Wirtschaftsinformatik Proceedings 2009 Wirtschaftsinformatik

2009

# SUPPORTING INVESTMENT MANAGEMENT PROCESSES WITH MACHINE LEARNING TECHNIQUES

Sven S. Groth
*E-Finance Lab Frankfurt*

Jan Muntermann
*Goethe-University Frankfurt*

# SUPPORTING INVESTMENT MANAGEMENT PROCESSES WITH MACHINE LEARNING TECHNIQUES

Sven S. Groth[1], Jan Muntermann[2]

***Abstract***

*IT support and especially IT innovations represent crucial success factors in investment management and financial decision making. The application of machine learning techniques for intraday decision support constitutes such a crucial factor, but existing literature has not been capable of melding latest research results from both the text and the capital market research field. Combining fundamental research results from both domains, it is possible to develop a forecasting model that performs better than both alternative internal (random) benchmarks and external prior research benchmarks. The approach is tested by means of "classical" evaluation methods such as well as by a simulation.*

## 1. Introduction

IT Innovations play an important role in the financial industry because most business processes are widely managed or supported by computer systems. In securities trading, decision makers rely on efficient information delivery including extensive information access, information management and decision support on the basis of information technology. Here, the provision of unstructured data such as news publications that need to be interpreted manually represent a problem domain being far from settled today. In the following, this problem field is addressed and a decision support approach is presented that is based on machine learning techniques, so-called Support Vector Machines (SVM). SVM provide capabilities for classifying unstructured data such as news publications. Until today, there exist only few contributions that present SVM-based approaches to support financial decision making. Generally, these contributions collect news documents and analyze their contents in order to forecast price effects following the news publication. In the existing literature that is presented in the following literature review, this forecasting objective has been addressed in a way in which fundamental research results from the finance research community (presented in Section 2.2) has been ignored. In the following Section 3, a novel SVM-based decision support approach for which forecasting objectives have been derived from empirical capital market research is presented. It is hypothesized that this approach is superior compared to existing approaches and it therefore can be beneficial to investment managers. An assessment of these benefits is provided by a simulation-based evaluation in Section 4. The evaluation compares

[1] E-Finance Lab Frankfurt

[2] Goethe-University Frankfurt

and statistically analyzes the outcomes of different trading strategies on the basis of empirical datasets including news stories and stock price series. The results provide evidence that the presented approach provides promising forecasting qualities and could therefore support decision processes in investment management.

## 2. Literature Review & Derived Hypotheses

### 2.1. Text Mining Research

The potentials of machine learning and text mining techniques to support financial investment processes and decision making are subject to research since the late 1990s. In this research field, a collection of news documents (i.e. unstructured data) is mapped to financial metrics such as stock returns following the news' publication dates (i.e. structured data). Then, text documents are analyzed in a way to detect patterns in order to estimate the structured data mapped (for example the following stock price reaction).

In the financial context, [25] present an early contribution being based on text mining techniques. They observe financial news articles being published on news portals and estimate the price trends of major market indices on a daily basis. Based on these market trend forecasts, they derive a trading strategy that achieves long-term profits that outperform several actively managed funds. In [11], a text mining approach is presented to identify news stories that will most likely induce a stock price trend. In contrast to [25], intraday price trends are analyzed during the period of one hour prior to and one hour subsequent to the news publications. Compared to a random trader strategy, realizable cumulative trading profits appears significantly higher when utilizing the trend information provided.

In [19], the problem of selecting only highly relevant company announcements is addressed in order to forward them wireless to mobile devices of retail investors. Document classification is realized with SAS Enterprise Miner in order to forecast the abnormal stock price behaviour observed during the publication dates (i.e. documents are mapped to daily stock returns). Achieving a classification error of 59%, the authors conclude that the analyzed text documents contain too much noise. The system design presented in [14] provides text mining-based functionalities to predict stock price adjustments following press releases. These were classified as good or bad news if a significant price reaction has been observed during 1 hour following the release date. By simulating a corresponding trading strategy and a noise trader, evidence is provided that the random trader can be outperformed. Another text mining-based system design is presented in [6] for which a forecasting model has been developed on the basis of 350,000 news releases and the corresponding stock price movements. Again, the authors present a trading strategy being based on long-term price effects (a holding period of 1 to 7 days following a news release).

