Institute of Information Systems

University of Bern

Working Paper No 184

Text Mining Systems for Market Response to News: A Survey

Marc-André Mittermayer

Gerhard F. Knolmayer

August 2006

Address: Engehaldenstrasse 8, CH-3012 Bern, Switzerland
Tel.: ++41 (0)31 631 38 09
Fax: ++41 (0)31 631 46 82
E-Mail: gerhard.knolmayer@iwi.unibe.ch

Text Mining Systems for Predicting Market Response to News: A Survey

Abstract

Several prototypes for predicting the short-term market reaction to news based on text mining techniques have been developed. However, no detailed comparison of the systems and their performances is available thus far. This paper describes the main systems developed 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 performance results cannot be achieved in reality. Furthermore, the adequacy of applying text mining techniques for predicting stock price movements in general and approaches for dealing with existing problems are discussed.

1 Introduction

The use of data mining techniques to predict financial markets has been extensively studied in numerous publications. Most of the studies use structured data like past prices, historical earnings, or dividends. Text mining approaches are comparatively rare due to the difficulty of extracting relevant information from unstructured data.

Wüthrich et al. were presumably the first to build a prototype using text mining techniques to predict stock markets by analyzing financial news articles [WCLP98]. In the meantime additional prototypes were developed, often without recognizing already existing systems. Thus, there is a need for a survey summarizing and comparing the existing prototypes as a basis for future research and applications.

The first goal of this paper is to briefly describe the prototypes developed thus far. Since all prototypes use different text mining techniques or distinct data sets, their comparison is not straightforward. Therefore, the second aim is to develop a framework which is helpful in comparing the prototypes described earlier. A third goal is to discuss the adequacy of text min-

ing, especially of automated text categorization, to predict stock price movements. The paper helps to see how the prototypes differ and in which direction they emerge.

2 Text Mining Approaches for Analyzing News Articles

Text mining is a research area positioned at the intersection of information retrieval, data mining, natural language processing, and machine learning. We subsume knowledge-detecting as well as knowledge-extracting techniques under the term text mining. Following this definition text mining embraces, among others, text preprocessing [Lewi92, Seba02, WIZD05], automated language detection of texts [MuSp97, Sull01], summarizing or abstracting texts [GeHK01, Sull01]], automated text categorization [Seba02, WIZD05], text clustering [JaMF99, WIZD05], extended information retrieval [BaRi99, Ferb03], and visualization of texts [LiSM91].

The forecasting prototypes discussed in Section 3 apply automated text categorization (which, of course, requires text preprocessing as a preliminary step). During a learning phase the algorithms capture the structures inherent in precategorized sample documents. This results in classifiers used in the operational phase to categorize other documents. Fig. 1 depicts this process in more detail. Often standard document collections with pre-categorized documents like Reuters 22173 (21578), RCV1 or OHSUMED are split into a training set used in the learning phase and a test set applied in the operational phase, particularly for evaluating the performance of the classifier. In these cases the labeling step is omitted. No standard document collection is available for the type of systems described in Section 3; therefore a labeling step had to be performed. Each of the systems uses a different data set, making a fair comparison quite challenging. Furthermore, replication studies applying identical data sets are inhibited because the codes of the systems are not available.

3 Prototypes and Reported Performances

To the best of our knowledge, eight prototypes exist today which are described in the following subsections in chronological order of their first appearance. A more detailed discussion of the systems is given by Mittermayer [Mitt05].

