-0.10755	-0.10940	-0.11085	-0.10993	-0.11113	-0.11254	-0.11059
dd3	0.01591	0.01591	0.01591	0.01591	0.00158	-0.10820	1.00000	-0.10797	-0.10982	-0.11128	-0.11035	-0.11156	-0.11298	-0.11102
dd4	0.01516	0.01516	0.01516	0.01516	0.00201	-0.10755	-0.10797	1.00000	-0.10917	-0.11062	-0.10970	-0.11090	-0.11230	-0.11036
dd5	0.00749	0.00749	0.00749	0.00749	-0.00051	-0.10940	-0.10982	-0.10917	1.00000	-0.11252	-0.11158	-0.11280	-0.11423	-0.11225
dd6	0.00133	0.00133	0.00133	0.00133	0.00198	-0.11085	-0.11128	-0.11062	-0.11252	1.00000	-0.11306	-0.11430	-0.11575	-0.11374
dd7	-0.00276	-0.00276	-0.00276	-0.00276	-0.00143	-0.10993	-0.11035	-0.10970	-0.11158	-0.11306	1.00000	-0.11335	-0.11478	-0.11279
dd8	-0.01496	-0.01496	-0.01496	-0.01496	-0.00286	-0.11113	-0.11156	-0.11090	-0.11280	-0.11430	-0.11335	1.00000	-0.11604	-0.11403
dd9	-0.02971	-0.02971	-0.02971	-0.02971	-0.00360	-0.11254	-0.11298	-0.11230	-0.11423	-0.11575	-0.11478	-0.11604	1.00000	-0.11547
dd10	-0.03082	-0.03082	-0.03082	-0.03082	-0.00435	-0.11059	-0.11102	-0.11036	-0.11225	-0.11374	-0.11279	-0.11403	-0.11547	1.00000

Table 6: Correlation coefficients of lagged sentiment values sent, sent(-1), sent(-2), sent(-3), open-to-open return oo, and variables representing companies' market capitalization decile dd2 to dd10 in the Verison 3b of the neural network

	sent	sent(-1)	sent(-2)	sent(-3)	oo	dd2	dd3	dd4	dd5	dd6	dd7	dd8	dd9	dd10
sent	1.00000	0.29556	0.22083	0.19849	0.01040	0.05883	0.05159	0.04564	0.03744	0.02799	0.01794	-0.00558	-0.03684	-0.25069
sent(-1)	0.29556	1.00000	0.29556	0.22083	-0.00172	0.05883	0.05159	0.04564	0.03744	0.02799	0.01794	-0.00558	-0.03684	-0.25069
sent(-2)	0.22083	0.29556	1.00000	0.29556	0.00431	0.05883	0.05159	0.04564	0.03744	0.02799	0.01794	-0.00558	-0.03684	-0.25069
sent(-3)	0.19849	0.22083	0.29556	1.00000	0.00141	0.05883	0.05159	0.04564	0.03744	0.02799	0.01794	-0.00558	-0.03684	-0.25069
oo	0.01040	-0.00172	0.00431	0.00141	1.00000	0.00459	0.00158	0.00201	-0.00051	0.00198	-0.00143	-0.00286	-0.00360	-0.00435
dd2	0.05883	0.05883	0.05883	0.05883	0.00459	1.00000	-0.10820	-0.10755	-0.10940	-0.11085	-0.10993	-0.11113	-0.11254	-0.11059
dd3	0.05159	0.05159	0.05159	0.05159	0.00158	-0.10820	1.00000	-0.10797	-0.10982	-0.11128	-0.11035	-0.11156	-0.11298	-0.11102
dd4	0.04564	0.04564	0.04564	0.04564	0.00201	-0.10755	-0.10797	1.00000	-0.10917	-0.11062	-0.10970	-0.11090	-0.11230	-0.11036
dd5	0.03744	0.03744	0.03744	0.03744	-0.00051	-0.10940	-0.10982	-0.10917	1.00000	-0.11252	-0.11158	-0.11280	-0.11423	-0.11225
dd6	0.02799	0.02799	0.02799	0.02799	0.00198	-0.11085	-0.11128	-0.11062	-0.11252	1.00000	-0.11306	-0.11430	-0.11575	-0.11374
dd7	0.01794	0.01794	0.01794	0.01794	-0.00143	-0.10993	-0.11035	-0.10970	-0.11158	-0.11306	1.00000	-0.11335	-0.11478	-0.11279
dd8	-0.00558	-0.00558	-0.00558	-0.00558	-0.00286	-0.11113	-0.11156	-0.11090	-0.11280	-0.11430	-0.11335	1.00000	-0.11604	-0.11403
dd9	-0.03684	-0.03684	-0.03684	-0.03684	-0.00360	-0.11254	-0.11298	-0.11230	-0.11423	-0.11575	-0.11478	-0.11604	1.00000	-0.11547
dd10	-0.25069	-0.25069	-0.25069	-0.25069	-0.00435	-0.11059	-0.11102	-0.11036	-0.11225	-0.11374	-0.11279	-0.11403	-0.11547	1.00000

