Information Processing and Management 45 (2009) 571–583

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/infoproman

A quantitative stock prediction system based on financial news

Robert P. Schumaker [a,*], Hsinchun Chen [b]

[a] Hagan School of Business, Information Systems Department, Iona College, New Rochelle, NY 10801, USA
[b] Artificial Intelligence Lab, Department of Management Information Systems, The University of Arizona, Tucson, AZ 85721, USA

ARTICLE INFO

Article history:
Received 4 April 2008
Received in revised form 22 April 2009
Accepted 1 May 2009
Available online 29 May 2009

Keywords:
Knowledge management
Prediction from textual documents
Quantitative funds

ABSTRACT

We examine the problem of discrete stock price prediction using a synthesis of linguistic, financial and statistical techniques to create the Arizona Financial Text System (AZFinText).

The research within this paper seeks to contribute to the AZFinText system by comparing AZFinText's predictions against existing quantitative funds and human stock pricing experts. We approach this line of research using textual representation and statistical machine learning methods on financial news articles partitioned by similar industry and sector groupings. Through our research, we discovered that stocks partitioned by Sectors were most predictable in measures of Closeness, Mean Squared Error (MSE) score of 0.1954, predicted Directional Accuracy of 71.18% and a Simulated Trading return of 8.50% (compared to 5.62% for the S&P 500 index). In direct comparisons to existing market experts and quantitative mutual funds, our system's trading return of 8.50% outperformed well-known trading experts. Our system also performed well against the top 10 quantitative mutual funds of 2005, where our system would have placed fifth. When comparing AZFinText against only those quantitative funds that monitor the same securities, AZFinText had a 2% higher return than the best performing quant fund.

1. Introduction

Predicting changes in the stock market has always had a certain appeal to researchers. While numerous attempts have been made (Chan, Jegadeesh, et al., 1996; Cho, Wuthrich, et al., 1999; Gidofalvi, 2001; Mittermayer, 2004; Seo, Giampapa, et al., 2002; Wuthrich, Cho, et al., 1998; Yoon & Swales, 1991), the difficulty has always centered on the behaviors of human traders within this socially constructed system. With parameters ill-defined and constantly shifting, prediction has been difficult at best. To further create confusion, there have been two diametrically opposed philosophies of stock market research: fundamental and technical analysis techniques (Technical-Analysis, 2005).

Acquiring relevant textual data is an important facet of stock market prediction. While many reports and articles are written on a daily basis, information flowing from these sources must take the form of numeric data before existing applications can process it. This limitation forces open a temporal gap between when information is acquired to when it can be acted on. Information of an unexpected nature can cause a significant impact on share prices and the ability to make accurate predictions from these textual documents would be a useful decision-making tool. Our research helps to close this gap by creating the Arizona Financial Text System (AZFinText), which can learn from textual financial documents and make price predictions based upon the news it contains.

While there have been several studies covering textual financial predictions, these studies have been limited to classifying price direction, i.e., the stock price will increase, decrease or stay the same (Cho et al., 1999; Lavrenko, Schmill, et al., 2000b;

* Corresponding author.
E-mail addresses: rschumaker@iona.edu (R.P. Schumaker), hchen@eller.arizona.edu (H. Chen).

0306-4573/$ - see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.ipm.2009.05.001

Mittermayer, 2004). Discrete analysis, or assigning an exact price with some degree of accuracy has not been a trivial task. In regards to textual document selection, prior studies have used either all financial news articles or only those articles of a specific company (Cho et al., 1999; Gidofalvi, 2001; Lavrenko, Schmill, et al., 2000a, 2000b; Mittermayer, 2004; Wuthrich et al., 1998). These studies have neglected the investigation of in-between categories, such as Industry sectors in their analyses. This paper will examine the worth of using textual financial news articles divided by similar industries and groupings. Similarly, we will also measure the value of our system versus existing quantitative funds and hedge-fund managers.

This paper is arranged as follows. Section 2 provides an overview of literature concerning Stock market prediction, textual representations and machine learning techniques and describes our research questions. Section 3 outlines the AZFinText system. Section 4 provides an overview of our experimental design. Section 5 details our experimental findings and discusses their impact on stock market prediction. Section 6 delivers our conclusions and a brief discourse on future research directions.

2. Literature review

In predicting stock market movement, two theories have had a significant impact on market research: Efficient Market Hypothesis (EMH) and Random Walk Theory. In EMH, the price of a security is a reflection of complete market information. Whenever a change in financial outlook occurs, the market will instantly adjust the security price to reflect the new information (Fama, 1964). EMH contained three different levels of information sharing: the weak form, the semi-strong and the strong form. Within weak EMH, only historical data is embedded within the current price. The semi-strong form goes farther by incorporating historical and current public information into its prices. The strong form includes historical and current public information as well as private information. From these three forms, it was believed that markets behaved efficiently and that instantaneous price corrections would obviate prediction models.

Random Walk Theory is slightly different in its theoretical underpinnings by focusing on an overall short-term random pattern of stock market movements (Malkiel, 1973). This random activity is believed to produce unpredictable prices and makes it impossible to consistently outperform the market. This view is similar to the semi-strong EMH model where all information is contained within the current price and is worthless for future prediction.

2.1. Fundamentalists and technicians

While EMH and Random Walk have served to discourage forecasting activity, the traders that persisted formed two distinctly different viewpoints on market prediction: fundamental and technical analysis. Fundamentalists are interested in the internal makeup of a security, such as numeric data on the overall economy, individual stock health ratios such as inflation, interest rates, return on assets, debt to equity and price to earnings among others. The focus of the fundamental trader is to make predictions from the current set of numeric data. As a consequence, historical or time-series data is not considered. By contrast, technicians rely heavily on time-series data and believe that market timing is crucial. Figures such as volume, volatility, support/resistance levels and charting techniques are all within the repertoires of a Technicians toolbox. However, technical analysis is considered to be more of an art form and is subject to interpretation.

In a study comparing the merits of fundamental and technical trading strategies, LeBaron created an artificial stock market with simulated traders (LeBaron, Arthur, et al., 1999). He introduced new pieces of information into the market and varied the amount of time between when an individual trader would receive information and act upon it. It was found that traders with longer-period waiting times formed fundamental strategies while those with shorter-period waits developed technical strategies. This study was more important from the standpoint that a lag was discovered between the time that information was introduced to when the market corrected itself. This apparent delay in market behavior helped to dispel the instantaneous correction notions of EMH and lent support of a weak ability to forecast the market. Subsequent research into this weak forecasting ability (Shmilovici, Alon-Brimer, et al., 2003) led to the discovery of a 20 min window of opportunity before and after a financial news article is released (Gidofalvi, 2001). Within this window, weak prediction of stock price direction was found to be possible.