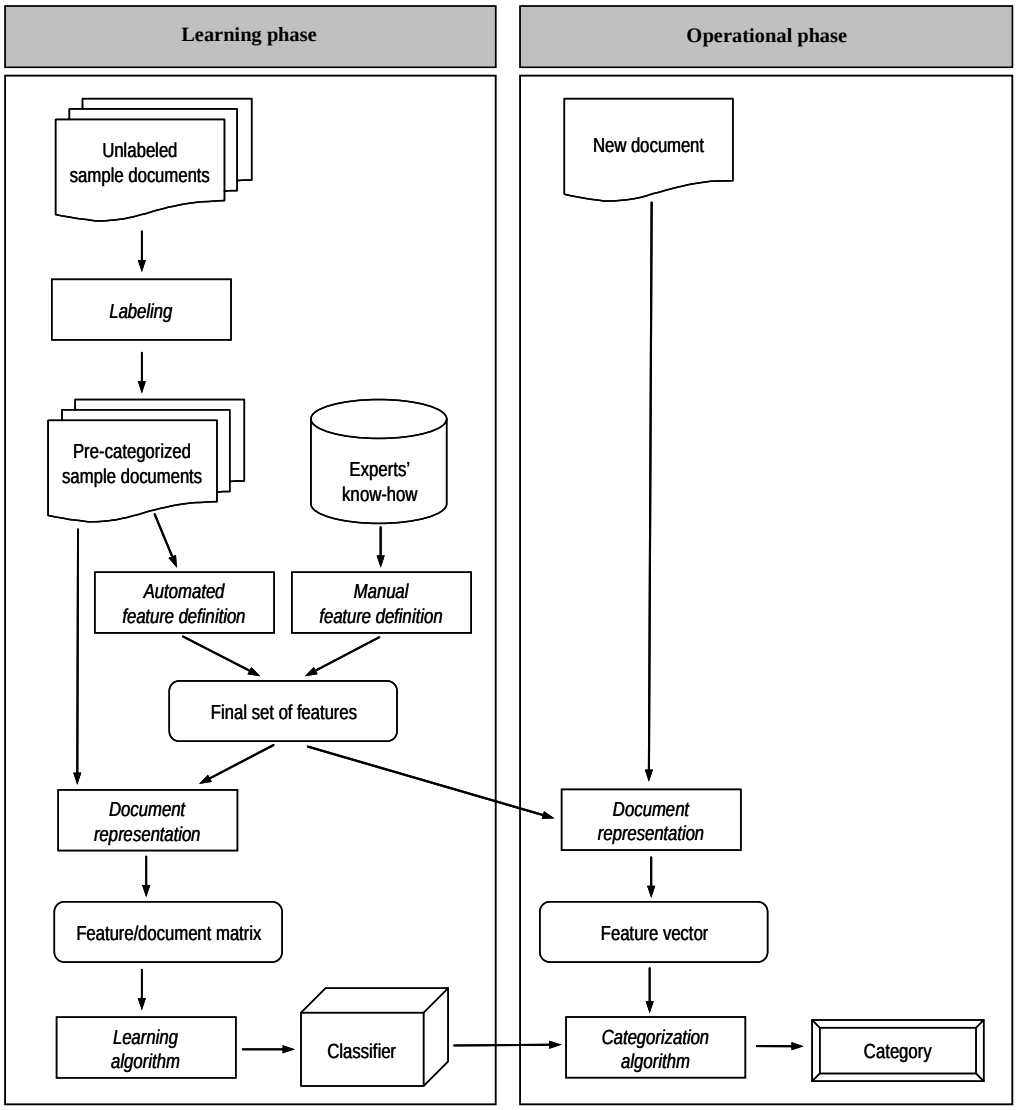

Fig. 1. Procedure Model for Automated Text Categorization.

3.1 Prototype developed by Wüthrich et al.

This prototype attempts to forecast the 1-day trend of five major equity indices, the Dow Jones, the Nikkei, the FTSE, the Hang Seng, and the Straits Times at 7:45 a.m. Hong Kong Time [Cho99, ChWZ99, WCLP98]. The daily forecast is based on news articles published overnight on web portals of, e.g., the Financial Times, Reuters, or the Wall Street Journal. Many of these articles, for instance After-the-Bell Reports summarizing or commenting the events of the previous trading day, contain no new information at least for the professional investor. The documents were 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) at least 0.5%. The third category contains the remaining news articles. The thresholds of +/- 0.5% were chosen so that roughly one third of the trading sessions fell in each of the three categories.

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

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

Averaged over all five indices a profit per roundtrip of 13 basis points (bps) was reported. At first glance the results are surprising because the information contained in the news articles 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 data: The researchers assumed next day's opening prices, on average, to be identical to the closing prices of the previous trading session. This does not make sense because the prototype decides better than a random trader and so, on average, buys too low and sells too high. As a consequence, the simulated performance obtained with this prototype cannot be achieved in reality.

