ISBN: 978-972-8924-40-9 © 2007 IADIS

TEXT MINING SYSTEMS FOR PREDICTING MARKET RESPONSE TO NEWS

Marc-André Mittermayer
Swiss Capital Group, CH 8039 Zurich, Switzerland

Gerhard F. Knolmayer
Institute of Information Systems, University of Bern, CH 3012 Bern, Switzerland

ABSTRACT

Over the last 10 years, several prototypes for predicting the short-term market reactions to news based on text mining techniques have been developed. Thus far no detailed comparison of the systems and their performance is available. This paper describes the main systems developed to forecast price trends and presents a framework for comparing the approaches. The prototypes differ in the text mining methods applied and the data sets used for performance evaluation. Some (mostly implicit) assumptions of these evaluations are rather unrealistic with respect to properties of financial markets and the simulated performance results cannot be achieved in reality. Best performance results are obtained with NewsCATS and we summarize main differences between earlier prototypes and this system.

KEYWORDS

Text Mining; Text Categorization; Forecasting; Financial Markets; Performance Evaluation; Comparison of Systems.

1. INTRODUCTION

Text mining approaches to predicting financial markets are comparatively rare due to the difficulty of extracting relevant information from unstructured data. Since 1998 several prototypes were developed, often without recognizing already existing systems.

The first goal of this paper is to briefly describe the prototypes developed thus far. The second aim is to develop a framework which is helpful in comparing the prototypes. The third goal is to discuss the adequacy of text mining, especially of automated text categorization, to predict stock price movements.

The forecasting prototypes discussed in Section 2 apply automated text categorization. During a learning phase the algorithms capture the structures inherent in pre-categorized sample documents. This results in classifiers used in the operational phase to categorize other documents. No standard document collection is available for the type of systems described in Section 2; therefore a labeling step has to be performed.

2. PROTOTYPES AND REPORTED PERFORMANCES

To the best of our knowledge, eight prototypes exist today which try to forecast price trends for single stocks or indices, volatilities, or exchange rates. In this paper we critically summarize the five systems that focus on price trends in chronological order of their first appearance. Thus, we do not cover the prototypes predicting volatilities (Schulz et al. 2003; Seo et al. 2004) or exchange rates (Peramunetilleke and Wong 2002). A more detailed discussion of all eight prototypes is provided by Mittermayer (2005).

2.1 Prototype Developed by Wüthrich et al.

This prototype attempts to forecast the 1-day trend of five major equity indices at 7:45 a.m. Hong Kong Time (Wüthrich et al. 1998; Cho 1999; Cho et al. 1999). The daily forecast is based on news articles published overnight on web portals. The documents are labeled according to a 3-category model. The first (second) category contains news articles followed by 1-day periods with the associated equity index increasing (decreasing) by at least 0.5%. The third category contains all other news articles. The threshold values are chosen so that roughly one third of the trading sessions fall into each of the three categories.

In the feature definition phase experts manually created a dictionary of 423 features, defined as tuples of words combined with the logical operator AND. The dictionary has not been published; however, a few examples like "bond AND strong", "dollar AND falter" or "dow AND rebound" are given. The authors trained a Naïve Bayes classifier, a Nearest Neighbor classifier, and a Neural Net.

During the operational phase the prototype categorized all newly published articles overnight. The numbers of news articles in each category were counted and, depending on the number of elements in each category, the prototype triggered a buy recommendation, a short recommendation, or advised to do nothing. Based on these recommendations the authors simulated roundtrips: They virtually bought (sold) an index as soon as the prototype triggered a buy (sell) recommendation. The positions were held for exactly one trading session; by the end of the day the system was back in cash. The prototype was tested with 60 days of data from December 1997 to beginning of March 1998. In 40% (Straits Times) to 46.7% (FTSE) of the cases it decided correctly, whereas a random trader simply guessing the next-day's trend, based on a uniform distribution for three categories, would achieve only 33.3%.