### 2.2. Capital Market Research

As illustrated in the previous section, existing text mining research to support financial decision-making has a focus on diverse news events and forecasting objectives. However, all papers aim at forecasting capital market effects that follow the publication of news stories, i.e. to present promising approaches to exploit subsequent price effects. In financial research, capital market efficiency and the question how quickly share prices adjust to newly published information represents a research domain attracting much attention. The underlying theoretical concept, i.e. the semi-strong-form of market efficiency, is addressed with so-called event studies that empirically observe and analyze the speed at which share prices adjust to new information available to the market. The empirical findings show that the price adjustment on capital markets is not a matter of hours or days but a process of minutes in which significant abnormal returns can be observed. The following Table 1 summarizes research results in the field of intraday event study analysis.

Table 1 illustrates that for different event types analyzed, significant abnormal price behaviour has been observed during the 30 to 35 minutes following the event dates. For ad hoc disclosures that were published by companies to comply with the Securities Trading Act in Germany (WpHG), the results of [15] show that most of the abnormal price behaviour (i.e. excess returns) can be observed within the 15 minutes following the event dates.

As illustrated in the previous section, existing text mining research to support financial decision-making has a focus on diverse news events and forecasting objectives. With forecasting and evaluation periods of 1 hour to several days, existent text mining research seems to disregard these empirical findings. Consequently, data selection and forecasting objectives have not been derived from this fundamental literature.

**Table 1: Price adjustments following the publication of news stories**

| | Barclay & Litzen-berger 1988 [1] | Gosnell et al. 1996 [7] | Patell & Wolfson 1984 [16] | Muntermann & Güttler 2007 [15] |
|---|---|---|---|---|
| Event type | Public offerings of corporate securities | Shifts in dividend policies | Earnings and dividend announcements | Ad hoc disclosures |
| Significant price effects observed within (from; to) minutes | (0; 30)<br>(5; 35) | (0; 30) bad<br>(0; 15) good | (0; 30)<br>(5; 35) | (0; 15)<br>(0; 30) |
| Return metric | Mean cumulative return | Mean abnormal return | Mean percent return | Median cumulative abnormal return |
| Price effect observed (from; to) | 2.65% (0; 30)<br>0.088% (5;35) | 0.09% (bad news)<br>0.04% (good news) | 0.225% (earnings)<br>0.384% (dividends) | 0.0221 (0; 15)<br>0.0019 (0; 30) |

## 2.3. Hypotheses

Literature on text mining is coined by many useful contributions that aim at improving forecasting methods. The majority of these proposals, however, are rather technical in nature and usually do not address specific application domains such as financial decision making. Moreover, scholars within the financial decision making domain – as summarized above – did not (or just to a small extent) incorporate prior capital market research into their text mining applications. The authors believe that this issue should not be disregarded. The following insights from capital market research are expected to improve results both with regard to internal benchmarks (*Hypothesis 1*; a simulation based-approach) and external benchmarks (*Hypothesis 2*; previous empirical studies):

- Limitation of news scope to those stories that are expected to contain pricing-relevant information (e.g. ad hoc disclosures).
- Use of a time frame for both the learning task and the investment strategy for which significant intraday abnormal price reactions have been observed (e.g. 15 minutes after publication).

## 3. Intraday Decision Support with Text Mining Techniques

### 3.1. General Setup & Dataset

Given above defined hypotheses, the set-up shown in Figure 1 is constructed. “Traditional” model evaluation by k-fold cross validation is thereby extended by a simulation to prove economic suitability of the “improved” text mining concept. To the authors’ knowledge, this is also the first study to conduct sound statistical tests in the course of simulation runs.

