Textual Analysis of Stock Market Prediction Using Breaking Financial News: The AZFinText System

ROBERT P. SCHUMAKER
Iona College
and
HSINCHUN CHEN
University of Arizona

Our research examines a predictive machine learning approach for financial news articles analysis using several different textual representations: bag of words, noun phrases, and named entities. Through this approach, we investigated 9,211 financial news articles and 10,259,042 stock quotes covering the S&P 500 stocks during a five week period. We applied our analysis to estimate a discrete stock price twenty minutes after a news article was released. Using a support vector machine (SVM) derivative specially tailored for discrete numeric prediction and models containing different stock-specific variables, we show that the model containing both article terms and stock price at the time of article release had the best performance in closeness to the actual future stock price (MSE 0.04261), the same direction of price movement as the future price (57.1% directional accuracy) and the highest return using a simulated trading engine (2.06% return). We further investigated the different textual representations and found that a Proper Noun scheme performs better than the de facto standard of Bag of Words in all three metrics.

Categories and Subject Descriptors: G.1.10 [**Numerical Analysis**]: Applications; H.4.2 [**Information Systems Applications**]: Decision Support; I.2.7 [**Natural Language Processing**]: Text analysis

General Terms: Algorithms, Design, Experimentation, Measurement, Performance

Additional Key Words and Phrases: SVM, stock market, prediction

ACM Reference Format:

Schumaker, R. P. and Chen, H. 2009. Textual analysis of stock market prediction using breaking financial news: The AZFin text system. ACM Trans. Inform. Syst. 27, 2, Article 12 (February 2009), 19 pages. DOI = 10.1145/1462198.1462204 http://doi.acm.org/10.1145/1462198.1462204

Portions of this article were previously published as Schumaker and Chen [2006], Textual analysis of stock market prediction using financial news articles. In *Proceedings of the 12th Americas Conference on Information Systems (AMCIS)*.

Authors' addresses: R. P. Schumaker, Information Systems Dept., Iona College., New Rochelle, NY, 10801; email: Rob.schumaker@gmail.com; H. Chen, Artificial Intelligence Lab, Department of Management Information Systems, University of Arizona, Tucson, AZ, 85721.

1. INTRODUCTION

Stock market prediction has always had a certain appeal for researchers. While numerous scientific attempts have been made, no method has been discovered to accurately predict stock price movement. The difficulty of prediction lies in the complexities of modeling market dynamics. Even with a lack of consistent prediction methods, there have been some mild successes.

Stock market research encapsulates two elemental trading philosophies; fundamental and technical approaches [Technical-Analysis 2005]. In fundamental analysis, stock market price movements are believed to derive from a security's relative data. Fundamentalists use numeric information such as earnings, ratios, and management effectiveness to determine future forecasts. In technical analysis, it is believed that market timing is key. Technicians utilize charts and modeling techniques to identify trends in price and volume. These later individuals rely on historical data in order to predict future outcomes.

One area of limited success in stock market prediction comes from textual data. Information from quarterly reports or breaking news stories can dramatically affect the share price of a security. Most existing literature on financial text mining relies on identifying a predefined set of keywords and machine learning techniques. These methods typically assign weights to keywords in proportion to the movement of a share price. These types of analyses have shown a definite, but weak ability to forecast the direction of share prices.

In this article we experiment using several linguistic textual representations, including bag of words, noun phrases, and named entities approaches. We believe that combining more precise textual representations with past stock pricing information will yield improved predictability results.

This article is arranged as follows. Section 2 provides an overview of literature concerning stock market prediction, textual representations, and machine learning techniques. Section 3 describes our research questions. Section 4 outlines our system design. Section 5 provides an overview of our experimental design. Section 6 expresses our experimental findings and discusses their implications. Section 7 delivers our experimental conclusions with a brief discussion of future directions for this stream of research.

2. LITERATURE REVIEW

When predicting the future prices of stock market securities, there are several theories available. The first is efficient market hypothesis (EMH) [Fama 1964]. In EMH, it is assumed that the price of a security reflects all of the information available and that everyone has some degree of access to the information. Fama's theory further breaks EMH into three forms: weak, semi-strong, and strong. In weak EMH, only historical information is embedded in the current price. The semi-strong form goes a step further by incorporating all historical and currently public information in the price. The strong form includes historical, public, and private information, such as insider information, in the share price. From the tenets of EMH, it is believed that the market reacts instantaneously to any given news and that it is impossible to consistently outperform the market.

A different perspective on prediction comes from random walk theory [Malkiel 1973]. In this theory, stock market prediction is believed to be impossible where prices are determined randomly and outperforming the market is infeasible. Random walk theory has similar theoretical underpinnings to semi-strong EMH where all public information is assumed to be available to everyone. However, random walk theory declares that even with such information, future prediction is ineffective.

It is from these theories that two distinct trading philosophies emerged; the fundamentalists and the technicians. In a fundamentalist trading philosophy, the price of a security can be determined through the nuts and bolts of financial numbers. These numbers are derived from the overall economy, the particular industry's sector, or most typically, from the company itself. Figures such as inflation, joblessness, return on equity (ROE), debt levels, and individual price to earnings (PE) ratios can all play a part in determining the price of a stock.

In contrast, technical analysis depends on historical and time-series data. These strategists believe that market timing is critical and opportunities can be found through the careful averaging of historical price and volume movements and comparing them against current prices. Technicians also believe that there are certain high/low psychological price barriers such as support and resistance levels where opportunities may exist. They further reason that price movements are not totally random, however, technical analysis is considered to be more of an art form rather than a science and is subject to interpretation.