2.2. Financial news articles

The extent of prediction between financial news articles and their impact on stock market prices is a complex avenue to investigate. While the information contained in financial news articles can have a visible impact on a security's price (Gidofalvi, 2001; Lavrenko et al., 2000a; Mittermayer, 2004; Wuthrich et al., 1998), sudden price movements can still occur from large unexpected trades (Camerer & Weigelt, 1991).

The first challenge of textual financial prediction is to process the large amounts of textual information which exist for securities. This material not only includes required reports such as periodic SEC filings, but also a wealth of financial news articles reporting unexpected events and routine news alike. Financial news articles can be automatically capitalized on by using Natural Language Processing (NLP) and text-processing techniques to identify specific terms which can lead to dramatic share price changes. This method can be repeatedly used to forecast price fluctuations and take advantage of arbitrage opportunities faster than human counterparts.

The means of obtaining timely financial news articles can come from a variety of Internet sources. One source is Comtex which offers real-time financial news in a subscription format. Another source is PRNewsWire which offers free real-time and subscription-based services. By contrast, Yahoo Finance is a compilation of 45 different news sources including the Associated Press, Financial Times and PRNewsWire among others. This source provides a variety of perspectives and timely news stories regarding financial markets.

2.3. Textual representation

Once news articles have been collected they must be represented. One technique is to use a *Bag of Words* approach which has been extensively used in textual financial research (Gidofalvi, 2001; Lavrenko et al., 2000a). This process involves the removal of meaningless stopwords such as conjunctions and declaratives from text and using the remaining terms as the textual representation. While this method has been popular, its drawbacks include noise from seldom-used terms and scalability problems where immense computational power is required for large datasets. An improved representational system which addresses these shortcomings is *Noun Phrases*. This representation focuses on retaining only the nouns and Noun Phrases present within a document and has been found to adequately represent the important article concepts (Tolle & Chen, 2000). As a consequence of its noun-centric activity, this technique uses fewer terms and can handle article scaling better than Bag of Words. A third representational technique is *Named Entities*, which extends *Noun Phrases* by selecting the article's Proper Nouns that fall within well-defined categories. This process uses a semantic lexical hierarchy (Sekine & Nobata, 2004) as well as a syntactic/semantic tagging process (McDonald, Chen, et al., 2005) to assign candidate terms to categories. Categorical definitions arise from the Message Understanding Conference (MUC-7) Information Retrieval task and encompass the entities of date, location, money, organization, percentage, person and time. This more abstract representational method allows for better generalization of previously unseen terms and does not possess the scalability problems associated with a semantics-only approach. A fourth representational technique is *Proper Nouns* which functions as an intermediary between *Noun Phrases* and *Named Entities*. This representation is a subset of *Noun Phrases* that names specific nouns while at the same time is a superset of *Named Entities* without the constraint of specific category assignment. This representation was found to be useful by removing the ambiguity associated with particular Proper Nouns which could either be represented by more than one named entity or fall outside one of the seven defined *Named Entity* categories. In a comparison study using these four representational techniques, it was found that *Proper Noun* representation was more effective in symbolizing financial news articles (Schumaker & Chen, 2006).

Simply assigning one representational mechanism is not sufficient to address the scalability issues associated with large datasets. One way to approach this problem is to introduce a threshold of term frequency (Joachims, 1998). This method uses a term frequency cut-off to represent terms that appear more frequently. It has the dual effect of eliminating noise from lesser used terms as well as reducing the number of features represented. Following this line of research, machine learning algorithms are unable to process these raw terms and require an additional layer of representation. One method is to represent the terms in binary where the term is either present or not in a given article (Joachims, 1998). This leads to sparse matrices where the number of represented terms throughout the dataset will greatly outnumber the terms used in an individual article.

Once article terms have been represented, machine learning algorithms can then be applied. One method, Support Vector Regression (SVR), a derivative of Support Vector Machines (SVM) (Vapnik, 1995), is the regression equivalent of SVM but without the aspect of classification. Like SVM, SVR attempts to minimize its fitting error while maximizing its goal function by fitting a regression estimate through a multi-dimensional hyperplane. This method is also well-suited to handling textual input as binary representations and has been used in similar financial news studies (Schumaker & Chen, 2008; Tay & Cao, 2001).

In a more specific study on the role of financial news articles to predict stock movement direction, Lavrenko tested the effects of training on the entire universe of news articles (Universal-training) and training based on articles for a specific company (Stock Specific-training) (Lavrenko et al., 2000a). From this study, it was found that training a system on Stock-Specific news articles led to more accurate predictions of price direction. It was reasoned that keywords specific to the company were more influential in determining price direction than the collection of terms from a universe of diverse sectors and industries. In a similar vein, it was also found that training a system on Universal keywords led to lower variances and more uniform predictions than the Stock-Specific counterparts. One of the limitations of this study was the absence of evaluation for in-between company groupings along similar sector and industrial pairings.

2.4. Industry classification

In order to investigate the effect of grouping similar companies together, there are several different Industry classification standards that group companies with similar outputs. The first of which is the Standard Industrial Classification (SIC) system that was developed during the 1930s as a way to categorize industrial production. Using a four digit numeric code, this system was a relatively flat structure without hierarchy that grouped similar industries together with somewhat close numeric codes. Replacing the SIC in the mid-1990s was the North American Industry Classification System (NAICS) which was designed for the industries of the United States, Canada and Mexico. This six-digit system utilized a 5-level hierarchy of sector, Sub-Sector, Industry Group, Industry and Nationality where sector used the first two digits and each successive level used an

additional digit. A similar classification system developed by Morgan Stanley specifically for global commerce was the Global Industry Classification Standard (GICS). This system, which is used by Standard and Poors, employs an eight-digit, 4-level hierarchy where each successive two digit pair indicates a deeper level in the hierarchy: sector, Industry Group, Industry and Sub-Sector. In a comparison study of various Industry classification systems, it was found that the GICS system of classification was more homogeneous and had lower variances in the calculated returns as compared to SIC and NAICS (Bhojraj, Lee, et al., 2003). In a similar study of Industry classification schemes versus analyst specialties, it was found that the GICS system best described the areas of expertise described by stock analysts (Boni & Womack, 2004). The homogeneity of company partitions was best matched against those Industry partitions followed by professional analysts.