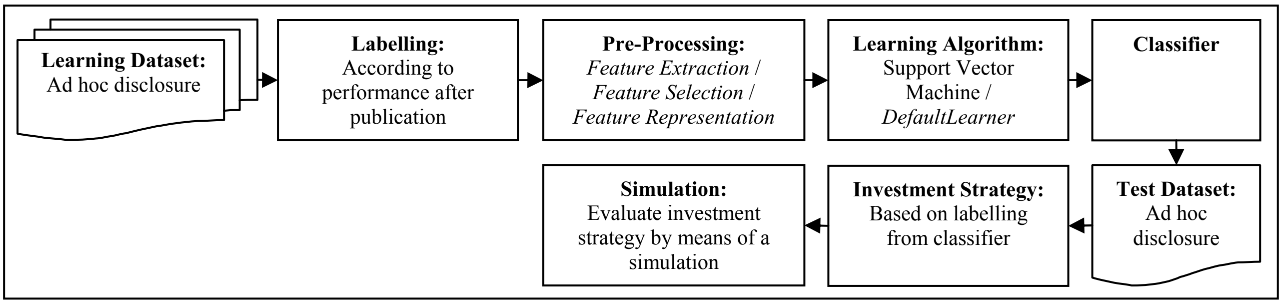


**Figure 1. Study set-up**

The dataset consists of ad hoc disclosures published by the Deutsche Gesellschaft für Ad-hoc-Publizität (DGAP) on behalf of the companies whose shares are admitted to trading on an organized market in Germany. To fulfil legal requirements, these companies have to publish immediately any insider information or other information being highly relevant to investors. The authors make use of a dataset that has – in part – already been evaluated by [4] and [15]. While the original dataset comprises of 160 ad hoc disclosures that were published between 2003-08-01 and 2004-08-31, the extended dataset comprises of 423 ad hoc disclosures that were published between 2003-08-01 and 2005-07-29. Both the original and the extended dataset include those ad hoc disclosures that were published during stock exchange trading hours only, which is illustrated in the following Figure 2.

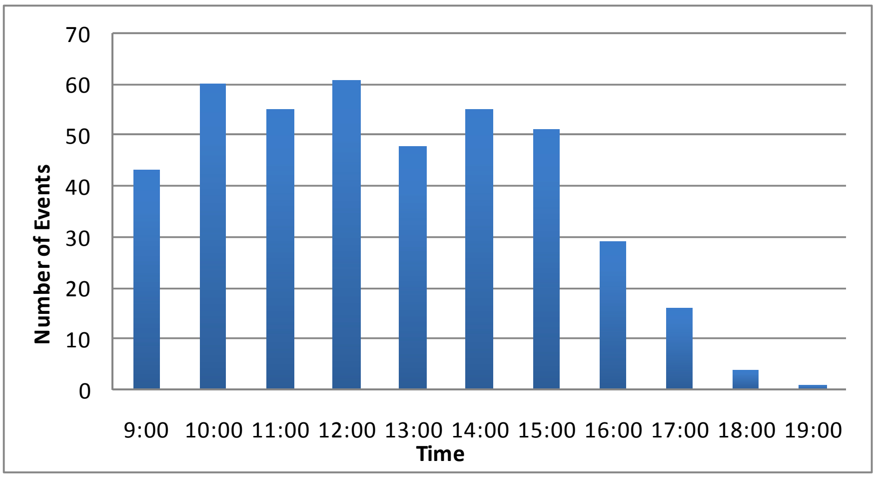


**Figure 2: Histogram of news events by daytime**

Similar publication numbers were observed from 9 a.m. to 3 p.m. and fewer after 3 p.m. Less publication numbers during the later periods also come from trading hours on XETRA where closing changed from 8 p.m. to 5:30 pm. during the observation period. For each of these ad hoc disclosures, the company’s intraday stock price series (of the publication date) were obtained. Consequently, the authors map ad hoc disclosure contents with the intraday price behaviour following their publication.