Both fundamentalists and technicians have developed certain techniques to predict prices from financial news articles. In one model that tested trading philosophies; LeBaron et al. [1999] posited that much can be learned from a simulated stock market with simulated traders. In their work, simulated traders mimicked human trading activity. Because of their artificial nature, the decisions made by these simulated traders can be dissected to identify key nuggets of information that would otherwise be difficult to obtain. The simulated traders were programmed to follow a rule hierarchy when responding to changes in the market; in this case it was the introduction of relevant news articles and/or numeric data updates. Each simulated trader was then varied on the timing between the point of receiving the information and reacting to it. The results were startling and found that the length of reaction time dictated a preference of trading philosophy. Simulated traders that acted quickly formed technical strategies, while traders that possessed a longer waiting period formed fundamental strategies [LeBaron et al. 1999]. It is believed that the technicians capitalized on the time lag by acting on information before the rest of the traders, which lent this research to support a weak ability to forecast the market for a brief period of time.

In similar research on real stock data and financial news articles, Gidofalvi [2001] gathered over 5,000 financial news articles concerning 12 stocks, and identified this brief duration of time to be a period of twenty-minutes before and twenty minutes after a financial news article was released. Within this period of time, Gidofalvi demonstrated that there exists a weak ability to predict the direction of a security before the market corrects itself to equilibrium. One reason for the weak ability to forecast is that financial news articles are typically reprinted throughout the various news wire services. Gidofalvi posits that a stronger predictive ability may exist in isolating the first release of an article. Using this twenty-minute window of opportunity and an automated textual news parsing system, the possibility exists to capitalize on stock price movements before human traders can act.

2.1 Textual Representation

There are a variety of methods available to analyze financial news articles. One of the more common methods is to apply a vector representation where article terms are indexed and then weighted. Selecting article terms can be as simple as tokenizing and using each word in the document. This technique assigns importance to determiners and prepositions which have little contribution to the overall meaning of the article. One method of circumventing these problems is to use a bag of words approach. In this approach, a list of semantically empty stop-words are removed from the article (e.g., the, a, and for). The remaining terms are then used as the textual representation. The bag of words approach has been used as the de facto standard of financial article research primarily because of its simple nature and its ability to produce a suitable representation of the text.

Building upon the bag of words approach, another tactic is to use a subset of terms as features [Moldovan et al. 2003], which can address issues related to article scaling while still encompassing the important concepts of an article [Tolle and Chen 2000]. One such method using this approach is noun phrasing. Noun phrasing is accomplished through the use of a syntax where parts of speech (i.e., nouns) are identified through the aid of a lexicon and aggregated using syntactic rules on the surrounding parts of speech, forming noun phrases.

A third method of article representation is named entities. This technique builds upon noun phrases by using lexical semantic/syntactic tagging where nouns and noun phrases can be classified under predetermined categories [Sekine and Nobata 2003]. This contrasts with using a differential approach, where concepts can be determined using a distributional analysis [Le Moigno et al. 2002]. An example of the predetermined category approach is the MUC-7 framework of entity classification, where categories include date, location, money, organization, percentage, person and time. The entity tagging procedure can happen in a number of ways. Typically, successful taggers have large lexicons of sample entities and/or word patterns, which may include both syntax and lexical information. Lexicons and pattern information can be used to designate features or machine learning approaches or be incorporated in rule in more of a rules-based approach. The large quantities of patterns are considered relatively cheap or shallow knowledge to obtain. Thus reusability of the extraction rules is not a priority. When input text is matched to the stored extraction patterns the corresponding input text gets assigned an entity tag. Because of the constrained categories, named entities in effect provide the smallest coverage of the document, but identify very specific types of phrases, which may or may not be helpful for stock price prediction.

Table I. Prior Algorithmic Research

Algorithm	Categories	Source Material	Examples
Genetic Algorithm	2 categories	Undisclosed number of chatroom posts	Thomas & Sycara, 2002
Naïve Bayesian	3 categories	5,000 articles from Lavrenko's collection	Gidofalvi et al. 2001
	5 categories	38,469 articles	Lavrenko et al. 2000
	5 categories	6,239 articles	Seo et al. 2002
SVM	3 categories	About 350,000 articles	Fung et al. 2002
	3 categories	6,602 articles	Mittermayer, 2004

Both noun phrases and named entities have shown limited success through previous comparison trials of tagging accuracy between differing algorithms. However, their usage as wide-scale textual representations for machine learning purposes remains somewhat unknown.

2.2 Machine Learning Algorithms

Like textual representation, there are also a variety of machine learning algorithms available. Almost all techniques start off with a technical analysis of historical security data by selecting a recent period of time and performing linear regression analysis to determine the price trend of the security. From there, a bag of words analysis is used to determine the textual keywords. Some keywords such as “earnings” or “loss” can lead to predictable outcomes which are then classified into stock movement prediction classes such as up, down, and unchanged. Much research has been done to investigate the various techniques that can lead to stock price classification. Table I illustrates a Stock Market prediction landscape of the various machine learning techniques.

From Table I, several items become readily noticeable. The first of these is that a variety of algorithms have been used. The second is that almost all instances commonly classify predicted stock movements into a set of classification categories, not a discrete price prediction. Lastly, not all of the studies were conducted on financial news articles, although a majority were.

The first technique of interest is the genetic algorithm. In this study, discussion boards were used as a source of independently generated financial news [Thomas and Sycara 2002]. In their approach, Thomas and Sycara attempted to classify stock prices using the number of postings and number of words posted about an article on a daily basis. It was found that positive share price movement was correlated to stocks with more than 10,000 posts. However, discussion board postings are quite susceptible to bias and noise.