2.5. Prediction experts and quantitative techniques

Among trading professionals and the Internet, there is no shortage of stock advice. While some of this free advice may be susceptible to bias and market manipulation, subscription-based forecasts may not necessarily be better. There has also been a recent resurgence of interest in quantitative investing where computer programs are given some degree of control over financial investments. These systems can be constrained to analyze financial data and either issue recommendations or complete trades. While these entities may be less susceptible to biases, they are vulnerable to variable movement outside of defined parameters and are unable to evaluate market intangibles (Jelveh, 2006).

2.5.1. Prediction experts

In a study between trading professionals and laypeople, both groups were given information on two stocks and were asked to select the better performer (Torngren & Montgomery, 2004). It was found that trading professionals averaged 40% accuracy, well-below chance as compared to 58% accuracy for laypeople. The study concluded that selection errors were attributable to overconfidence and the failure to weight information uncertainty. A similar study pitted trading professionals against two groups of students in pricing a series of options (Abbink & Rockenbach, 2006). This study resulted in the student groups having an 88% and 90% payoff, respectively, compared to an 80% payoff for the trading professionals. In both the laypeople and student studies, it was discovered that participants without much investing experience were employing simple heuristics to derive their answers. By contrast, trading professionals relied more on intuition than training.

As for trading professionals, there are many available and across a broad spectra of media that provide free advice and are reasonably easy to web mine. One of which is Jim Cramer, host of the CNBC Mad Money television show. Cramer is a former hedge-fund manager who boasts 24% returns over his 13 year tenure. He provides a daily list of recommendations for stocks he feels should be bought and sold. Jim Jubak, Senior Markets Editor for MSN Money, is another professional advice expert. Jubak has served as editor and contributor to several print and Internet publications with a self-reported 14% annual return on his selections. He provides a list of buy recommendations three times a week. DayTraders.com is an Internet stock recommendation service with the self-reported goal of achieving 2–3% returns per week. This service provides daily buy/sell recommendations.

2.5.2. Quantitative techniques

The computational prediction of security markets follows two distinct paths, the first of which parallels information markets where an artificial market is constructed and predictions are made by varying system inputs (LeBaron et al., 1999; Raberto, Cincotti, et al., 2001, 2003). These inputs could be as simple as varying the time in which new information is received and acted upon (LeBaron et al., 1999) or more complex such as the modeling of an entire stock market exchange (Raberto et al., 2003). The second computational security prediction type is that of a quantitative nature. In real market predictions, quantitative systems, or quants, follow various stock parameters and are essentially automated versions of existing market strategies (e.g., look for high growth, unvalued securities, etc.) except with the ability to follow all stocks in real-time. This advantage has led quants to steadily outperform market averages by 2–3% for the past several years (Jelveh, 2006).

While the exact strategies used are a closely guarded secret, some quantitative funds do disclose the parameters they track. The exact number and weights assigned to these parameters fluctuate frequently to keep pace with market conditions and to tweak model performance. Quant programs are also becoming a part of the individual investor's toolbox as well. Wealth Lab Pro software (www.wealth-lab.com) allows individual investors to track upwards of 600 parameters through 1000 pre-set investment strategies (Lucchetti & Lahart, 2006) and many brokerage houses are giving their investors quantitative software as a customer retention tool.

The number of quant funds has increased from just a few in 2001 to over 150 by the beginning of 2006 (Burke, 2006). These funds have also branched themselves out, able to cover worldwide financial markets or focus exclusively on a select boutique of securities.

2.6. Research gaps

From our survey of the literature, we identified several potential gaps. First, machine learning systems have only been tested on either Universal financial news articles or on a Stock-Specific basis. Training a system using a hybrid of the two may result in more accurate results and a better-tuned prediction. Second, we are unaware of any system which has been compared against both trading professionals and existing quantitative systems. While there are several studies which pit

professionals against non-professionals and some mechanized techniques against others, it would be useful to make a broad comparison as a baseline evaluation to see just how well a financial news prediction system compares.

From these gaps, we have formulated a couple of research questions with which to explore. The first of which is:

- What effect does GICS partitioning of articles have on the prediction of stock price?

Following Bhojraj's conclusion that GICS classification is superior in its homogeneity of industries (Bhojraj et al., 2003), we propose to investigate the prediction accuracy of a system that explores the in-between areas of sector and Industry-level grouping and what impact it may have on prediction results.

The second research question we pursue is:

- How effective is a discrete prediction model versus the market and human traders?

Given our system, how does its predictive ability compare to the advice given by trading professionals, quantitative funds and the overall market in general?

3. System design

The AZFinText system builds upon several fields of prior research in linguistics, textual representation, machine learning and application of itself to a problem in finance. While the AZFinText system does not add anything new to these fields themselves, its contribution is the creation of the system itself which weaves together these disparate fields in the pursuit of solving a discrete prediction problem. The main goal of AZFinText is to learn what article terms are going to have an impact on stock prices, how much of an impact they will have, and then make an estimate of what the stock price is going to be 20 min into the future. This period of 20 min was derived from the prior work of Gidofalvi (2001). The premise is that certain article terms such as "factory exploded" or "workers strike" will have a depressing effect on stock prices whereas other article terms like "earnings rose" will have an increasing effect.

To properly evaluate our research questions, we designed the AZFinText system. Fig. 1 illustrates the AZFinText system design.

In examining the AZFinText system from Fig. 1, there are several major components to describe. The *Textual Analysis* component gathers financial news articles from Yahoo! Finance and represents them by their Proper Nouns. This module further limits extracted features to three or more occurrences in any document, which cuts down the noise from rarely used terms (Joachims, 1998). To identify the Proper Nouns we chose a modified version of the Arizona Text Extractor (AzTeK) system which performs semantic/syntactic word level tagging as well as phrase-based aggregation. AzTeK works by using a syntactic tagger to identify and aggregate the document's Noun Phrases and was found to have an 85% F-measure for both precision and recall, which is comparable to other tools (Tolle & Chen, 2000). Although the AzTeK system was selected due to availability, it performs adequately for Proper Noun extraction. However, there are many other such systems as reported in the Message Understanding Conference (McDonald et al., 2005), which can be adopted for financial news text analysis. The second major component is *Stock Quotations* which gathers stock price data in 1 min increments. The third major component is *Model Building*. This component has derived from prior empirical testing and includes article term representations and the stock price at the time the news article was released. This combination of parameters was previously tested and judged to provide superior performance to all combinations tested (Schumaker & Chen, 2006). The fourth major component is the *Trading Experts* which gathers the daily buy/sell recommendations from a variety of trading experts. Lastly, the metrics component evaluates system output.