Table 7: Estimated coefficients of multivariate OLS regression according to Equation 1 with open-to-open return R_{OO}, as a dependent variable and lagged text sentiment values sent, sent(-1), sent(-2), sent(-3) and dummy variables representing companies' market capitalization decile dd2 to dd10 as independent variables, expressed in basis points

	Version 1a	Version 1b	Version 2a	Version 2b	Version 3a	Version 3b	RNSE Data
sent	0.8929	-37.6926 **	1.0978	-2.1921	78.4174 *	137.5680 **	103.8900 ***
sent(-1)	6.6190	13.6879 *	8.3696	8.9637	3.8260	-86.6401 ***	21.8671 **
sent(-2)	-2.4927	-6.3603 ***	-3.7937 *	-4.0681 **	4.7288	39.6125 ***	-9.3731 **
sent(-3)	-4.6630	10.7841 ***	0.8915	-0.8002	2.1339	-17.6992	-8.7930 **
dd2	1.8038 ***	1.8260 ***	1.8208 ***	1.8086 ***	1.8624 ***	1.8145 ***	1.8276 ***
dd3	-1.4862 ***	-1.4570 ***	-1.4798 ***	-1.4849 ***	-1.3985 ***	-1.4305 ***	-1.3898 ***
dd4	-1.0331 ***	-0.9787 ***	-1.0405 ***	-1.0359 ***	-0.9392 ***	-0.9431 ***	-0.8733 ***
dd5	-3.7841 ***	-3.7181 ***	-3.7826 ***	-3.7855 ***	-3.5898 ***	-3.6373 ***	-3.7142 ***
dd6	-1.0743 ***	-0.9901 ***	-1.0750 ***	-1.0781 ***	-0.8031 ***	-0.8683 ***	-0.8188 ***
dd7	-4.7189 ***	-4.6113 ***	-4.7515 ***	-4.7336 ***	-4.3984 ***	-4.4555 ***	-4.5275 ***
dd8	-6.2846 ***	-6.1298 ***	-6.2886 ***	-6.2921 ***	-5.8138 ***	-5.8784 ***	-5.6995 ***
dd9	-7.0089 ***	-6.8404 ***	-6.9948 ***	-7.0105 ***	-6.3620 ***	-6.4181 ***	-6.6953 ***
dd10	-7.8744 ***	-7.4664 ***	-8.0534 ***	-7.9492 ***	-7.2037 ***	-5.9884 ***	-6.6596 ***
Constant	20.0064 ***	20.0466 ***	20.0009 ***	20.0032 ***	20.1147 ***	20.0834 ***	20.0346 ***

The results are presented in Table 7. In columns we have results of three different versions of neural network, each of them represented with two different runs, ordered by they statistical significance (the first run produced results that are less significant than the second run). The last column gives the values of the same benchmark applied to sentiment data produced by Reuters NewsScope Sentiment Engine (RNSE). The results are expressed in basis points, representing one hundredth of a percent.

The coefficients in the table are estimations of the following parameters from Equation 1: $\alpha_0 \ldots \alpha_3$ for lagged *sent* variables, $\beta_2 \ldots \beta_{10}$ for variables *dd2-dd10*, and γ for *Constant* factor. The statistical significance is expressed according to Table 1. Given the observed data, p value represents the probability that the null hypothesis is true. In our case, the null hypothesis states that the particular coefficient is zero, hence the daily return doesn't depend on the observed variable, or in other words that the observed variable doesn't predict daily return.

Table 8: Statistical significance of the results

	p value
***	< 1%
**	< 5%
*	< 10%
otherwise	>= 10%

The coefficients in Table 7 support our hypothesis for the version 1 of the neural network and the second run. All regression coefficients are significantly different from zero and they have the following meaning: if the text sentiment extracted using neural network version 1 increased 1 unit, the next day’s open-to-open return would increase in average 13.69 basis points, having all other influences constant. This is the most important result related to our hypothesis, because it states that changes in text sentiment can predict next day’s change in open-to-open return.