Another machine learning technique, naïve Bayesian, represents each article as a weighted vector of keywords [Seo et al. 2002]. Phrase cooccurrence and price directionality is learned from the articles which lead to a trained classification system. One such problem with this style of machine learning is from a company mentioned in passing. An article may focus its attention on some other event and superficially reference a particular security. These types of problems can cloud the results of training by unintentionally attaching weight to a casually mentioned security.

One of the more interesting machine learners is support vector machines (SVM). In the work of Fung et al. [2002] regression analysis of technical data is used to identify price trends while SVM analysis of textual news articles is used to perform a binary classification in two predefined categories; stock price rise and drop. In cases where conflicting SVM classification ensues, such as both rise and drop classifiers are determined to be positive, the system returns a "no recommendation" decision. From their research using 350,000 financial news articles and a simulated buy-hold strategy based upon their SVM classifications, they showed that their technique of SVM classification was mildly profitable.

Mittermayer [2004] also used SVM is his research to find an optimal profit trading engine. While relying on a three tier classification system, this research focused on empirically establishing trading limits. It was found that profits can be maximized by buying or shorting stocks and taking profit on them at 1% up movement or 3% down movement. This method slightly beat random trading by yielding a 0.11% average return.

Many of the prior studies were classification oriented with questions asked such as; will this article cause the stock price to increase/decrease? These studies were all tests of directional movement and not the predictors of stock prices. Discrete prediction from numeric trends is hardly new. However, the application of this regression technique to SVM mechanics is rather recent [Gao et al. 2002]. One such method is sequential minimal optimization (SMO) [Platt 1999], where many of the scalability problems from using large training sets has been obviated through a more simplistic SVM solving technique. This combination of techniques has lead to the completely numeric prediction studies for futures contracts [Tay and Cao 2001], but discrete prediction has not been coupled with a systematic study of various textual analysis methods before.

From prior studies on the textual representation of documents, Joachims posits that limiting the inclusion of features to three or more instances per document will avoid the problem of unmanageably large feature spaces [Joachims 1998]. Extending this to textual representation, each feature is further represented in binary as either a zero or one; the term is either present or not present in the article [Vanschoenwinkel 2003]. This simple representational scheme is easy to implement and will lead to a sparse dataset with many zero features.

Applying these regression based methods and textual representation techniques to a supervised machine learning algorithm such as SVM can lead to a trained system with discrete numeric output.

Evaluation of output has been generally focused on only one of the following three metrics; measures of closeness, directional accuracy, or simulated trading. In measures of closeness, the estimated value from machine learning is compared against the actual value in a mean squared error (MSE) measure [Pai and Lin 2005]. Directional accuracy was the more common measure of previous financial studies, where the direction of the predicted value is compared with the movement direction of the actual value [Cho et al. 1998]. Whereas simulated trading initiates a simple trading engine to capitalize on large predicted value differences [Lavrenko et al. 2000a].

Table II. Examples of Textual Financial Data

Textual Source	Types	Examples	Description
Company Generated Sources	SEC Reports	8K	Reports on significant changes
		10K	Annual reports
Independently Generated Sources	Analyst Created	Recommendations	Buy/Hold/Sell assessments
		Stock Alerts	Alerts for share prices
	News Outlets	Financial Times	Financial News stories
		Wall Street Journal	Financial News stories
	News Wire Services	PRNewsWire	Breaking financial news articles
		Yahoo Finance	45 financial news wire sources
	Discussion Boards	The Motley Fool	Forum to share stock-related information

2.3 Financial News Article Sources

In real-world trading applications, the amount of textual data available to stock market traders is staggering. This data can come in the form of required shareholder reports, government-mandated forms, or news articles concerning a company's outlook. Articles and reports are also routinely cross-posted in many different locations leading to problems of uniqueness and database selection [Conrad and Claussen 2003]. Reports of an unexpected nature can lead to wildly significant changes in the price of a security. Table II illustrates some examples of textual financial data.

Textual data can arise from two sources: company generated and independently generated sources. Company generated sources such as quarterly and annual reports can provide a rich linguistic structure that if properly read can indicate how the company will perform in the future [Kloptchenko et al. 2004]. This textual wealth of information may not be explicitly shown by financial ratios but rather encapsulated in forward-looking statements or other textual locations. Independent sources such as analyst recommendations, news outlets, and wire services can provide a more balanced view of the company and have a lesser potential to bias news reports. Discussion boards can also provide independently generated financial news; however, they can be suspect sources.

News outlets can be differentiated from wire services in several different ways. One of the main differences is that news outlets are centers that publish available financial information at specific time intervals. Examples include *Bloomberg, Business Wire, CNN Financial News, Dow Jones, Financial Times, Forbes, Reuters*, and the *Wall Street Journal* [Cho 1999; Seo et al. 2002]. In contrast, news wire services publish available financial information as soon as it is publicly released or discovered. News wire examples include PRNewsWire, which has free and subscription levels for real-time financial news access, and Yahoo Finance, which is a compilation of 45 news wire services including the Associated Press and PRNewsWire. Besides their relevant and timely release of financial news articles, news wire articles are also easy to automatically gather and are an excellent source for computer-based algorithms.

Stock quotations are also an important source of financial information. Quotes can be divided into various increments of time from minutes to days, however, one minute increments provide sufficient granularity for machine learning.