### 3.2. Labelling

Conducting a *supervised learning* task, the respective learning algorithms need to be provided with a set of correctly labelled examples to learn from. Whenever the learning task is to mimic human behaviour or thinking (e.g. sentiment analysis), labelling naturally needs to be conducted by hand and therefore might turn out to be quite labour intensive. Contrary, the authors aim to enable the learning algorithm to identify (previously unknown) patterns in news announcements that are capable of indicating abnormal price reactions. Pursuing this attempt, automatic labelling according to the price reaction following ad hoc disclosure publication will be conducted. The simple return measure *sr* (see Formula 1) is used as a proxy for abnormal price reactions. It takes into account the underlying stock's price $p_e$ at event (publication) time and at event time plus an interval of 15 minutes. This measure obviously neither takes into account price movements within the interval nor price movements by an index (market trend). Analogue to [14], it is assumed that the interval, in our case the 15 minutes following the news publication during which [15] detected abnormal price behaviour, is too short to be significantly influenced by simultaneous market fluctuations. The documents are labelled either "positive" or "negative" depending on the simple return following the disclosure publication.

$$sr = \frac{p_{e+15}}{p_e} \tag{1}$$

### 3.3. Text Pre-Processing

Text mining techniques are based on "traditional" data mining algorithms that are not able to cope with plain text. Against this background, the main pre-processing task is to be found with the transformation of textual documents into a numeric representation. The applied pre-processing steps *Feature Extraction*, *Feature Selection*, and *Feature Representation* [1] will shortly be explained.

During *Feature Extraction*, a dictionary of words and phrases is generated that describes the document adequately. The authors make use of a simple *StringTokenizer* [26] which uses the Unicode specification to identify separators by non-letter characters. Moreover, words with little meaning, but frequent appearances, are dealt with by means of a *Stopword List*. Aiming at a reduction of features within the feature set [22], both the *Porter Stemmer* [17] and a simple *GermanStemmer* [26] are made use of. Regarding *Feature Selection*, i.e. the elimination of those tokens that contain few or relatively less important information, [4] finds that available feature selection methods do not necessarily "perform better than using all features available". Therefore, these methods are not applied in the following. Finally, during *Feature Representation* each document is represented by previously extracted and selected number of features. The respective feature weightings are generated by frequency measure *tfidf*. Thereby *tf* denotes *term frequency* and *idf* denotes inverse *document frequency* [12].

### 3.4. Text Mining and Classification

Comparative empirical studies [10] [27] provide evidence that the performance of SVMs for classification tasks seems to be superior to other methods and this is the reason why the authors make use of this algorithm. SVM was firstly introduced by [21] for solving two-class recognition problems. In contrast to parametric classification methods, the addressed non-parametric method SVM interprets the input matrix spatially. The general idea is to find the decision surface that maximizes the margin between the data points (classes). Thereby maximizing the margin between a separating hyperplane and the nearest data points is undertaken by means of structural risk minimization. In case of originally linearly non-separable data points, the original data vectors may be mapped to higher dimensional space to achieve linear separability again [27]. For calculation efficiency reasons a linear kernel, i.e a function in lower-dimensional space that exhibits similar behaviour as the original functions in higher-dimensional space, is made use of. [8] have shown that the use of a linear kernel seems sufficient whenever the number of features is exceptionally large as it is the case with text mining.

### 3.5. Classic Model Evaluation

Model evaluation is conducted by means of a *k-fold (k=10) cross validation*, which seems to be the preferred method of choice in situations with small datasets at hand [23]. In the course of cross validation, the performance measures (overall) accuracy, recall and precision are calculated for each round and are averaged at the end (*micro averaging*). In order to sensitize for classes of unequal size the results of an additional *DefaultLearner* [13] are presented as well. The *DefaulLearner* creates a model based on a default value (e.g. mode of the true labels in case of classification) for all examples.

The results for "classical" model evaluation can be found in Table 2. Regarding overall accuracy, SVM does not perform better than the *DefaultLearner*. The *DefaultLearner* reaches an overall accuracy level of 60.76% because the classes "positive" and "negative" are unequal in size, i.e. 257 vs. 166. Therefore, class "positive" recall is by definition 100% and class "positive" precision is 60.76%. These values are again better than the ones observed with SVM. The *DefaulLearner* class "negative" recall and precision, however, are by definition both 0%. Contrary, SVM reveals better – but not necessarily good – class "negative" recall and precision figures.