3.2 Prototype developed by Lavrenko et al.

The prototype Ænalyst was developed around 2000 at the University of Massachusetts Amherst [LSLO00a, LSLO00b, OgSc99]. Ænalyst aims to forecast very short-term (intraday) price trends of a subset of U.S. stocks by analyzing news articles published on the homepage of YAHOO!Finance. One difference to the prototype described above is that there is no manual feature definition. The features were determined automatically using TFxIDF (Term Frequency times Inverse Document Frequency) as feature selection technique. A 5-category model with 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 ≥ 0.75 were put into the category "Surge". Articles leading to a slope between 0.5 to 0.75 were assigned to the category "Slight+"; the other categories were determined accordingly. The authors claim that a parameter value h between 5 and 10 hours provides best results.

The classifier was trained with a Naïve Bayes approach. During the operational phase the prototype triggered a buy recommendation if an incoming news article was assigned to the categories "Surge" or "Slight+". On the other hand, if a news article fell into the "Slight-" or "Plunge" category, the prototype triggered a short recommendation. Based on these recommendations, the authors performed virtual roundtrips in the U.S. stock market. They assumed one could enter the stock market at the time the news appeared. The market was exit once 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 asymmetric exit strategy.

The prototype was tested in a simulation based on 10-minute stock price data between mid-March and April 2000. In each roundtrip USD 10,000 are invested. After a testing period of 40 days a result of 280,000 USD is achieved by performing about 12,000 transactions, resulting in a profit per roundtrip of 23 bps. The authors are humble when they regard this figure a "very modest gain" [LSLO00b]. 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 that showed the largest positive or negative price movements in the period under investigation. 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 in 40 days 127 stocks generate 12,000 different news that trigger a buy or short recommendation. Maybe certain events were reported in several articles and the system reacted to all of them. The authors also assumed that they may obtain unlimited funds for trading. Of course, institutional investors may leverage their investment by borrowing money and, thus, investing 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 larger 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.

3.3 Prototype developed by Thomas et al.

This prototype was developed at the Robotics Institute of Carnegie Mellon University between 2000 and 2004 [SeGS02, SeGS04, Thom03, ThSy00]. During this period the setup as well as the goal of the prototype changed remarkably. In the beginning the system should predict stock price trends but the results were rather poor. Later it focused on forecasting volatility. The authors developed a rather surprising strategy in which they temporarily exit the market for a particular stock once news are published that may increase volatility. The decision to re-enter the market is based on technical indicators. The result of such a strategy is an improved return/risk profile since the long-term return of a stock is accompanied by lower volatility.

The current version of the prototype consists of a rule-based classifier for 39 categories each representing a specific type of news, e.g., acquisition, earnings outlook, or stock split. The classifier was handcrafted and, thus, in a strict sense has little in common with text mining. The classifier of the category "Lawsuit", for instance, contains the following expressions:

- "class action[s] OR class period[s] OR arbitration OR su[es|ed|ing]"
- "patent[s] AND (dispute OR settle[d] OR suit OR protect)"
- "court battle OR restraining order OR litigation OR injunction OR lawsuit[s]"
- "(fil[e|es|ing] OR throw[s] out OR launch[es] OR dismiss[es] OR toss[es] OR win[s] OR lose[s] OR announce[s]) AND [law]suit NAND class"
- "seek damages".

If for a certain news article at least one of the expressions holds, it is assigned to the category "Lawsuit". Multicategory assignments are conceivable but the authors do not state how they deal with them. The prototype was tested with the members of the Russell 3000 index during the period December 2001 to April 2002. Important information regarding the simulation is missing. For instance, the trading rule for the category "Lawsuit" is

<< IF [News Article ∈ "Lawsuit"] THEN [Stay out of the market for this stock for 15 days] >>

No information is given at which point in time such a stock is sold: At the closing of the previous session or of the session which contained the triggering event? An immediate reaction cannot be implemented since the authors operate with daily data and the system will not be able to circumvent the short-term volatility resulting from news if the stock is sold after the triggering event.

3.4 Prototype developed by Elkan/Gidófalvi

Another prototype aiming to forecast stock price trends [Gidó91, GiEl03] 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 at least to a price increase (decrease) of 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. 1000 words with highest MI values were used as features. The authors list the first 100 words in the appendix [GiEl03]. It is surprising that most of the words do not refer to stock prices. The top five words are "sbc", "msft", "websphere", "db", and "index". Thus, the features used differ completely from those applied in the rule-based classifier described in Section 3.3. 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 3.2) was applied for triggering the market exit. The prototype was trained with data between end of July 2001 and mid-January 2002 and tested with data of the next two months. The authors used 10-minute intraday data for the members of the Dow Jones index and achieved a performance of 10 bps per roundtrip.