While previous studies have mainly focused on the classification of stock price trends, none has been discovered to harness machine learning to determine a discrete stock price prediction based on breaking news articles. Prior techniques have relied solely on a bag of words approach and not other textual representations. Finally, there is no consensus on what information to include in a model that will lead to better performance. From these gaps in the research we form the crux of our study with the following questions.

3. RESEARCH QUESTIONS

Given that prior research in textual financial prediction has focused solely on the classification of stock price direction, we ask whether the prediction of discrete values is possible. This leads to our first research question.

—How effective is the prediction of discrete stock price values using textual financial news articles?

We expect to find that discrete prediction from textual financial news article is possible. Since prior research has indicated that certain keywords can have a direct impact on the movement of stock prices, we believe that predicting the magnitude of these movements is likely.

Prior research into stock price classification has almost exclusively relied on a bag of words approach. While this de facto standard has led to promising results, we feel that other textual representation schemes may provide better predictive ability, leading us to our second research question.

—Which combination of textual analysis techniques is most valuable in stock price prediction?

Since prior research has not examined this question before, we are cautious in answering such an exploratory issue. However, we feel that other textual representation schemes may serve to better distill the article into its essential components.

4. SYSTEM DESIGN

From these questions we developed the AZFinText system illustrated in Figure 1.

In this design, each financial news article is represented using three textual analysis techniques; bag of words, noun phrases, and named entities. These representations identify the important article terms and store them in the database. To limit the size of the feature space, we selected terms that occurred three or more times in a document [Joachims 1998].

To perform our textual analysis we chose a modified version of the Arizona Text Extractor (AzTeK) system which performs semantic/syntactic word level tagging as well as phrasal aggregation. AzTeK's noun phrasing component

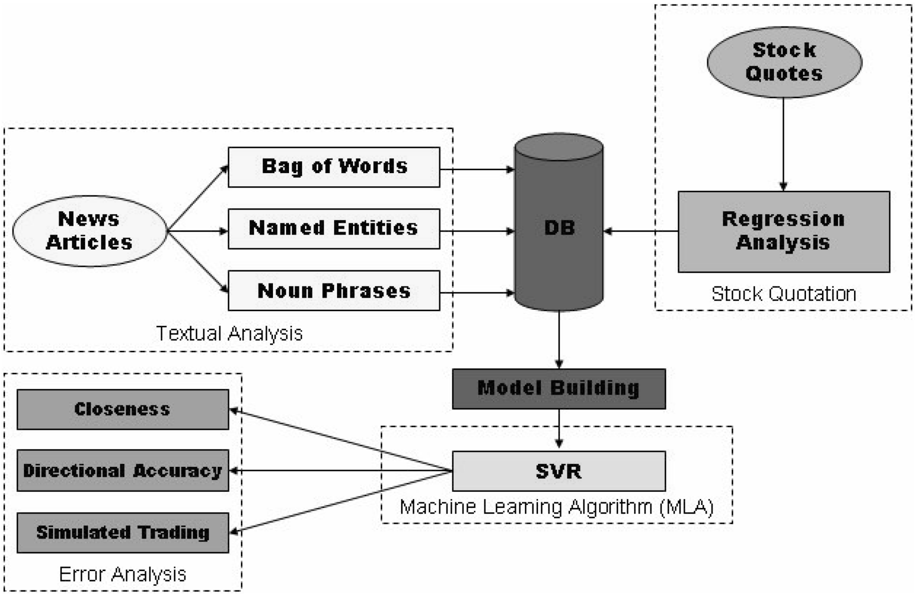

Fig. 1. AZFinText system design.

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 and Chen 2000]. The entity extractor portion goes one step further by assigning hybrid semantic/syntactic tags to document terms and phrases in one of the seven predefined categories of date, location, money, organization, percentage, person and time [McDonald et al. 2005]. These entities are then identified through the usage of a lexicon. Although the AzTeK system was selected due to availability, it performs adequately for noun phrase and named entity extraction. However, there are many other such systems, as reported in the Message Understanding Conference [McDonald et al. 2005], that can be adopted for financial news text analysis.

Stock quotes are gathered on a per minute basis for each stock. When a news article is released, we estimate what the stock price would be 20 minutes after the article was released. To do this we perform linear regression on the quotation data using an arbitrary 60 minutes prior to article release and extrapolate what the stock price should be 20 minutes in the future.

To test the types of information that need to be included, we developed four different models and varied the data given to them. The first model, regress, was a simple linear regression estimate of the +20 minute stock price. Assuming that breaking financial news articles have no impact on the movement of stock prices, we would expect a reasonable performance from this model. While we acknowledge the obvious violation of random walk theory, within such a compressed amount of time weak predictive ability remains [Gidofalvi 2001]. Next the three models use the supervised learning of SVM regression to compute their +20 minute predictions. Model M1 uses only extracted article terms for its prediction. While no baseline stock price exists within this model, we chose it because of its frequent usage in prior studies on directional classification of stock prices. Model M2 uses extracted article terms and the stock price at the time the article was released. We feel that given a baseline of stock price that this model will fare better. Model M3 uses extracted terms and a regressed estimate of the +20 minute stock price. This model may lead to better predictive results should the article terms have no impact on the movement of the stock price. All three models rely on using article terms in their prediction. SVM learns what terms lead to share price changes and adjust their weights depending on the severity of price changes.

To illustrate how the AZFinText system works, we offer a sample news article [Burns and Wutkowski 2005] and step through the logic of our system.