To illustrate how the AZFinText system works, we offer a sample news article (Burns & Wutkowski, November 15, 2005) and step through the logic of our system.

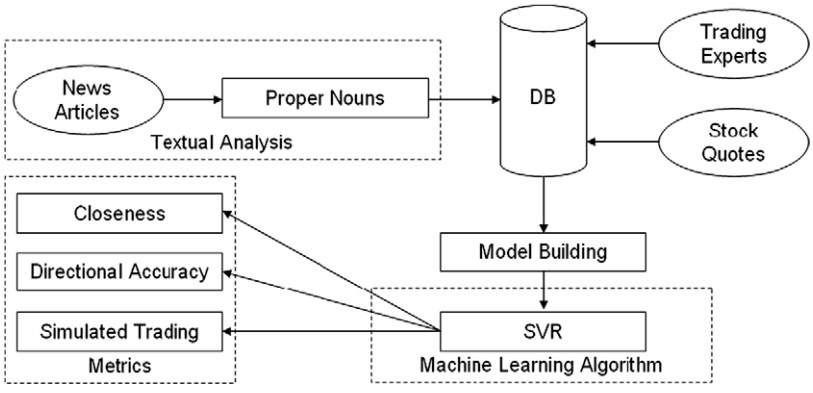

Fig. 1. AZFinText system design.

Schwab shares fell as much as 5.3 percent in morning trading on the New York Stock Exchange but later recouped some of the loss. San Francisco-based Schwab expects fourth-quarter profit of about 14 cents per share two cents below what it reported for the third quarter citing the impact of fee waivers a new national advertising campaign and severance charges. Analysts polled by Reuters Estimates on average had forecast profit of 16 cents per share for the fourth quarter. In September Schwab said it would drop account service fees and order handling charges its seventh price cut since May 2004. Chris Dodds the company s chief financial officer in a statement said the fee waivers and ad campaign will reduce fourth quarter pre-tax profit by $40 million while severance charges at Schwab s U.S. Trust unit for wealthy clients will cut profit by $10 million. The NYSE fined Schwab for not adequately protecting clients from investment advisers who misappropriated assets using such methods as the forging of checks and authorization letters. The improper activity took place from 1998 through the first quarter of 2003 the NYSE said. This case is a stern reminder that firms must have adequate procedures to supervise and control transfers of assets from customer accounts said Susan Merrill the Big Board s enforcement chief. It goes to the heart of customers expectations that their money is safe. Schwab also agreed to hire an outside consultant to review policies and procedures for the disbursement of customer assets and detection of possible misappropriations the NYSE said. Company spokeswoman Alison Wertheim said neither Schwab nor its employees were involved in the wrongdoing which she said was largely the fault of one party. She said Schwab has implemented a state-of-the-art surveillance system and improved its controls to monitor independent investment advisers. According to the NYSE Schwab serves about 5000 independent advisers who handle about 1.3 million accounts. Separately Schwab said October client daily average trades a closely watched indicator of customer activity rose 10 percent from September to 258 900 though total client assets fell 1 percent to $1.152 trillion. Schwab shares fell 36 cents to $15.64 in morning trading on the Big Board after earlier falling to $15.16. (Additional reporting by Dan Burns and Karey Wutkowski.)

Fig. 2 shows a sample run using the previous article. The extracted terms are represented in binary as either present or not. Supposing our corpus also contained the term Reuters which appeared in a different article and not in this instantiation, the term is given a zero for not being present in the current article. Terms are then filtered and only the Proper Nouns continue on for further analysis. For stock quotation data, we lookup what the stock price was at the time of article release ($15.65), and lookup the actual +20 min stock price for training and later evaluation ($15.59). This data (e.g., the Proper Noun matrix and stock price at the time the article was released) is then taken to the Model Building stage where the various models are given their appropriate data. These models are the different categories of GICS classification. For GICS Sectors, a model of GICS Sector 10 – Energy is built using only those articles (and prices at the time the article was released) of S&P 500 companies within this sector. Likewise additional models of other GICS Sectors, Industry Groups, Industries, and Sub-Industries are similarly built. The model is then given to the SVR algorithm where machine learning takes place and an estimate of the +20 min stock price is produced ($15.645). We can see from the stock prices given in this example that Schwab's share price dropped 6 cents while the model estimate figures a more conservative half penny drop.

The heart of the AZFinText system is the SVR algorithm. It receives a matrix of Proper Noun terms, represented in binary as being present in the financial news article or not, the stock price at the time the article was released and is further restricted by the model chosen to include only the articles/terms/data within the sector, group, Industry or Sub-Industry.

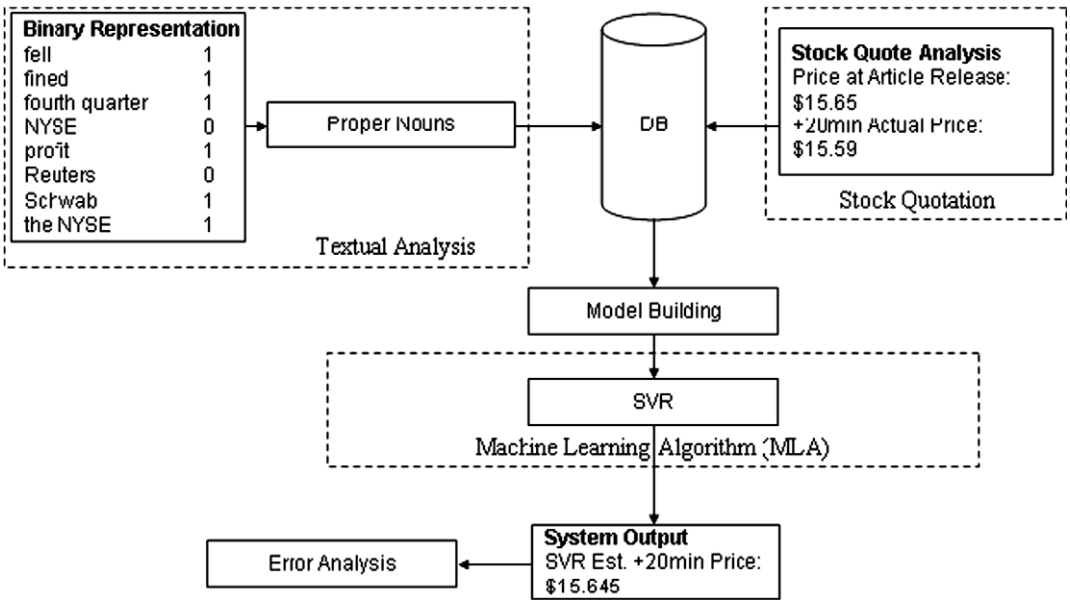

Fig. 2. Example of AZFinText representation.