3.5 Prototype developed by Peramunetilleke/Wong

The prototype developed at the University of New South Wales was developed in collaboration with currency traders from UBS around 2001 [PeWo02]. This cooperation is probably the main reason why the prototype aims to forecast trends in USD/DEM and USD/JPY exchange rates. The system was tested with relatively few and quite old data from September 1993. The prototype relies like the one described in Section 3.1 on a handcrafted dictionary (in this case defined by currency experts). It contains more than 400 features, each one consisting of two to five words that are combined with the logical operator AND. Also this dictionary has not been made accessible to the public.

In the learning phase the authors create a rule-based classifier based on the three categories "dollar up >0.23%", "dollar down >0.23%", and "dollar steady" (in the 60 minutes following the news release). The threshold 0.23% was chosen to obtain on average a uniform assignment to the three categories. The authors do not provide a profit per roundtrip but only mention that the system is right in 50% of the predictions. Again, a random trader would have achieved only 33.3%.

The authors claim that professional currency traders are also right in about 50% of their decisions and, as a consequence, the prototype achieves similar decision quality. However, decision speed should be significantly improved with the system compared to a manual analysis.

3.6 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 [FuYL02, FuYL03]. 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 by a similar approach as described in Section 3.2. In a first step the authors segmented the price time series around the publication of a news article 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".

The authors were among the first who partially used commercial text mining software instead of programming a prototype completely on their own. For instance, the 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). The authors also provide a graph showing the cumulative profit obtained for various parameter settings; however, the y-axis lacks a scale, leaving the reader in the dark about the size of the profit.

3.7 Prototype developed by Schulz/Spiliopoulou/Winkler

This prototype was collaboratively created by several institutes of German universities around 2002 [ScSW03, SpSW03]. The authors of the prototype were again using commercial text mining software, in this case SAS Enterprise Miner, to fulfill some of the tasks. The goal is to predict volatility after the publication of press releases of German public companies. It is not attempted to make profits but only to sort press releases into price relevant (those leading to higher volatility) and price irrelevant news. As a consequence the authors do not calculate profits. The advantage for an investor using this system is that he can focus on potentially price relevant news and skip others, reducing the information overload.

The labeling of the press releases was performed as follows: For each stock a regression according to the capital asset pricing model was calculated, based on daily data. If the performance of

the stock on the publication day of a press release was located outside the 90% confidence interval, the news was regarded as leading to increased volatility. The system was tested with data from 1999 to 2002. In 61% of the cases the prototype categorizes correctly. On a first glance this looks like a good result since a random categorization would result only in 50%. However, the prototype achieves correct assignments with 68% for price irrelevant news but only with 43% for price relevant press releases. Since the goal of the prototype is to identify price relevant press releases, the authors qualify the results as unsatisfactory.

3.8 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 [Mitt04; Mitt05; MiKn06]. The system forecasts the short-term stock price trend following the publication of press releases in the US. Articles from editorial newswires (like Reuters or Dow Jones) are neglected since they typically do not contain new information.

Main differences of the prototype's present version from the previously described ones 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 (and not all types of news articles) are used as categorization objects.
- A heuristic was developed to separate types of press releases that, in the past, led to higher volatility from others, less relevant news. The rationale behind this procedure is to disburden the learning algorithm from irrelevant information.
- Stock prices are used with a temporal granularity of only 15 seconds. Thus, NewsCATS handles high-frequency data to allow more realistic performance evaluations.
- NewsCATS allows choosing among several categorization algorithms.
- A more sophisticated and more conservative exit strategy is applied which triggers also stop-loss trades [MiKn06].

Similar to other systems, NewsCATS uses a 3-category model with categories "Buy", "Short", and "No Recommendation". 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 in order to prevent the learning algorithm from being confused. Each press release is represented by a vector 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.

More detailed descriptions of the present version of NewsCATS are given in [Mitt05; MiKn06].

4 Properties and Comparison of the Prototypes

Table 1 summarizes the properties of the systems described above. It is organized in four sections: The first section provides a rough 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.

Most of the prototypes aim to forecast price trends (especially of stock prices). All of them predict only the trend and not the price level itself. A 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.