> 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 5 000 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

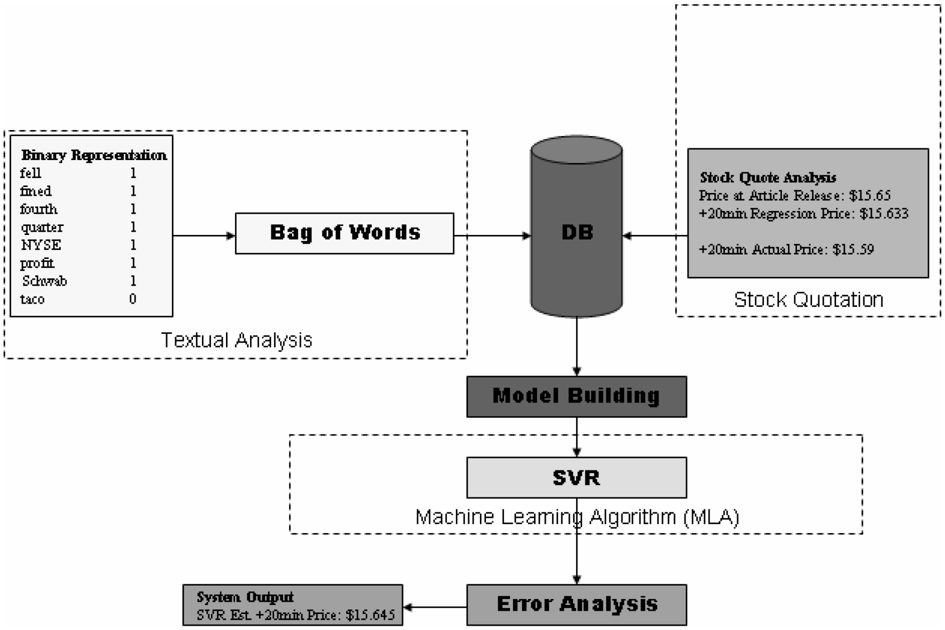

Fig. 2. Example AZFinText representation.

> after earlier falling to $15.16. (Additional reporting by Dan Burns and Karey Wutkowski)

The first step in our system is to extract the text from each article using our three textual representations, independently. This builds up 3 separate corpora for each representation. Then each article is passed through the system, one at a time, as is shown in Figure 2 with the prior Schwab article using the bag of words representation. In the Textual Analysis box, extracted terms are represented in binary as either present or not in this article. Supposing our corpus also contained the term Reuters that appeared in many different articles but not this one, the term is given a zero for not being present in the current article. For stock quotation data, we lookup what the stock price was at the time of article release ($15.65), calculate a regression estimate of the +20 minute stock price over the past hour ($15.633) and look up the actual +20 minute stock price for training and later evaluation ($15.59). This data is then taken to the model building stage where the various models are given their appropriate data. Following that, machine learning takes place and an estimate of the +20 minute stock price is produced ($15.645). We can see from the stock prices given in this example that Schwab's share price dropped six cents while the model estimate figures a more conservative half penny drop.

5. EXPERIMENTAL DESIGN

For our experiment we picked a research period of Oct. 26 to Nov. 28, 2005, to gather news articles and stock quotes. We further focused our attention only on companies listed in the S&P 500 as of Oct. 3, 2005. We acknowledge that several mergers and acquisitions did take place during this period of time; however, this only had an effect on less than 2% of the stocks tracked. In order to eliminate the "company in passing" problem, we gathered the news articles from Yahoo Finance using a company's stock ticker symbol. This resulted in articles on 484 of the 500 companies listed in the S&P 500. Articles were further constrained to a time frame of one hour after the stock market opened to twenty minutes before the market closed. This period of time allows for sufficient data to be gathered for prior regression trend analysis and future estimation purposes. We further limited the influence of articles such that we did not use any two or more articles that occurred within twenty minutes of each other. This measure eliminated several possible avenues of confounding results.

By performing these actions we gathered 9,211 candidate financial news articles and 10,259,042 stock quotes over the five-week period. From this pool of news articles we analyzed them using the three textual representations and retained only those terms that appeared three or more times in an article, which results in a differing number of articles. The filtering process resulted in the following breakdown:

—Bag of words used 4,296 terms from 2,839 articles

—Noun phrases used 5,283 terms from 2,849 articles

—Named entities used 2,856 terms from 2,620 articles

Article and stock quote data was then processed by a Support Vector Machine derivative, using Sequential Minimal Optimization [Platt 1999] in a form of regression, which can handle discrete number analysis .[Tay and Cao 2001].

Following training, we chose three evaluation metrics; closeness, directional accuracy, and a simulated trading engine. The closeness metric evaluated the difference between the predicted value and the actual stock price, measured using mean squared error (MSE). Directional accuracy measured the up/down direction of the predicted stock price compared with the actual direction of the stock price. While the inclusion of directional accuracy may not seem intuitive given the measure of closeness, it is possible to be close in prediction yet predict the wrong direction of movement. This leads us to a third evaluation measure using a simulated trading engine that invests $1,000 per trade and follows simple trading rules. The rules implemented by our trading engine are a modified version of those proposed by Mittermayer [2004] to maximize short-term trading profit. Our simulated trading engine evaluates each news article and will buy/short the stock if the predicted +20 minute stock price is greater than or equal to 1% movement from the stock price at the time the article was released. Any bought/shorted stocks are then sold after 20 minutes. This assumes a zero transaction cost which is consistent with the research of Lavrenko [Lavrenko et al. 2000a, 2000b] and Mittermayer [2004] who argue that trading in volume will offset the costs of trading.