Averaged over all five indices a profit of 5.2% resulted. Since the prototype triggered signals only on 41 out of the 60 days, the profit per roundtrip can be calculated as 5.2%/41 = 0.13% or 13 basis points (bps). At first glance this seems surprising since the information released overnight should be fully included in the next day's opening prices. However, when looking a bit closer at the experimental setup, one recognizes that there is a bias in estimating the next day's opening price and, thus, in the performance results: The researchers assumed next day's opening prices, on average, to be identical to the closing prices of the previous trading session. This is inappropriate because the prototype, on average, buys too low and sells short too high. Thus, the performance achieved in the simulation cannot be achieved in reality.

2.2 Prototype Developed by Lavrenko et al.

The prototype Ænalyst was developed around 2000 at the University of Massachusetts Amherst (Lavrenko et al. 1999; Lavrenko et al. 2000a; Lavrenko et al. 2000b). Ænalyst aims to forecast very short-term (intraday) price trends of a subset of U.S. stocks by analyzing news articles published by YAHOO!Finance. The features were determined automatically using TFxIDF (Term Frequency times Inverse Document Frequency) as feature selection technique. A 5-category model with the categories "Surge", "Slight+", "No Recommendation", "Slight-", and "Plunge" was applied. To label the news articles, the authors first segmented the stock price time series with a piecewise linear regression into small trend windows. News articles published in the h hours preceding the start of a trend window in which the price trend had a slope $s \geq 0.75$ ($0.5 \leq s \leq 0.75$) were put into the category "Surge" ("Slight+"); the other categories were determined accordingly. The authors mention that parameter values h between 5 and 10 provide best results.

The classifier was trained with a Naïve Bayes approach. During the operational phase the prototype triggered a buy (short) recommendation if an incoming news article was assigned to the categories "Surge" or "Slight+" ("Slight-" or "Plunge"). Based on these recommendations, the authors performed virtual roundtrips in the U.S. stock market, assuming that one could enter the stock market at the time the news appeared. The market was exited when the investment was 1% or more in the profit zone or at the latest after 60 minutes. This rule is asymmetric because no stop loss limit was defined; no rationale was given for this exit strategy.

The prototype was tested based on 10-minute stock price data between mid-March and April 2000. In each roundtrip USD 10,000 were invested. After a testing period of 40 days a result of 280,000 USD was achieved by performing about 12,000 transactions, resulting in a profit per roundtrip of 23 bps. However, this quite impressive result is put into perspective if we consider some details of the simulation.

One shortcoming is the fact that the authors included only those 127 U.S. stocks showing the largest positive or negative price movements in the period investigated. Such a selection cannot be conducted ex ante and leads to a substantial bias towards highly volatile stocks, reducing the risk of noise trades. Furthermore, it seems to be very unrealistic that within 40 days, 127 stocks can generate 12,000 different news articles that trigger a buy or short recommendation. Certain events were possibly reported in several articles and the system reacted to all of them. The authors also assumed that unlimited funds are available for trading. Of course, institutional investors may leverage their investment by borrowing money and, thus, invest a multiple of the originally available amount. But even for highly creditworthy institutions this multiple is typically in the single digits. By contrast Ænalyst, on average, invests a capital of more than USD 400,000 (multiple greater than 40!). As usual, transaction costs are neglected in the simulation; however, considering the huge number of transactions executed, this omission is more critical than in other papers.

2.3 Prototype Developed by Elkan/Gidófalvi

Another prototype aiming to forecast stock price trends (Gidófalvi 2001; Gidófalvi and Elkan 2003) divides the stock price time series around the publication of a news article into windows of influence. For instance, a window ranging from 0 to 20 means that most of the price adjustment occurs in the 20 minutes following the publication. In the learning phase the documents were labeled according to a 3-category model. The first (second) category consists of news leading to a price increase (decrease) of at least 0.2% during the window of influence. The remaining news fell into the third category.