	Prototype 3.1.	Prototype 3.2.	Prototype 3.3.	Prototype 3.4.	Prototype 3.5.	Prototype 3.6.	Prototype 3.7.	Prototype 3.8.
Prototype idea								
Aims to forecast…	price trends	price trends	volatilities	price trends	price trends	price trends	volatilities	price trends
Underlying	equity index	single stock	single stock	single stock	exchange rate	single stock	single stock	single stock
Forecasting horizon	24 hours	1 hour	N/A	1 hour	3 hours	1 hour	N/A	15 minutes
Text mining parameter								
Feature definition	manually	automated	manually	automated	manually	automated	automated	semi-automated
Number of features	423	N/A	145	1000	400	N/A	200	85
Feature granularity	tuple (words)	terms	tuple (terms)	single words	tuple (words)	single words	single words	tuple (terms)
Primary classifier	Naïve Bayes	Naïve Bayes	decision rules	Naïve Bayes	decision rules	linear SVM	regression	polynomial SVM
Number of categories	3	5	39	3	3	5 (training: 3)	2	4 (training: 3)
Input data								
Information age	2 - 15 hours	0 hours	0 - 24 hours	0 hours	0 - 2 hours	0 hours	0 hours	0 hours
Text analyzed	headline, body	headline, body	headline	headline, body	headline	headline, body	headline, body	headline, body
Labeling	automated	automated	manually	automated	automated	automated	automated	automated
Price frequency	daily close	10 min.	daily close	10 min.	60 min.	intraday	daily close	15 sec.
Test								
Period investigated	1997 - 1998	1999 - 2000	2001 - 2002	2001 - 2002	1993	2002 - 2003	1999 - 2002	2002
Training/Test split	3 months rolling	3 / 1.5 months	8 / 5 months	5.5 / 2 months	1 month rolling	6 / 1 month(s)	cross validation (90% / 10%)	cross validation (90% / 10%)
Prototype vs. random	44% vs. 33%	N/A	N/A	40% vs. 33%	50% vs. 33%	N/A	61% vs. 50%	45% vs. 33%
Roundtrips per year	< 600	> 100'000	(200)	< 6000	N/A	N/A	N/A	< 500
Profit per roundtrip as reported	13 bps	23 bps	(first phase: 10 bps)	10 bps	N/A	N/A	N/A	29 bps
Market	DJIA, Nikkei, FTSE, HS, ST	127 stocks (USA)	constituents Russell 3000	constituents DJIA	USD/DEM and USD/JPY	614 stocks (Hong Kong)	constituents DAX100	constituents S&P500

Table 1. Comparison of Main Properties of the Prototypes.

Apart from the admittedly interesting concepts of the prototypes some deficiencies exist. Most of them use daily closing data to measure the impact of news on prices. However, it is questionable whether one is able to capture market reactions following the publication of news with an hourly or even daily resolution. Only in three systems the resolution is lower than 1 hour. Most of the prototypes use only a few months of data to train the classifier and/or to test the system. This is comprehensible for prototypes using intraday data but hard to understand for systems using daily data, since the volume of daily data is fairly easy to handle.

Most of the financial performances obtained are rather moderate. This holds especially because none of the prototypes considers any costs in the performance simulation. 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). Since most of the prototypes achieve a gross profit of 10-15 bps it is likely that net of all costs their return boils down to zero.

5 Adequacy of Text Mining to Predict Market Response to News

From a finance perspective it is difficult to argue that stock price movements can be predicted without considering market expectation. The problem with using plain-vanilla text mining is that one cannot account for it. If we develop a model which includes market expectation, we end up using a rule-based model which has little in common with the basic idea of text mining. In addition to that, news often contains price-relevant information as numbers and text mining cannot put them in an economic perspective. Thus, the question of whether text mining is the right way to predict market response to news is fairly adequate.

However, there are several types of news where the dominant information is coded in an unstructured format and/or the market's expectation is of little relevance. For instance, if a company reports that it received a takeover bid, the crucial data is the unstructured information of the bid. One could argue that also in this case the bid price impacts the stock price movement. But generally, a premium has to be paid in a takeover and therefore takeovers typically lead to an increase of the stock price of the takeover candidate. A text mining system could therefore concentrate on news without important numerical information by, for instance, neglecting earning news.

Another problem in using text mining techniques for predicting stock price movements is the fact that they measure the importance of the words finally used to represent the training data (i.e. the features) through selection statistics like IDF, MI, or information gain. These statistics assume that the importance

of words is either a function of their occurrence in the document collection or the specific category. But none of these assumptions is necessarily true when the importance of words for predicting the category of the investigated document has to be determined.