6. EXPERIMENTAL FINDINGS AND DISCUSSION

In order to answer our research questions on the effectiveness of discrete stock prediction and the best textual representation; we tested our three models

Table III. Closeness Results

MSE	Regress	M1	M2	M3
Bag of Words	0.07279	930.87	0.04422	0.12605
Noun Phrases	0.07279	863.50	0.04887	0.17944
Named Entities	0.07065	741.83	0.03407	0.07711
Average	0.07212	848.15	0.04261	0.12893

Table IV. Directional Accuracy Results

Directional Accuracy	Regress	M1	M2	M3
Bag of Words	54.8%	52.4%	57.0%	57.0%
Noun Phrases	54.8%	56.4%	58.0%	56.9%
Named Entities	54.2%	55.0%	56.4%	56.7%
Totals	54.6%	54.6%	57.1%	56.9%

against a regression-based predictor using the three dimensions of analysis; measures of closeness, directional accuracy, and a simulated trading engine. Table III shows the results of the Closeness measures where smaller numbers indicate less error in price prediction. Table IV illustrates directional accuracy, where 50.0% could be achieved by chance alone. Finally, Table V displays the returns obtained from the simulated trading engine.

6.1 Model M2 with Articles Terms and Baseline Stock Price Performed the Best

From looking at the average results in Table III, Model M2 which used both article terms and the stock price at the time of article release, had the lowest MSE score (0.04261) of any of the models (p-values < 0.05). This result signifies that Model M2's predictions were closer to the actual +20 minute stock price than any of the other models including linear regression (regress). Looking deeper into the results, we find that Model M2 performed better than regress in each of the three textual representations, which supports Gidofalvi's claim of weak short-term predictability.

Model M1, which used only article terms, had a difficult time in its estimation of future stock prices with an average closeness score of 848.15. While this model may have been appropriate for prior classification-only studies, this unpleasant value was somewhat expected given the lack of baseline stock prices.

The other item of interest is that Model M2's named entities representation had the best performance at 0.03407 (p-values < 0.05). We will further investigate the effects of textual representation in a later section.

Turning our attention back to Model M2, we examined the weighting scheme that SVM assigned to the training variables. The stock price at the time of article release was given a weight of 0.9997 by SVM, while the article terms had a combined weight of 0.0003. While the weighting of article terms may appear superficially light, these terms are important because they provide the final touches to the estimated +20 minute stock price. If we were to rely on stock price alone without article terms, we would have the values of regress and Model M2, which used both the stock price and article terms performed better than regress. This signifies that the 0.0003 combined weighting of article terms is an important element in providing more accurate price results. If instead we used

Table V. Simulated Trading Engine Results

Trading Engine	Regress	M1	M2	M3
Bag of Words	−1.81%	−0.34%	1.59%	0.98%
Noun Phrases	−1.81%	0.62%	2.57%	1.17%
Named Entities	−2.26%	−0.47%	2.02%	2.97%
Totals	−1.95%	−0.05%	2.06%	1.67%

a regressed price estimate plus article terms, we would have Model M3, and M2 performed better than M3. However, article terms alone were not sufficient in estimating the future stock price, as demonstrated by Model M1.

In order to gain insight into the performance of our results, we can compare them to Pai and Lin, who conducted a similar study on forecasting stock prices [Pai and Lin 2005]. In their study, they attempted stock price prediction one day in advance, using a small set of stocks and only close-of-day prices. They managed an MSE score of 0.3001 and comparing that to our average MSE scores in Table III, our findings were an order of magnitude better.

In evaluating the directional accuracy results of Table IV, we again note that Model M2 performed better on average (57.1%) than the other models (p-values < 0.05). Regress did not perform so well (54.6%), which would seem to indicate that unexpected stock swings were captured by article terms. Comparing our results to previous studies shows that our values are somewhat reasonable. Cho et al. [1988], who used 100 days of training articles and 392 keywords, had an average directional accuracy of 46.8%.

In the simulated trading results of Table V, Model M2 using article terms and the stock price at the time of article release again had the best performance at 2.06% return (p-values < 0.05). This result would imply that Model M2 was better able to capitalize on trading opportunities given article terms and baseline stock price. Comparing this model against Model M1 with a trading return of −0.05%, we see that using article terms alone was insufficient. The results from regress (−1.95% return) were unexpected. We believe that this finding was from the news articles themselves affecting major changes in the share price of stocks. Correlating our results with prior studies, Lavrenko et. al. [2006b] claimed a 2% return from tracking four stocks over a forty-day period. In a similar study, Lavrenko et. al. [2000a] expanded the number of stocks to 127 over the same 40 day period and had a much lower return of 0.23. Both of these studies used essentially the same trading mechanism as we did which leads to an interesting observation; that perhaps more stocks lead to lower returns, although our study tracked 500 companies over 23 trading days. In a third simulated trading study, Mittermayer obtained a 0.11% average return using all of the stocks from Nasdaq, NYSE, and AMEX over a one year period of time. From our results, it would appear that our system is achieving fairly reasonable results at a 2.09% return.

Compiling all of the model performances together, Model M2 using article terms and the stock price at the time of article release performed best in all three metrics; measures of closeness (0.04261), directional accuracy (57.1%), and simulated trading (2.06% return). This model was better able to capture stock price movements and further bolsters the idea of weak short-term predictability. Our

Table VI. Closeness Results

MSE	Regress	M1	M2	M3	Average
Bag of Words	0.07279	930.87	0.04422	0.12605	232.77789
Noun Phrases	0.07279	863.50	0.04887	0.17944	215.95020
Named Entities	0.07065	741.83	0.03407	0.07711	185.50404

Table VII. Directional Accuracy Results

Directional Accuracy	Regress	M1	M2	M3	Totals
Bag of Words	54.8%	52.4%	57.0%	57.0%	55.3%
Noun Phrases	54.8%	56.4%	58.0%	56.9%	56.5%
Named Entities	54.2%	55.0%	56.4%	56.7%	55.6%

Table VIII. Simulated Trading Engine Results

Trading Engine	Regress	M1	M2	M3	Totals
Bag of Words	−1.81%	−0.34%	1.59%	0.98%	0.10%
Noun Phrases	−1.81%	0.62%	2.57%	1.17%	0.64%
Named Entities	−2.26%	−0.47%	2.02%	2.97%	0.57%

results were also inline with those from prior studies and mostly performed better. With some tweaking to how we classify directional movement, we feel that our system could produce better Directional Accuracy results as well.