The SVR algorithm then builds a hyperplane using the article terms and stock price and makes a prediction of what the price should be 20 min into the future for each news article. Once the predictions are made, measurements are taken to evaluate the accuracy of the results.

The first metric we use is a measure of *Closeness* which evaluates the accuracy of the predicted discrete value to the actual future price (Cho et al., 1999). It does so by calculating the mean square error (MSE) of the predicted values versus the actual +20 min stock price. The second metric is *Directional Accuracy*. This metric simply analyzes the direction of the predicted value versus the direction of the actual +20 min stock price with respect to the price at the time the article was released (Gidofalvi, 2001). The third metric we use is a *Simulated Trading Engine* which evaluates the potential gain/loss of our predictions. This metric examined the difference between the predicted price and the price at the time the article was released. If an absolute value was more than 1%, the system will then buy or short the stock depending upon the direction of the prediction (Lavrenko et al., 2000a). Our system differed slightly from Lavrenko's in that we invested blocks of $1000 compared to Lavrenko's $10,000. We then divest our holdings and take the +20 min stock price to derive our potential gain/loss.

Since the actual +20 min stock price in our example was $15.59, our measures of Closeness for this particular example is 0.003025. With the stock price at the time the article was released at $15.65 and going in the same direction as our prediction, we pass the directional accuracy test for this article. Since the predicted price was less than 1% movement from the original stock price, our Simulated Trading Engine declines to make a trade based on this article.

4. Experimental design

For our experiment, we selected a consecutive period of time to serve as our experimental baseline. We selected a 5-week research period of October 26, 2005–November 28, 2005, which incorporates 23 trading days. The 5-week period of study was selected because it gathered a comparable number of articles in comparison to prior studies: 6602 for Mittermayer (2004) and 5500 for Gidofalvi (2001). We also observe that the 5-week period chosen did not have unusual market conditions and would be a good testbed for our evaluation. To identify companies with more financial news, we further limited the scope of activity to focus on companies listed in the S&P 500 as of October 3, 2005. Articles gathered during this period were restricted to occur between the hours of 10:30 am and 3:40 pm. While trading starts at 9:30 am, we felt it important to reduce the impact of overnight news on stock prices and selected a period of 1-h to allow prices to adjust. The 3:40 pm cut-off for news articles was selected to disallow any +20 min stock predictions to occur after market hours. A further constraint was introduced to reduce the effects of confounding variables, where two articles on the same company cannot exist within 20 min of each other or both will be discarded.

The above processes had filtered the 9211 candidate news articles gathered during this period to 2809, where the majority of discarded articles occurred outside of market hours, as shown in Table 1. Similarly, 10,259,042 per-minute stock quotations were gathered during this period. This large testbed of time-tagged articles and fine-grain stock quotations allow us to perform our evaluation systematically.

For the machine learning algorithm we chose to implement the SVR Sequential Minimal Optimization (SMO) (Platt, 1999) function through Weka (Witten & Eibe, 2005). This function allows for discrete numeric prediction instead of classification. We selected a linear kernel and 10-fold cross validation. A similar prediction method was employed in the forecasting of futures contracts (Tay & Cao, 2001). To test the effects of GICS partitioning, we trained our system on keywords for all stocks, each GICS Sector, Industry Group, Industry and Sub-Industry, as well as trained on articles for each specific company. Output from these models is then evaluated on a three metric platform consistent with prior research (Schumaker & Chen, 2006).

To provide an overall sense of the data used throughout this experiment for the different classification models, we present Table 2 to illustrate some basic statistics on our dataset.

There are several facets of the above table which deserve further explanation. The first of which is that within the categories of Industry, Sub-Industry and Stock Specific, the number of GICS categories does not equal the number of usable categories. This is because of the GICS categories having fewer than ten news articles available for training which reduces the number of categories that can be used. We examined the role of decreasing articles and terms with each increasing level of classification and found that they had little impact on the results.

To test our second research question on the effectiveness of the AZFinText against professional traders and quantitative funds, we arbitrarily selected a group of experts and funds for comparison. The experts we chose were Jim Cramer, Jim Jubak and DayTraders.com. Selection criteria mainly focused on the availability and ease of gathering recommendations in an automated fashion. However, we were conscious to spread out experts across a variety of media sources. We then invest $1000 into each buy/short recommendation, buying at the opening price and then selling it at the close of trading. We further

Table 1
Article collections.

# Articles	
Raw number over study period	9211
Filtered article used by AZFinText	2809
Articles in Mittermayer's study	6602
Articles in Gidofalvi's study	5500

Table 2
Basic statistics on the AZFinText datasets.

	Universal	Sector	Industry Group	Industry	Sub-Industry	Stock Specific
Number of GICS categories	1	10	24	61	119	500
Usable categories	1	10	24	50	76	71
Minimum number of firms per category		9	6	1	1	
Maximum number of firms per category		89	36	22	16	
Average number of firms per category		500	50	21	10	5
Standard deviation of firms per category		26.7	9.8	5.6	3.6	
Minimum number of articles per category		100	16	11	10	10
Maximum Number of Articles per Category		518	266	150	139	58
Average number of articles per category	2809	281	117	55	34	18
Standard deviation of articles per category		160.8	66.0	37.6	28.3	9.1
Minimum number of terms per category		242	57	42	23	15
Maximum number of terms per category		974	606	415	377	194
Average number of terms per category	3710	567	288	158	105	61
Standard deviation of terms per category		291.7	134.4	88.5	67.8	30.8

assume that there is a zero transaction cost, consistent with Lavrenko et al. (2000a). Table 3 shows the recommendation breakdown for each of these Trading Professionals during our period of study.

While Jubak and DayTraders.com had few trading recommendations during our period of study, we felt that their contribution would help add depth to the overall trading professionals' advice. Similarly, since the gathered recommendations did not provide any predicted price information, we can only compare them to AZFinText using the Directional Accuracy metric.