Feature definition was done automatically by using MI (Mutual Information) as selection criterion. 1,000 words with highest MI values were used as features. The first 100 words are listed (Gidófalvi and Elkan 2003); it is surprising that most of them do not refer to stock prices. The top five words are "sbc", "msft", "websphere", "db", and "index". The learning phase was finished by training a Naïve Bayes classifier. If the prototype sorts an incoming article into one of the first two categories, a virtual roundtrip is performed. The asymmetric exit strategy of Lavrenko et al. (cf. Section 2.2) was applied for triggering the market exit. The prototype was trained with data between the end of July 2001 and mid-January 2002 and tested with data of the following two months. It used 10-minute intraday data for the members of the Dow Jones index and achieved a performance of 10 bps per roundtrip.

2.4 Prototype Developed by Fung/Lam/Yu

This prototype was developed around 2002 at the Department of Systems Engineering and Engineering Management of the Chinese University of Hong Kong (Fung et al. 2002; Fung et al. 2003). The universe consists of 614 stocks listed at the Hong Kong Stock Exchange and the goal is to forecast price trends after publication of news. The documents are labeled similarly as described in Section 2.2. In a first step the price time series around the publication of a news article was segmented into time windows with longest possible monotonic price increase/decrease. In the next step they used a clustering algorithm to divide the sample of time windows into the three most discriminating clusters. The cluster in which the time windows showed the steepest positive (negative) average slope was named "Rise" ("Drop"); the third one was called "Steady". Preprocessing of the news articles was performed with IBM's Intelligent Miner for Text and the SVM^{light} software from the University of Dortmund was used as classifier.

Unfortunately essential information to understand the simulation is missing. For instance, the selection criteria for determining the 614 stocks from the Hong Kong Stock Exchange are not explained (the number of stocks traded on this exchange has been well above this number since 1998). In a graph showing the cumulative profit obtained for various parameter settings the y-axis lacks a scale, leaving the reader in the dark about the profits achieved.

2.5 Prototype Developed by Mittermayer/Knolmayer

NewsCATS (News Categorization and Trading System) is a prototype developed at the Institute of Information Systems of the University of Bern since 2002 (Mittermayer 2004; Mittermayer 2005; Mittermayer and Knolmayer 2006). The system forecasts the short-term stock price trend following the publication of press releases in the U.S.

The main differences of the prototype's present version from the systems described above are:

- The Feature Selection phase is extended by an additional step in which features from a handcrafted thesaurus were added to the set of features. The thesaurus contains words, phrases, and tuples of words/phrases assumed to influence market prices of securities. Only press releases of publicly traded companies are used as categorization objects. Articles from editorial newswires are neglected since they typically do not contain new information.
- A heuristic was developed to separate types of press releases that, in the past, led to higher volatility from other, less relevant news. The rationale behind this procedure is to disburden the learning algorithm of irrelevant information.
- NewsCATS allows choosing amongst several categorization algorithms.
- A more sophisticated and more conservative exit strategy is applied which also triggers stop-loss trades.
- Stock prices with a temporal granularity of only 15 seconds are used. Thus, News¬CATS handles high-frequency data to allow more realistic performance evaluations (Mittermayer and Knolmayer 2006).

To be labeled with "Buy" ("Short"), a press release must lead to an increase (decrease) of the stock price of at least 3% during the 15 minutes following the publication. A fourth category "Unclear" is defined but neglected in the learning phase. Each press release is represented in a vector space with the 85 most important features, selected by CTF. The classifier used is the polynomial variant of SVM. In the performance simulation a stock is bought (sold short) for 15 minutes if NewsCATS assigns an incoming press release to the category "Buy" ("Short"). The simulation yields a profit per roundtrip of 27 bps. If the simulation is performed with an asymmetric exit strategy (which takes profits greater than 0.5% and losses greater than 2%) the performance increases to 29 bps per roundtrip. This is a remarkable improvement compared to the performance achieved by all other prototypes.

3. A FRAMEWORK FOR COMPARING THE PROTOTYPES

Table 1 summarizes the properties of the systems described above. It is organized in four sections: The first section provides an outline idea about each prototype, the second section details the parameter settings for the techniques used, the third section summarizes the data used for training, and the final section gives an overview of the major performance figures reported.