6.2 A Superset of Named Entities Was the Best Textual Representation

To answer our second research question, which combination of textual analysis techniques is most valuable in stock price prediction, we compare the averages of each textual representation using our three metrics. Table VI presents the results of closeness measures, Table VII displays directional accuracy, and Table VIII illustrates the simulated trading engine.

From these tables, named entities had the lowest score in measures of closeness (185.50404) and noun phrases had the better score in both directional accuracy (56.5%) and simulated trading (0.64%), all p-values < 0.05. These seemingly confusing results were not as clear-cut as our model selection in the previous section as no one textual representation dominated the results.

However, it must be noted that these averaged results contain noise from previously failed models. If we were to focus only on the textual results for Model M2 and discard the other models, noun phrases performed the best in 2 of the 3 metrics and named entities in the remaining one.

These results ran contrary to our expectations. We had assumed that a named entity representation would generate better performance because of its ability to abstract the article terms and discard the noise of terms picked up by both bag of words and noun phrases. This MUC-7 textual representation was not sufficient to adequately model our article terms and lead us to ask the question, What were the differences between noun phrases and named entities? The answer was that named entities are essentially specialized proper nouns. The AzTeK system we used for part of speech tagging, identifies select terms in one of seven categories; date, location, money, organization, percentage, person and time [McDonald et al. 2005]. Words in these categories are basically a subset of

Table IX. Closeness Results

MSE	M2
Bag of Words	0.04422
Noun Phrases	0.04887
Proper Nouns	0.04433
Named Entities	0.03407

Table X. Directional Accuracy Results

Directional Accuracy	M2
Bag of Words	57.0%
Noun Phrases	58.0%
Proper Nouns	58.2%
Named Entities	56.4%

Table XI. Simulated Trading Engine Results

Trading Engine	M2
Bag of Words	1.59%
Noun Phrases	2.57%
Proper Nouns	2.84%
Named Entities	2.02%

noun phrases. We believe that expanding the number of categories for named entities will lead to a better representational scheme.

In order to investigate this we took a subset of terms from noun phrases that were tagged as proper nouns and introduced a fourth, hybrid, textual representation of proper nouns. This selection of terms is a comparable superset of named entities but without the entity categories. Proper nouns captured 3,710 article terms from 2,809 articles compared to the 5,283 terms in 2,849 articles for noun phrases and 2,856 terms in 2,620 articles for named entities.

To give the reader an understanding of what types of terms would be captured as proper nouns and not named entities, we refer back to the sample news article immediately preceding Figure 2. It is important to remember that named entities are derived using a semantic lexicon of previous input. Therefore, terms such as NYSE, which do not appear as a named entity, will be depicted in a proper noun representation.

Restating the metrics in terms of Model M2 to clear up some of the noise from the other failed models, we introduce the following data. Table IX shows measures of closeness, Table X, directional accuracy and Table XI, simulated trading.

The first item of interest is that the proper nouns subset performed better than noun phrases in all three metrics; 0.04433 to 0.04887 in measures of closeness, 58.2% to 58.0% in directional accuracy and 2.84% to 2.57% in simulated trading (all p-values < 0.05). This would seem to back up our initial expectation that a more abstract textual representation would perform better. In comparison to named entities, proper nouns performed better in two of the three metrics, directional accuracy and simulated trading whereas named entities had better success at measures of closeness. This would indicate that the direction we have undertaken is perhaps correct, but is still in need of refinement. We would suggest that future research should evaluate expanding the number of entity categories and evaluating the optimal mix for business-related news articles.

Overall, bag of words performed poorly by comparison. While this textual representation may be the de facto standard used in other studies, its weak performance is believed to arise from its reliance on too many noisy article terms. Noun phrases performed much better, with good performance in both directional accuracy and simulated trading. However, it suffered from poor closeness measures. We believe that this is the result of using a better tuned representational scheme of news articles, as compared to the bag of words approach. Yet noun phrases still possessed some elements of noise that led to less than desirable closeness scores. Named entities had some problems and did not perform as expected. While this representation had the best closeness score in prediction accuracy, it was unable to translate those gains into both directional accuracy and simulated trading returns. This is probably the result of using a limited set of entity categories which was unable to fully represent the content of financial news articles. Finally, proper nouns had the better performance results. While this textual representation can be thought of as the hybrid go-between for noun phrases and named entities, it had a solid performance on both directional accuracy and simulated trading. This result is likely attributable to proper nouns adequately using the article terms in a manner that was freer of the noise plaguing noun phrases and free of the constraining categories used by named entities.

7. CONCLUSIONS AND FUTURE DIRECTIONS

Our first conclusion was that model M2, using both article terms and the stock price at the time of article release, had a dominating performance in all three metrics; measures of closeness at 0.04261, directional accuracy at 57.1% and simulated trading at a 2.06% return. These results were the direct consequence of this model's ability to capitalize on the article terms and stock price for machine learning.