As for comparisons against quantitative funds, we selected the top ten performing quantitative mutual funds of 2005 (Burke, 2006). Since quant trading strategies and predictions are closed-source, we gathered the observable fund prices at the beginning and end of our research period to make performance comparisons against our system using the Simulated Trading metric.

5. Experimental findings and discussion

5.1. Sector-based training has the best performance

To answer our first research question "*What effect does GICS partitioning of articles have on the prediction of stock price?*," we trained our system on the different GICS classification levels and evaluated them with our three aforementioned metrics. Results are presented in Table 4.

The first notable result is that Universal-training has the lowest average Closeness score of 0.0443. Sector-based training had the highest Directional Accuracy and Simulated Trading scores of 71.18% and 8.50%, respectively, (p-values <0.05). Comparing our results to the Universal versus Stock Specific research conducted by Lavrenko, our Stock Specific model should have a lower Closeness score than Universal-training (Lavrenko et al., 2000a). However, we observed the opposite result where Universal keywords had a lower Closeness score of 0.0443 as compared to Stock Specific's 1.0443. Returning to Lavrenko's work, we should also expect to see lower variances with Universal-training. Confirming this, Universal-training had a standard deviation of 0.1081 as compared to 2.7615 for Stock Specific. We believe that the observed uniformity is a result of Universal's homogenous keywords which behave similarly across all stocks. However, when we expand this work to include in-between GICS categories, Closeness scores gradually increase with each successive level and suddenly drop at Stock Specific training. Directional Accuracy and Simulated Trading appear to spike at the sector level and then steadily decrease. We believe that this behavior is the result of small inconsistencies within the GICS categories. We further investigated whether the reduction of either the number of articles or terms were a part of this phenomenon. However, when looking at the results of Stock Specific training, which used the least number of articles and terms, its numbers were somewhat similar to Universal, leaving the GICS categories with spiking behavior. Although both Bhojraj and Boni found the GICS classification system to be superior to similar systems (Bhojraj et al., 2003; Boni & Womack, 2004), its classifications are not perfect.

Analyzing model performance across all levels including those of the GICS, we found that sector-based training performed better than similar classifications. While sector had the best Directional Accuracy score of 71.18% and Simulated Trading of

Table 3
Recommendation breakdown of each trading professional.

	Cramer	Jubak	DayTraders.com
Number of buy recommendations	322	13	10
Number of sell recommendations	164	0	1
Number of unique companies mentioned	291	13	5

Table 4
Three metric evaluation of the different GICS levels.

Universal	Closeness		Directional accuracy		Simulated Trading	
	Average	Standard deviation	Average (%)	Standard deviation	Average (%)	Standard deviation
Universal	**0.0443**	0.1081	58.17	0.4933	2.86	0.1085
Sector	0.1954	3.7102	**71.18**	0.4530	**8.50**	0.2192
Group	3.3129	23.1133	66.12	0.4734	4.57	0.1789
Industry	16.1087	71.2319	62.37	0.4845	2.02	0.1572
Sub-Industry	26.1330	102.6304	57.50	0.4944	1.09	0.1501
Stock Specific	1.0443	2.7615	56.92	0.4954	1.01	0.1295

8.50% return, sector also had the second-lowest Closeness score of 0.1954 to Universal's 0.0443 (p-values <0.05). This would seem to indicate that sector-based training was better able to forecast stock price direction and capitalize on forecasted trades, yet was not as precise as Universal-training in obtaining a future price.

To further investigate the reasons behind sector's predictive success, Fig. 3 is a histogram of trade returns for the sector model.

From this figure, approximately 69% of trades had returns between -$100 and $100, however, the average is skewed towards the positive. It was also interesting to observe that 29 trades resulted in excess of $300 returns while only five trades lost $100 or more. This phenomenon was found to be a function of article length where longer articles contained more Proper Nouns and hence were more predictable. These returns arose from an outlay of $342,000 (i.e., 342 trades) and an excess return of $29,059.

We present Table 5, which contains results for each metric and compares each to a composite index return over the 5 week period of study, for a deeper analysis into the components of each sector.

From this table, the Financials sector (Sector 40), had the lowest Closeness score of 0.0189, the highest Directional Accuracy, 76.02% and a Simulated Trading return of 6.60% (p-values <0.05). However, this trading return did not outperform the market's composite Financials return of 8.25%. In a similar vein, the Materials sector, Sector 15, performed the worst with a Closeness score of 3.8269, Directional Accuracy of 63.00% and Simulated Trading return of 5.50% compared to the market's Materials composite of 8.76%. Sector 25, Consumer Discretionary, had the best Simulated Trading return of 19.40% while Sector 55, Utilities, had the worst return of −4.30% but did manage to lessen the loss of the composite average at −7.64%.

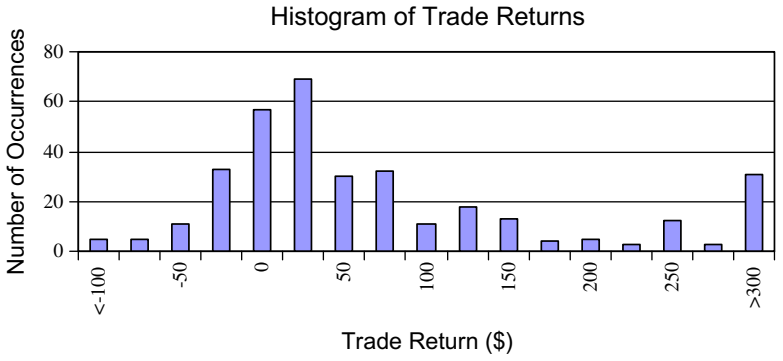

Fig. 3. Histogram of trade returns.

Table 5
Sector breakdown and evaluation.

Sector	Sector name	AZFinText			Index
		Closeness	Directional accuracy (%)	Simulated trade (%)	% Return
10	Energy	0.1951	68.79	17.50	1.60
15	Materials	**3.8269**	**63.00**	5.50	8.76
20	Industrials	0.0206	72.64	2.10	5.96
25	Consumer Discretionary	0.0616	70.04	**19.40**	6.05
30	Consumer staples	0.0218	64.81	2.60	2.07
35	Health care	0.0239	70.15	−1.00	4.28
40	Financials	**0.0189**	**76.02**	6.60	8.25
45	Information technology	0.0220	72.59	12.70	6.47
50	Telecommunication services	0.4684	72.36	0.50	7.66
55	Utilities	0.0475	68.24	−**4.30**	−7.64

To answer the follow-up question of why the Consumer Discretionary sector performed better than Utilities, we further investigated the Simulated Trading results of the constituent companies within each sector. Table 6 illustrates the component companies for Consumer Discretionary while Table 7 shows the same for Utilities. While not salient from Tables 5 and 6, Consumer Discretionary had 11 trades for nine companies with $0 gain/loss and Utilities had two trades for two companies with $0 gain/loss.