All prototypes considered predict only the trend and not the price level itself. 3-category model and Naïve Bayes approaches are most commonly used. In contrast to the vast majority of papers in automated text categorization, the feature definition in some of the prototypes is performed manually. Labeling is almost always done automatically which again differs from classical text categorization. The features are mainly bags of words or tuples of single words combined with the logical operator AND.

Most of the systems use daily closing data to measure the impact of news on prices. However, it is questionable whether one can capture market reactions following the publication of news with an hourly or even daily resolution. Many financial performances obtained are rather moderate. This holds especially because none of the prototypes consider any costs in the performance simulation. Transaction costs are typically much smaller for institutional investors than for retail clients; however, U.S. discount e-brokers in particular have started to offer small transaction costs for the public also. In addition to transaction costs the systems also have to cover costs of immediate execution (the bid/ask spread) and may be even indirect costs for effects of illiquidity (e.g., limited volume at the bid/ask price). For prototypes that achieve a gross profit of 10-15 bps it is likely that, net of all costs, their return boils down to zero. Interest should therefore focus on those prototypes that perform above this critical threshold.

Table 1. Comparison of Main Properties of the Prototypes described in Section 2

	Prototype 2.1.	Prototype 2.2.	Prototype 2.3..	Prototype 2.4.	Prototype 2.5.
Prototype idea					
Aims to forecast…	price trends	price trends	price trends	price trends	price trends
Underlying	equity index	single stock	single stock	single stock	single stock
Forecasting horizon	24 hours	1 hour	1 hour	1 hour	15 minutes
Text mining parameter					
Feature definition	manually	automated	automated	automated	semi-automated
Number of features	423	N/A	1000	N/A	85
Feature granularity	tuple (words)	terms	single words	single words	tuple (terms)
Primary classifier	Naïve Bayes	Naïve Bayes	Naïve Bayes	linear SVM	polynomial SVM
Number of categories	3	5	3	5 (training: 3)	4 (training: 3)
Input data					
Information age	2 - 15 hours	0 hours	0 hours	0 hours	0 hours
Text analyzed	headline, body	headline, body	headline, body	headline, body	headline, body
Labeling	automated	automated	automated	automated	Automated
Price frequency	daily close	10 min.	10 min.	intraday	15 sec.
Test					
Period investigated	1997 - 1998	1999 - 2000	2001 - 2002	2002 - 2003	2002
Training/Test split	3 months rolling	3 / 1.5 months	5.5 / 2 months	6 / 1 month(s)	cross validation (90% / 10%)
Prototype vs. random	44% vs. 33%	N/A	40% vs. 33%	N/A	45% vs. 33%
Roundtrips per year	< 600	> 100'000	< 6000	N/A	< 500
Profit per roundtrip as reported	13 bps	23 bps	10 bps	N/A	29 bps
Market	DJIA, Nikkei, FTSE, HS, ST	127 stocks (USA)	constituents DJIA	614 stocks (Hong Kong)	constituents S&P500

4. CONCLUSION

This paper summarizes and compares prototypes developed for predicting the market response to news by text mining techniques. Most of the prototypes forecast price trends, and these systems are compared in some detail. Other systems aim to predict volatilities; however, the strategies for using this information are less convincing.

The prototypes differ particularly in their labeling procedures and the details of the data (like source, market, age, and frequency of the data). Some prototypes also use a list of features handcrafted by experts to restrict the domain vocabulary. Unfortunately these lists are not available to the public.

In the performance studies, the success of the systems is typically measured by looking at the basis points per roundtrip; in some cases we were able to determine this indicator retrospectively. More technical performance criteria like F1 (the harmonic mean of macro-averaged precision and macro-averaged recall) or overall accuracy α, the percentages of correct predictions, are often missing. However, is seems that the NewsCATS system provides superior financial performance results compared to all previously developed prototypes (Mittermayer and Knolmayer, 2006).

The performance studies neglect some important features of financial markets. Low granularity of data and a practically infeasible selection of the universe of stocks may be some reasons for too favorable results. Such weaknesses should be avoided in future developments and comparisons of forecasting systems.