Let us regard a company stating that current earnings per share are higher compared to the last quarter and also higher compared to the same quarter last year. However, next quarter earnings are expected to be lower than declared in a previously published guidance. In such a profit warning all information is in plain text, but most of it is irrelevant. It is not too important that the company increased earnings per share in this quarter. More important is the fact that in the future it assumes to end up below earlier forecasts. So, in this news article the rarest word ("lower") in the word set {"higher", "lower"} is dominant for deciding whether it is good or bad news. Now assume that the earnings per share are also lower than in the last quarter, then "lower" is the most frequent word, inverting the logic completely. It is therefore important to analyze words always in the context of the surrounding words. The word "higher" next to the string "last quarter" might not be relevant if the word "lower" is followed closely by "than predicted". This problem can only be solved if document preprocessing is adjusted. One option would be to create a thesaurus which contains words and phrases recognized as being capable of moving stock prices. During the feature selection, words could be forced into the final set if they appear in any of the training documents (no matter how often); for instance, the term "delisting" will occur very rarely in news articles but if it exists it may be the catalyst for a sharp price drop. The advantage of such a concept is that the feature selection still happens automatically for large parts of the feature set but is overruled if an informative word would be overlooked. This concept has been successfully applied in the redesigned version of NewsCATS [Mitt05; MiKn06].

6 Summary and Outlook

This paper summarizes and compares the prototypes developed for predicting the market response to news by text mining techniques. Most of the prototypes forecast price trends, in particular trends of stock prices, exchange rates, and equity indices. Typically three categories are distinguished: A category with positive news, one with negative, and one with news without impact on prices. 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 hand-crafted 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 percentage of correct predictions, are often missing. Furthermore, the performance studies neglect some important features of the financial markets like transaction costs, bid/ask spreads, and limited volume at given prices. Sometimes the low granularity of data or a practically infeasible selection of the universe of stocks is the reason for too favorable results. Such weaknesses should be avoided in future developments of forecasting systems.

Literature

[BaRi99] Baeza-Yates, R.; Ribeiro-Neto, B.: Modern Information Retrieval. ACM Press, New York 1999.

[Cho99] Cho, V.: Knowledge Discovery from Distributed and Textual Data. Dissertation Hong Kong University of Science and Technology. Hong Kong 1999.

[ChWZ99] Cho, V.; Wüthrich, B.; Zhang, J.: Text Processing for Classification. In: Journal of Computational Intelligence in Finance 7 (1999) 2, pp. 6-22.

[Ferb03] Ferber, R.: Information Retrieval - Suchmodelle und Data-Mining-Verfahren für Textsammlungen und das Web. dpunkt.verlag, Heidelberg 2003.

[FuYL02] Fung, G.P.C.; Yu, J.X.; Lam, W.: News Sensitive Stock Trend Prediction. In: Proceedings 6th Pacific-Asia Conference on Knowledge Discovery and Data Mining. Taipei 2002, pp. 481-493.

[FuYL03] Fung, G.P.C.; Yu, J.X.; Lam, W.: Stock Prediction: Integrating Text Mining Approach Using Real-time News. In: Proceedings IEEE Int. Conference on Computational Intelligence for Financial Engineering. Hong Kong 2003, pp. 395-402.

[GeHK01] Gerstl, P.; Hertweck, M.; Kuhn, B.: Text Mining: Grundlagen, Verfahren und Anwendungen. In: HMD - Praxis der Wirtschaftsinformatik 39 2001, pp. 38-48.

[Gidó01] Gidófalvi, G.: 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.

[GiEl03] Gidófalvi, G.; Elkan, C.: 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.

[JaMF99] Jain, A.K.; Murty, M.N.; Flynn, P.J.: Data Clustering: A Review. In: ACM Computing Surveys 31 (1999) 3, pp. 264-323.

[Lewi92] Lewis, D.D.: Representation and Learning in Information Retrieval. Dissertation University of Massachusetts. Amherst 1992.

[LiSM91] Lin, X.; Soergel, D.; Manchionini, G.: A Self-organizing Semantic Map for Information Retrieval. In: Proceedings 14th Annual Int. ACM SIGIR Conference on Research and Development in Information Retrieval. Chicago 1991, pp. 262-269.