From these tables, Simulated Trading made trades in 40 of the 89 Consumer Discretionary companies and 12 of the 33 Utility companies. Table 6 shows that company TJX, The TJX Companies Inc., posted the largest gains with a 3.70% return on investment compared to a 10.15% increase over the 5 week period. For the Utilities sector of Table 7, there were quite a few low return transactions and the CPN, Calpine Corporation, trade with a −5.00% return further harmed the results of this sector. It is also notable that a good portion of trades in both sectors netted returns of $5 or less.

5.2. Sector-based training outperforms professional traders and quants

To answer our second research question "*How effective is a discrete prediction model versus the market and human traders?*," we measured the Directional Accuracy of our sector-based approach versus the Trading Professionals. Results are given in Table 8.

From this table, our sector-based training performed better in Directional Accuracy (71.18%) than Jim Cramer at 57.00% and Jim Jubak at 69.23%, however, our system did not perform as well against DayTraders.com at 81.82% (p-values <0.05). Even with statistical significance we must be mindful of the sparsity of stock recommendations from both Jubak and DayTraders.com. This sparsity of recommendations, as previously shown in Table 3, would suggest that Jim Jubak and DayTraders.com may be more conservative in their selection approaches.

Comparing our system against both Trading Professionals and the top 10 quant funds according to their trailing 1-year returns (Burke, 2006), results in the Simulated Trading are given in Table 9. Since quant trading strategies and predictions are closed-source, we could only gather observable price movements over our trading period and make comparisons using our Simulated Trading metric.

As evidenced by this table, AZFinText with its 8.50% return outperformed the overall market, 5.62% and the trading professionals (p-values <0.05). Comparing AZFinText against the top 10 quants shows AZFinText performed well, outperforming

Table 6
Company components of the Consumer Discretionary sector.

Company	Simulated Trading Engine		
	# Trades	Gain/loss	% Return
Consumer Discretionary sector			
BBY	2	$2	0.20
BLI	4	−$6	−0.60
CC	1	$1	0.10
CCL	1	$1	0.10
CTB	3	−$1	−0.10
CTX	1	$10	1.00
DCN	1	$2	0.20
DDS	1	$6	0.60
DHI	2	$4	0.40
DJ	1	$6	0.60
EBAY	3	$8	0.80
EK	2	$9	0.90
F	13	−$5	−0.50
FD	1	$3	0.30
FO	2	$12	1.20
GM	4	$24	2.40
GPS	3	$8	0.80
GT	1	$11	1.10
HAS	1	$2	0.20
HOT	1	$5	0.50
HRB	1	$2	0.20
IGT	1	$6	0.60
JCP	1	$9	0.90
KSS	1	−$1	−0.10
NYT	1	$4	0.40
OMX	1	$3	0.30
TGT	2	$4	0.40
TJX	**4**	**$37**	**3.70**
TWX	3	−$5	−0.50
VC	2	$22	2.20
YUM	1	$1	0.10

Table 7
Company components of the Utilities sector.

Company	Simulated Trading Engine		
	# Trades	Gain/loss	% Return
Utilities sector			
AES	1	$2	0.20
CIN	2	$8	0.80
CPN	**6**	**−$50**	**−5.00**
DTE	1	−$1	−0.10
DUK	1	$2	0.20
DYN	4	−$17	−1.70
KSE	1	$7	0.70
NI	1	$2	0.20
PPL	1	$1	0.10
TE	1	$1	0.10

Table 8
Comparison of directional accuracy results.

Directional accuracy (%)	
Sector	71.18
Cramer	57.00
Jubak	69.23
DayTraders	81.82

Table 9
Simulated Trading results of professionals and quants.

Simulated Trading (%)	
AZFinText	**8.50**
S&P 500	5.62
Trading professionals (%)	
Cramer	0.15
Jubak	−0.14
DayTraders.com	0.46
Quantitative funds (%)	
ProFunds Ultra Japan Inv (UJPIX)	24.73
ProFunds Ultra Japan Svc (UJPSX)	24.59
American Century Global Gold Adv (ACGGX)	12.96
American Century Global Gold Inv (BGEIX)	12.93
Quantitative Advisors Emerging Markets Instl (QEMAX)	8.16
Quantitative Advisors Emerging Markets Shs (QFFOX)	8.15
Lord Abbett Small-Cap Value Y (LRSYX)	5.22
Lord Abbett Small-Cap Value A (LRSCX)	5.19
Quantitative Advisors Foreign Value Instl (QFVIX)	4.99
Quantitative Advisors Foreign Value Shs (QFVOX)	4.95

six of the top 10 quant funds. It is interesting to note that the four better performing quants were trading in the Nikkei and gold markets where AZFinText was constrained to the companies in the S&P 500. In making a more direct performance comparison, Table 10 shows the trade returns of AZFinText versus several quant funds which are also operating within the S&P 500.

As shown in this table, AZFinText performed better than its peer quant funds. It is worthwhile to point out that AZFinText's success came mostly from making predictions from financial news articles and stock quotes, whereas quants used sophisticated mathematical models on a large set of financial variables. We believe that our research helps identify a promising research direction in financial text mining. However, more research is critically needed.

Table 10
Simulated Trading results of S&P 500 quants.

Return (%)	
AZFinText	**8.50**
Vanguard growth and income (VQNPX)	6.44
BlackRock investment trust portfolio Inv A (CEIAX)	5.48
RiverSource disciplined equity fund (ALEIX)	4.69

6. Conclusions and future directions

The contribution of this research is to examine the feasibility of a textual-based processing system that can process large amounts of financial news articles and make discrete predictions from them (AZFinText). In particular, this piece of research looked at comparing AZFinText's predictions against existing quantitative funds and human stock pricing experts.

Using a triangulation of evaluation methods which have been used mostly independently of each other, we determined that sector-based training had the better performance of the models tested. Sector had the best Directional Accuracy at 71.18% and Simulated Trading of 8.50% return on investment. Sector also had the second-lowest Closeness score, 0.1954, as compared to Universal, 0.0443. This would seem to indicate that sector-based training was better able to forecast stock price direction and capitalize on forecasted trades, yet was not as precise as Universal-training in obtaining a future price. In an analysis of the individual sectors we found that the Financials sector was unusually predictive with a Closeness score of 0.0189 and Directional Accuracy at 76.02%. This would mean that this sector was more sensitive to our representations than others.