REFERENCES

Cho, V., 1999. *Knowledge Discovery from Distributed and Textual Data. Dissertation Hong Kong University of Science and Technology.* Hong Kong.

Cho, V. et al, 1999. Text Processing for Classification. *In Journal of Computational Intelligence in Finance*, Vol. 7, No. 2, pp. 6-22.

Fung, G.P.C. et al, 2002. News Sensitive Stock Trend Prediction. *Proceedings 6th Pacific-Asia Conference on Knowledge Discovery and Data Mining,* Taipei, pp. 481-493.

Fung, G.P.C. et al, 2003. Stock Prediction: Integrating Text Mining Approach Using Real-time News. *Proceedings IEEE Int. Conference on Computational Intelligence for Financial Engineering.* Hong Kong, pp. 395-402.

Gidófalvi, G., 2001. *Using News Articles to Predict Stock Price Movements. Project Report*. Department of Computer Science and Engineering, University of California, San Diego.

http://www-cse.ucsd.edu/users/elkan/254spring01/gidofalvirep.pdf, 2001-06-15

Gidófalvi, G. and Elkan, C., 2003. *Using News Articles to Predict Stock Price Movements. Technical Report*. Department of Computer Science and Engineering. University of California, San Diego.

http://www.cs.aau.dk/~gyg/docs/financial-prediction-TR.pdf, 2003-03-26

Lavrenko, V. et al, 1999. *Ænalyst – Electronic Analyst of Stock Behavior. Project Proposal 791m.* Department of Computer Science, University of Massachusetts, Amherst.

http://ciir.cs.umass.edu/~lavrenko/aenalyst/pitch.pdf, 1999-10-23

Lavrenko, V. et al, 2000a. Mining of Concurrent Text and Time Series. *Proceedings 6th ACM SIGKDD Int. Conference on Knowledge Discovery and Data Mining.* Boston, pp. 37-44.

Lavrenko, V. et al, 2000b. Language Models for Financial News Recommendation. *Proceedings 9th Int. Conference on Information and Knowledge Management.* Washington, pp. 389-396.

Mittermayer, M.-A., 2004. Forecasting Intraday Stock Price Trends with Text Mining Techniques. *Proceedings 37th Annual Hawaii Int. Conference on System Sciences (HICSS).* Big Island, p. 64.

Mittermayer, M.-A., 2005. *Einsatz von Text Mining zur Prognose kurzfristiger Trends von Aktienkursen nach der Publikation von Unternehmensnachrichten. Dissertation.* University of Bern 2005; dissertation.de, Berlin 2006.

Mittermayer, M.-A. and Knolmayer, G.F., 2006. NewsCATS: A News Categorization And Trading System*. Proceedings of the 6th IEEE International Conference on Data Mining, Hong Kong, pp. 1002-1007.*

Peramunetilleke, D. and Wong, R.K., 2002. Currency Exchange Rate Forecasting from News Headlines. *Proceedings 13th Australasian Database Conference.* Melbourne, pp. 131-139.

Schulz, A. et al, 2003. Kursrelevanzprognose von Ad-hoc-Meldungen: Text Mining wider die Informationsüberlastung im Mobile Banking. *In Uhr, W., Esswein, W., Schoop, E. (eds.): Wirtschaftsinformatik*. Physica, Heidelberg, pp. 181-200.

Seo, Y. et al, 2004. *Financial News Analysis for Intelligent Portfolio Management. Technical Report CMU-RI-TR-04-04.* Robotics Institute, Carnegie Mellon University, Pittsburgh.

http://www.ri.cmu.edu/pub_files/pub4/seo_young_woo_2004_2/seo_young_woo_2004_2.pdf, 2004-01-07

Wüthrich, B. et al, 1998. Daily Prediction of Major Stock Indices from Textual WWW Data. *Proceedings 4th ACM SIGKDD Int. Conference on Knowledge Discovery and Data Mining.* New York, p. 364-368.