Further confidence in our regression results can be gained by looking at Tables 1 to 6 and noting that there are no high values of correlation coefficients. High values of correlation coefficients would signify a high degree of multicolinearity between independent variables, what could deteriorate robustness of the regression model.

The performance of this type of neural network is strongly dependent on the initial values of the weights, which were randomly assigned in this case. This influences the instability of performance and different results between training sessions. It is represented by big difference between best and worst result of the same network. Having a benchmark at hand, we can solve this problem by using distinct train and development datasets. This is common practice in natural language processing and offers a possibility to train the neural network on one set of data and to run the test and observe the results on the distinct development dataset. Then we can simply discard the training sessions with unsatisfactory performance.

5. Conclusion

We presented a system for automatic financial news analytics by determining text sentiment using a Neural Network predictor. The employed machine learning method uses a feedforward Neural Network with two hidden layers. The performance is assessed by an empirical finance approach, which offers a possibility to prove statistical significance of the results.

From the presented results it is visible that, if measured by a benchmark we proposed, some of the neural network structures can achieve performance that is comparable to other state of the art systems. Neural network with two hidden layers and one neuron in output layer produces a text sentiment measure that is highly significantly related to next day's return. The relation between the text sentiment extracted in this way and open-to-open return two and three days ahead is significant to the 1% level, while the state of the art proprietary industrial system achieves lower significance of 5%. Future work would be extending these results by using Deep Multilayer Neural Networks with more than two hidden layers for determining text sentiment.

References

Antweiler, W & Frank, MZ 2004, 'Is All That Talk Just Noise? The Information Content of Internet Stock Message Boards', *The Journal of Finance*, 59(3), pp. 1259-1294

Bozic, C 2009, 'FINDS - Integrative services", *Computer Systems and Applications, IEEE/ACS International Conference*, pp. 61-62

Das, S & Chen, M 2007, 'Yahoo! for Amazon: Sentiment extraction from small talk on the web', *Management Science*, 53(9), pp. 1375-1388

Fung, GPC, Xu Yu, J & Lam, W 2003, 'Stock prediction: Integrating text mining approach using real-time news', *Proceedings of IEEE International Conference on Computational Intelligence for Financial Engineering*, pp. 395 - 402

Gidófalvi, G & Elkan, C 2003, *Using news articles to predict stock price movements*

Groth, S & Muntermann, J 2010, *Discovering Intraday Market Risk Exposures in Unstructured Data Sources: The Case of Corporate Disclosures*, pp.1-10.

Hellinger, U 2008, 'Event and Sentiment Detection in Financial Markets', *5th European Semantic Web Conference ESWC 2008 Ph. D. Symposium,* pp. 31-35

Hinton, G & Salakhutdinov, R 2006, 'Reducing the dimensionality of data with neural networks', *Science*, 313(5786), p. 504.

Joachims, T 1998, *Text categorization with Support Vector Machines: Learning with many relevant features*, pp. 137-142.

Lavrenko, V, Schmill, M, Lawrie, D, Ogilvie, P, Jensen, D & Allan, J 2000, 'Mining of Concurrent Text and Time-Series', *Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*

Liang, X 2005, *Impacts of Internet Stock News on Stock Markets Based on Neural Networks*, p. 811.

Liang, X & Chen, RC 2005*, Mining Stock News in Cyberworld Based on Natural Language Processing and Neural Networks*, pp. 893-898.

Mittermayer, M & Knolmayer, G 2006, 'NewsCATS: A News Categorization and Trading System', *IEEE International Conference on Data Mining*, pp. 1002-1007

Mittermayer, M & Knolmayer, G 2006b, *Text mining systems for market response to news: A survey*.

Pfrommer, J, Hubschneider, C, & Wenzel S 2010, *Sentiment Analysis on Stock News using Historical Data and Machine Learning Algorithms*

Rachev, S, Hoechstoetter, M, Fabozzi, F & Focardi, S 2010, *Probability and Statistics for Finance,* Wiley

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

Wüthrich, B, Permunetilleke, D, Leung, S, Cho, V, Zhang, J, & Lam W 1998, *Daily prediction of major stock indices from textual www data* .

Yang, Y & Liu X 1999, *A re-examination of text categorization methods*, pp. 42-49.