Above results show that the evaluation of text mining classification tasks needs to be conducted by means of several evaluation metrics, including benchmark learners such as the guessing-equivalent *DefaultLearner*. Derived insights, however, are not clear enough to either corroborate or discard the usefulness / profitability of above introduced intraday text mining approach. It is, for example, not clear whether or not, in the course of an investment strategy based on the SVM learner, the rather good classification results for the class "positive" offset expected losses from the poor classification quality of the class "negative".

**Table 2. Learning task results**

| Method | Accuracy (micro) | Precision (micro) | | Recall (micro) | |
|---|---|---|---|---|---|
| | | Class "Positive" | Class "Negative" | Class "Positive" | Class "Negative" |
| *SVM* | 56.50% +/- 5.15% (56.50%) | 60.41% +/- 3.60% (60.40%) | 35.14% +/- 22.33% (41.18%) | 82.49% +/- 8.15% (82.49%) | 16.95% +/- 11.81% (16.87%) |
| *Default-Learner* | 60.75% +/- 1.20% (60.76%) | 60.76% | 0% | 100% | 0% |

For above results, ad hoc disclosures were labelled "positive", if sr-15 >= 0 and "negative" if sr-15 < 0.

Against this background, model evaluation methods need to be expanded if text mining techniques are applied to support intraday financial decision-making. An appropriate methodology, i.e. a simulation approach, is introduced below.

## 4. Simulation-Based Evaluation

To evaluate the forecasting quality of the SVM-based forecasting model in the proposed application context, the authors compare a SVM-based trading strategy with a *DefaultLearner*-based strategy and with a random trader. On the basis of a testing dataset (236 events that were not used as model learning set), the outcome of the different trading patterns (model-based strategies or random trader) are simulated in the following (Section 4.1). The resulting return populations then provide the basis for statistical hypotheses testing in order to address the question if the developed SVM-based model can provide a promising technology basis and whether or not it can outperform random traders or other trading strategies (Section 4.2).

### 4.1. Simulation Setup

The simulation is intended to provide further insights into the classification quality of above introduced intraday text mining framework. Pursuing this attempt, opposed to previous k-fold cross validation, a one-time split into a learning dataset and a test dataset is conducted. The models that were developed on the basis of the learning dataset are applied on the test dataset and the outcomes will be used as input to investment strategies. Thereby both the *DefaultLearner* and the random trader strategy should serve as benchmarks to the SVM strategy.

All three investment strategies have in common that an investment decision, on the company the ad hoc disclosure is about, needs to be made. (Dis-)investment is undertaken at event time, i.e. time of ad hoc disclosures publication, and the liquidation of positions is conducted at event time plus 15 minutes. If there are no tick data available exactly at event time (or event time plus 15 minutes), the next tick is made use of. For the calculation of returns, above introduced simple return measure *sr* is used. If a news event from the test dataset is assigned a "positive" label, the respective security is bought at event time and sold at event time plus 15 minutes (see Figure 3). In case of a "negative" label, a negative price movement is expected and therefore a short position is being taken in this security. While both the *DefaultLearner* and the *SVM* models are created by means of above described learning process, the random trader investment decision is based on a stochastic process. At each event the probability of receiving a positive or negative label, i.e. either going long or short, is equal. The 5,000 simulation runs for the random trader result in 1,180,000 return figures.