When comparing the sector-based approach to trading professionals, we found that AZFinText had a Directional Accuracy of 71.18%, which was second-best to DayTraders.com's 81.82%. However, in Simulated Trading, sector-based training performed the best with an 8.50% return and even outperformed six of the top ten quantitative funds. We believe that our system was better able to capitalize on the Proper Nouns presented at the sector-level. This ability translated to consistently better predictions than comparable human experts while reacting to market changes faster than the history-dependent quantitative funds.

We feel it would be interesting to explore other machine learning techniques which can take advantage of the historical probabilities associated with stock prices. It would also be of merit to investigate stocks outside of the S&P 500. While S&P 500 stocks are widely followed by analysts and traders alike, smaller cap stocks may be more susceptible to share price movements from unexpected financial news.

References

Abbink, K., & Rockenbach, B. (2006). Option pricing by students and professional traders: A behavioural investigation. *Managerial and Decision Economics, 27*(6), 497–510.

Bhojraj, S., Lee, C. M. C., et al (2003). What's my line? A comparison of industry classification schemes for capital market research. *Journal of Accounting Research, 41*(5), 745–774.

Boni, L., & Womack, K. L., 2004. Industries, analysts, and price momentum. Tuck school of business working paper no. 03-12, Dartmouth College.

Burke, K., 2006. Not the man, but the machine. Retrieved 01.11.06.

Burns, D., & Wutkowski, K., 2005. Schwab to miss forecast, fined by NYSE (November 15, 2005). <http://biz.yahoo.com/rb/051115/financial_schwab.html?.v=3>. Retrieved from Yahoo! News.

Camerer, C., & Weigelt, K. (1991). Information mirages in experimental asset markets. *Journal of Business, 64*(4), 463–493.

Chan, L., Jegadeesh, N., et al (1996). Momentum strategies. *The Journal of Finance, 51*(5), 1681–1713.

Cho, V., Wuthrich, B., et al (1999). Text processing for classification. *Journal of Computational Intelligence in Finance, 7*(2).

Fama, E. (1964). *The behavior of stock market prices*. Graduate School of Business, University of Chicago.

Gidofalvi, G. (2001). *Using news articles to predict stock price movements*. San Diego: Department of Computer Science and Engineering, University of California.

Jelveh, Z. (2006). *How a computer knows what many managers don't*. The New York Times.

Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. In *Proceedings of the 10th European conference on machine learning* (pp. 137–142). Springer-Verlag.

Lavrenko, V., Schmill, M., et al., 2000a. Language models for financial news recommendation. In *International Conference on Information and Knowledge Management*, Washington, DC.

Lavrenko, V., Schmill, M., et al (2000b). *Mining of concurrent text and time series*. Boston, MA: International Knowledge Discovery and Data Mining.

LeBaron, B., Arthur, W. B., et al (1999). Time series properties of an artificial stock market. *Journal of Economic Dynamics and Control, 23*(9–10), 1487–1516.

Lucchetti, A., & Lahart, J. (2006). Your portfolio on autopilot; brokerages roll out software to automate trading strategies; risks of becoming a 'quant'. *Wall Street Journal*, B1 (New York).

Malkiel, B. G. (1973). *A random walk down wall street*. New York: W.W. Norton and Company Ltd.

McDonald, D. M., Chen, H., et al., 2005. Transforming open-source documents to terror networks: The Arizona terrornet. In *American association for artificial intelligence conference spring symposia*, Stanford, CA.

Mittermayer, M., 2004. Forecasting intraday stock price trends with text mining techniques. In *Hawaii international conference on system sciences*, Kailua-Kona, HI.

Platt, J. C. (1999). Fast training of support vector machines using sequential minimal optimization. In *Advances in kernel methods: Support vector learning* (pp. 185–208). MIT Press.

Raberto, M., Cincotti, S., et al (2001). Agent-based simulation of a financial market. *Physica A: Statistical Mechanics and its Applications, 299*(1–2), 319–327.

Raberto, M., Cincotti, S., et al (2003). Traders' long-run wealth in an artificial financial market. *Computational Economics, 22*(2), 255–272.

Schumaker, R. P., & Chen, H., 2006. Textual analysis of stock market prediction using financial news articles. In *12th Americas conference on information systems (AMCIS-2006)*, Acapulco, Mexico.

Schumaker, R. P., & Chen, H. (2008). Evaluating a news-aware quantitative trader: The effects of momentum and contrarian stock selection strategies. *Journal of the American Society for Information Science, 59*(2), 247–255.

Sekine, S., & Nobata, C., 2004. Definition, dictionaries and tagger for extended named entity hierarchy. In *Proceedings of the LREC*, Lisbon, Portugal.

Seo, Y.-W., Giampapa, J., et al., 2002. Text classification for intelligent portfolio management. Technical report CMU-RI-TR-02-14, Carnegie Mellon University.

Shmilovici, A., Alon-Brimer, Y., et al (2003). Using a stochastic complexity measure to check the efficient market hypothesis. *Computational Economics, 22*(2–3), 273–284.

Tay, F., & Cao, L. (2001). Application of support vector machines in financial time series forecasting. *Omega, 29*, 309–317.

Technical-Analysis, 2005. The trader's glossary of technical terms and topics. Retrieved 15.03.05.

Tolle, K. M., & Chen, H. (2000). Comparing noun phrasing techniques for use with medical digital library tools. *Journal of the American Society for Information Science, 51*(4), 352–370.

Torngren, G., & Montgomery, H. (2004). Worse than chance? Performance and confidence among professionals and laypeople in the stock market. *The Journal of Behavioral Finance, 5*(3), 148–153.
Vapnik, V. (1995). *The nature of statistical learning theory*. New York: Springer.
Witten, I. H., & Eibe, F. (2005). *Data mining: Practical machine learning tools and techniques*. San Francisco: Morgan Kaufmann.
Wuthrich, B., Cho, V., et al., 1998. Daily stock market forecast from textual web data. In *IEEE international conference on systems, man, and cybernetics*, San Diego, CA.
Yoon, Y., & Swales, G., 1991. Predicting stock price performance: A neural network approach. In *24th Hawaii international conference on system sciences*, Waikoloa, HI.