[LSLO00a] Lavrenko, V.; Schmill, M.; Lawrie, D.; Ogilvie, P.; Jensen, D.; Allan, J.: Mining of Concurrent Text and Time Series. In: Proceedings 6th ACM SIGKDD Int. Conference on Knowledge Discovery and Data Mining. Boston 2000, pp. 37-44.

[LSLO00b] Lavrenko, V.; Schmill, M.; Lawrie, D.; Ogilvie, P.; Jensen, D.; Allan, J.: Language Models for Financial News Recommendation. In: Proceedings 9th Int. Conference on Information and Knowledge Management. Washington 2000, pp. 389-396.

[MiKn06] Mittermayer, M.-A.; Knolmayer, G.F.: NewsCATS: A News Categorization And Trading System. In: Proceedings of the International Conference in Data Mining (ICDM06), Hong Kong 2006, in print.

[Mitt04] Mittermayer, M.-A.: Forecasting Intraday Stock Price Trends with Text Mining Techniques. In: Proceedings 37th Annual Hawaii Int. Conference on System Sciences (HICSS). Big Island 2004, p. 64.

[Mitt05] Mittermayer, M.-A.: Einsatz von Text Mining zur Prognose kurzfristiger Trends von Aktienkursen nach der Publikation von Unternehmensnachrichten. Dissertation University of Bern. Bern 2005.

[MuSp97] Muthusamy, Y.K.; Spitz, A.L.: Automatic Language Identification. In: Cole et al. (eds.): Survey of the State of the Art in Human Language Technology. Cambridge University Press, Cambridge 1997, pp. 273-276.

[OgSc99] Ogilvie, P.; Schmill, M.: 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.

[PeWo02] Peramunetilleke, D.; Wong, R.K.: Currency Exchange Rate Forecasting from News Headlines. In: Proceedings 13th Australasian Database Conference. Melbourne 2002, pp.131-139.

[Seba02] Sebastiani, F.: Machine Learning in Automated Text Categorization. In: ACM Computing Surveys 34 (2002) 1, pp. 1-47.

[ScSW03] Schulz, A.; Spiliopoulou, M.; Winkler, K.: Kursrelevanzprognose von Ad-hoc-Meldungen: Text Mining wider die Informationsüberlastung im Mobile Banking. In: Uhr, W., Esswein, W., Schoop, E. (eds.): Wirtschaftsinformatik 2003. Physica, Heidelberg 2003, pp. 181-200.

[SeGS02] Seo, Y.; Giampapa, J.A.; Sycara, K.: Text Classification for Intelligent Portfolio Management. Technical Report CMU-RI-TR-02-14, Robotics Institute, Carnegie Mellon University, Pittsburgh. http://www.ri.cmu.edu/pub_files/pub4/seo_young_woo_2002_1/seo_young_woo_2002_1.pdf, 2002-05.

[SeGS04] Seo, Y.; Giampapa, J.A.; Sycara, K.: 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.

[SpSW03] Spiliopoulou, M.; Schulz, A.; Winkler, K.: Text Mining an der Börse: Einfluss von Ad-hoc-Mitteilungen auf die Kursentwicklung. In: Becker, C., Redlich, H. (eds.): Data Mining und Statistik in Hochschule und Wirtschaft. Shaker, Aachen 2003, pp. 215-228.

[Sull01] Sullivan, D.: Document Warehousing and Text Mining. Wiley, New York 2001.

[Thom03] Thomas, J.D: News and Trading Rules. Dissertation Carnegie Mellon University. Pittsburgh 2003.

[ThSy00] Thomas, J.D.; Sycara, K.: Integrating Genetic Algorithms and Text Learning for Financial Prediction. In: Proceedings GECCO-2000 Workshop on Data Mining with Evolutionary Algorithms. Las Vegas 2000, pp. 72-75.

[WIZD05] Weiss, S.M.; Indurkhya, N.; Zhang, T.; Damerau, F.J.: Text Mining – Predictive Methods for Analyzing Unstructured Information. Springer, New York 2005.

[WCLP98] Wüthrich, B.; Cho, V.; Leung, S.; Peramunetilleke, D.; Sankaran, K.: Zhang, J.; Lam, W.: Daily Prediction of Major Stock Indices from Textual WWW Data. In: Proceedings 4th ACM SIGKDD Int. Conference on Knowledge Discovery and Data Mining. New York 1998, S. 364-368.