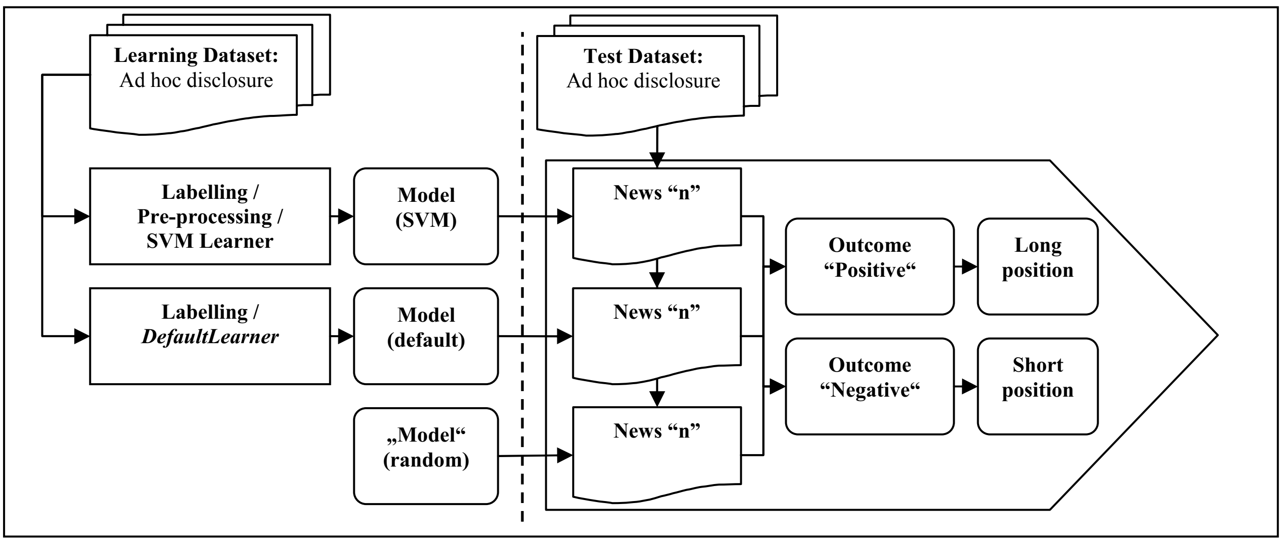


**Figure 3. Simulation set-up**

The simulation results, i.e. the return populations $R$ of the (1) SVM-based trading strategy $R_{SVM}$, (2) *DefaultLearner*-based strategy $R_{DefaultLearner}$, and (3) the random trader $R_{RandomTrader}$, provide the basis for the following statistical evaluations.

## 4.2. Empirical Evaluation Results

Descriptive statistics of the return populations $R_{SVM}$, $R_{DefaultLearner}$, and $R_{RandomTrader}$ are given in the following Table 3.

**Table 3. Descriptive statistics for return populations**

| | $R_{SVM}$ | $R_{DefaultLearner}$ | $R_{RandomTrader}$ |
|---|---|---|---|
| Population Size | 236 | 236 | 1,180,000 |
| Mean in % | 1.0490 | 0.3717 | 0.0139 |
| Standard Deviation | 0.05330 | 0.05420 | 0.05421 |

Compared to [11] and [14], the observed mean return of the SVM-based trading strategy appears significantly higher (*Hypothesis 2*). Furthermore, the population means provide first evidence that the SVM-based trading strategy outperforms both the *DefaultLearner*-based strategy and the random trader (given a lower return variance). To statistically explore this finding, in line with above defined main *Hypothesis 1* corresponding null- and alternative hypotheses are formulated as follows:

$$H_0 : \mu(R_{SVM}) \le \mu(R_{DefaultLearner}) \qquad H_1 : \mu(R_{SVM}) > \mu(R_{DefaultLearner})$$

and

$$H_0 : \mu(R_{SVM}) \le \mu(R_{RandomTrader}) \qquad H_1 : \mu(R_{SVM}) > \mu(R_{RandomTrader})$$

If one or both null hypotheses can be rejected, the authors can statistically prove a higher population mean return of $R_{SVM}$ compared to $R_{DefaultLearner}$ and/or $R_{RandomTrader}$ at a given level of significance. Since the populations feature common standard deviation variances, pooled-variances *t*-tests for comparing the means of two independent samples have been applied and test results are summarized in Table 4.

**Table 4. *t*-Test statistics for comparing return population means**

| | Null Hypotheses | |
|---|---|---|
| | $\mu(R_{SVM}) \le \mu(R_{DefaultLearner})$ | $\mu(R_{SVM}) \le \mu(R_{RandomTrader})$ |
| Pooled variance | 0.2938% | 0.2889% |
| Degrees of freedom | 1180234 | 470 |
| *t*-value | 1.369** | 2.933*** |

*** and ** indicate significance at the 1% and 5% level.

Since both null hypotheses can be rejected at high levels of significance, the test results provide evidence that a trading strategy being based on the developed SVM-based forecasting model and being applied to our event testing set significantly outperforms both the *DefaultLearner*-based strategy as well as the random trader behaviour (*Hypothesis 1*).

## 5. Summary and Conclusion

Having conducted a literature review on both text mining and capital market research that is relevant in the context of the investment process of intraday decision-making, it turned out that existing research has primarily been aiming at improving text mining applications by means of rather technical refinements. However, in this paper it is argued that findings from capital market research are at least as important as results from relevant text mining research. Against this background, two major alterations were proposed: First, the choice of event window is determined by the period in which event studies observed significant intraday abnormal price reactions. As price adjustment on capital markets is not a matter of days or hours, the 15 minute time period after news publication is used. Second, pricing-relevant information is not expected to be inherent to each and every news story. Therefore, investors should – in the first place – concentrate on those news publications that contain pricing-relevant information such as regulatory-driven ad hoc disclosures. It is hypothesized that these alterations improve results.

The evaluation of the proposed (text mining) system's quality, however, also revealed that the application of "classical" domain-independent measures such as accuracy, precision or recall are – in isolation – not sufficient to exclusively judge on its' quality. That is the reason why a domain specific simulation has been conducted. Thereby an SVM-based trading strategy is compared to both a *DefaultLearner*-based strategy and a random trader strategy. Statistical test procedures provide evidence that the SVM-based trading strategy significantly outperforms the other strategies. It follows that above alterations with regard to capital market research insights proved to be valuable – especially in comparison to [11] and [14].

Further research aims at including transaction costs into the simulation-based evaluation, too. Moreover, the relative advantageous of – by definition – pricing-relevant news types need to be tested in other countries as well.

## 6. References


[1] BARCLAY, M. J., LITZENBERGER, R. H. (1988) Announcement Effects of New Equity Issues and the Use of Intraday Price Data, in Journal of Financial Economics, 21, 1, 71-99.

[2] BRÜCHER, H., KNOLMAYER, G. and MITTERMAYER, M. A. (2002) Document Classification Methods for Organizing Explicit Knowledge, in: Proceedings of the 3rd European Conference on Organizational Knowledge, Learning and Capabilities (ALBA), Athens.

[3] DHAR, V.,STEIN, R. (1997) Intelligent Decision Support Methods, Prentice Hall, Upper Saddle River.

[4] GROTH, S. S.,MUNTERMANN, J. (2008) A Text Mining Approach to Support Intraday Financial Decision Making, in: Proceedings of the Fourteenth Americas Conference on Information Systems (AMCIS), Toronto, Canada.

[5] FORMAN, G. (2003) An Extensive Empirical Study of Feature Selection Metrics for Text Classification, in: Journal of Machine Learning Research, 3, 1289-1305.

[6] FUNG, G. P. C., YU, J. X., LU, H. (2005) The Predicting Power of Textual Information on Financial Markets, in: IEEE Intelligent Informatics Bulleting, 5, 1, 1-10.

[7] GOSNELL, T. F., KEOWN, A. J., PINKERTON, J. M. (1996) The Intraday Speed of Stock Price Adjustment to Major Dividend Changes: Bid-Ask Bounce and Order Flow Imbalances, in: Journal of Banking & Finance, 20, 2, 247-266.

[8] HSU, C. W., CHANG, C. C., LIN, C. J. (2007) A Practical Guide to Support Vector Classification, Working Paper, National Taiwan University.

[9] HULL, D. (1996) Stemming Algorithms: A Case Study for Detailed Evaluation, in: Journal of the American Society for Information Science, 47, 1, 70-84.

[10] JOACHIMS, T. (1998) Text Categorization with Support Vector Machines: Learning with Many Relevant Features, in: Proceedings of the 10th European Conference on Machine Learning (ECML), Chemnitz, Germany, 137-142.

[11] LAVRENKO, V., SCHMILL, M., LAWRIE, D., OGILVIE, P., JENSEN, D., ALLAN, J. (2000) Language Models for Financial News Recommendation, in: Proceedings of the 9th International Conference on Information and Knowledge Management, November 6-11, McLean, VA, USA, 389-396.
[12] LEWIS, D. D. (1992) Representation and Learning in Information Retrieval, Dissertation University of Massachusetts, Amherst.
[13] MIERSWA, I., WURST, M., KLINKENBER, G. R., SCHOLZ, M., EULER, T. (2006) YALE: Rapid Prototyping for Complex Data Mining Tasks, in: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2006).
[14] MITTERMAYER, M. (2004) Forecasting Intraday Stock Price Trends with Text Mining Techniques, in: R. H. Sprague (Eds.) Proceedings of the 37th Annual Hawaii International Conference on System Sciences, Computer Society Press, Los Alamitos, CA, USA.
[15] MUNTERMANN, J., GÜTTLER, A. (2007) Intraday Stock Price Effects of Ad Hoc Disclosures: The German Case, in: Journal of International Financial Markets, Institutions and Money, 17, 1, 1-24.
[16] PATELL, J. M., WOLFSON, M. A. (1984) The Intraday Speed of Adjustment of Stock Prices to Earnings and Dividend Announcements, in: Journal of Financial Economics, 13, 2, 223-252.
[17] PORER, M. F. (1980) An Algorithm for Suffix Stripping, in: Program, 14, 3, 130-137.
[18] ROBERSTON, C., GEVA, S., WOLFF, R. (2007) Predicting the Short-Term Market Reaction to Asset Specific News: Is Time Against Us?, in: J. Z. Huang and Y. Ye (Eds.) Proceedings Industry Track Workshop, 11th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD 2007), Nanjing, China, 1-13.
[19] SCHULZ, A., SPILIOPOULOU, M., WINKLER, K. (2003) Kursrelevanzprognose von Ad-hoc-Meldungen: Text Mining wider die Informationsüberlastung im Mobile Banking, in: W. Uhr, W. Esswein and E. Schoop (Eds.) Wirtschaftsinformatik 2003: Medien, Märkte, Mobilität, Band II, Physika, Heidelberg, Germany.
[20] THOMAS, J. D. (2003) News and Trading Rules, Dissertation Carnegie Mellon University, Pittsburgh.
[21] VAPNIK, V. (1995) The Nature of Statistical Learning Theory, Springer, New York.
[22] WITTEN, I. H., MOFFAT, A., BELL, T. C. (1999) Managing Gigabytes: Compressing and Indexing Documents and Images, 2nd edition, San Francisco.
[23] WITTEN, I. H., FRANK, E. (2001) Data Mining: Praktische Werkzeuge und Techniken für das maschinelle Lernen, Munich / Vienna.
[24] WOODRUFF, C. S., SENCHAK, A. (1988) Intradaily Price-Volume Adjustments of NYSE Stocks to Unexpected Earnings, in: Journal of Finance, 43, 2, 467-491.
[25] WUTHRICH, B., CHO, V., LEUNG, S., PERMUNETILLEKE, D., SANKARAN, K., ZHANG, J. (1998) Daily stock market forecast from textual web data, in: Proceedings of IEEE International Conference on Systems, Man, and Cybernetics, October 11-14, San Diego, CA, USA, 2720-2725.
[26] WURST, M. (2007) The Word Vector Tool: User Guide, Operator Reference, Developer Tutorial, URL: "http://wvtool.sf.net", [Access date: December 2007].
[27] YIMING, Y., XIN, L. (1999) A re-examination of text categorization methods, in: Proceedings of ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'99), 42-49.