Our second conclusion was that proper nouns had the better textual representation performance. While it performed best in 2 of the 3 metrics, directional accuracy at 58.2% and simulated trading at 2.84%, it pulled up short on measures of closeness, 0.04433, as compared to named entities with 0.03407, all p-values < 0.05. However, this subset representation performed better than its parent, noun phrases, in all three metrics. We believe that proper nouns can attribute its success to being freer of the term noise used by noun phrases and free of the constraining categories used by named entities. Although more research into what constitutes an optimum mix of entity categories is encouraged.

Future research includes using other machine learning techniques such as relevance vector regression, which promises to have better accuracy and fewer vectors in classification [Bishop and Tipping 2003]. It would also be worthwhile to pursue expanding the selection of stocks outside of the S&P 500. While the S&P 500 is a fairly stable set of companies, perhaps more volatile and less tracked companies may provide interesting results. Another worthwhile approach would be to test a model based on article terms and percentage of stock price change. While our models relied on fixed stock prices that traded within a consistent range, penny stocks with wild fluctuations may prove worthy of further research. Lastly, while we trained our system on the entire S&P 500, it would be a good idea to try more selective article training such as industry groups or company peer group training and examine those results in terms of prediction accuracy.

Finally, there are some caveats to impart to readers. While the findings presented here are certainly interesting, we acknowledge that they rely on a small dataset. Using a larger dataset would help offset any market biases that are associated with using a compressed period of time, such as the effects of cyclic stocks, earnings reports, mergers and other unexpected surprises.

REFERENCES

Bishop, C. M. and Tipping, M. E. 2003. *Bayesian Regression and Classification*. IOS Press, Amsterdam.

Burns, D. and Wutkowski, K. Nov. 15, 2005. Schwab to miss forecast, fined by NYSE. http://biz.yahoo.com/rb/051115/financial_schwab.html?.v=3.

Cho, V. 1999. Knowledge Discovery from Distributed and Textual Data. Tech. rep. Department of Computer Science. Hong Kong University of Science and Technology.

Cho, V., Wuthrich, B., and Zhang, J. 1998. Text processing for classification. *J. Computat. Intel. Fin. 26*.

Conrad, J. G. and Claussen, J. R. S. 2003. Early user-system interaction for database selection in massive domain-specific online environments. *ACM Trans. Inform. Syst.* 21, 1, 94–131.

Fama, E. 1964. The behavior of stock market prices. Tech. rep. Graduate School of Business, University of Chicago.

Fung, G. P. C., Yu, J. X., Yu, X., and Lam, W. 2002. News sensitive stock trend prediction. In *Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD)*.

Gao, J. B., Gunn, S. R., Harris, C. J., and Brown, M. 2002. A probabilistic framework for SVM regression and error bar estimation. *Mach. Learn.* 46,1–3, 71–89.

Gidofalvi, G. 2001. Using news articles to predict stock price movements. Tech rep. Department of Computer Science and Engineering, University of California, San Diego.

Joachims, T. 1998. Text categorization with support vector machines: Learning with many relevant features. In *Proceedings of the 10th European Conference on Machine Learning*. Springer-Verlag, 137–142.

Kloptchenko, A., Eklund, T., Karlsson, J., Back, B., Vanharanta, H., and Visa, A. 2004. Combining data and text mining techniques for analysing financial reports. *Intel. Syst. Account. Fin. Manage.* 12, 1, 29–41.

Lavrenko, V., Schmill, M., Lawrie, D., and Ogilvie, P. 2000b. Mining of concurrent text and time series. In *Proceedings of the 6th ACM International Conference on Knowledge Discovery and Data Mining (KDD)*.

Lavrenko, V., Schmill, M., Lawrie, D., Ogilvie, P., Jensen, D., and Allan, J. 2000a. Language models for financial news recommendation. In *Proceedings of the 9th International Conference on Information and Knowledge Management*.

Le Moigno, S., Charlet, J., Bourigualt, D., Degoulet, P., and Jaulent, M.-C. 2002. Terminology extraction from text to build an ontology in surgical intensive care. In *Proceedings of the AMIA Symposium*.

LeBaron, B., Arthur, W. B., and Palmer, R. 1999. Time series properties of an artificial stock market. *J. Econ. Dynam. Contr. 23*, 9–10, 1487–1516.

Malkiel, B. G. 1973. *A Random Walk Down Wall Street*. W.W. Norton, New York.

McDonald, D. M., Chen, H., and Schumaker, R. P. 2005. Transforming open-source documents to terror networks: The Arizona TerrorNet. In *Proceedings of the American Association for Artificial Intelligence Conference Spring Symposia*.

Mittermayer, M.-A. 2004. Forecasting intraday stock price trends with text mining techniques. In *Proceedings of the 37th Hawaii International Conference on Social Systems*.

Moldovan, D., Pasca, M., Harabagiu, S., and Surdeanu, M. 2003. Performance issues and error analysis in an open-domain question answering system. *ACM Trans. Inform. Syst.* 21, 2, 133–154.

Pai, P.-F. and Lin, C.-S. 2005. A hybrid ARIMA and support vector machines model in stock price forecasting. *Omega* 33, 6, 497–505.

Platt, J. C. 1999. Fast training of support vector machines using sequential minimal optimization. In *Advances in Kernel Methods: Support Vector Learning*, MIT Press, 185–208.

Sekine, S. and Nobata, C. 2003. Definition, dictionaries and tagger for extended named entity hierarchy. In *Proceedings of the International Conference on Language Resources and Evaluation.*

Seo, Y.-W., Giampapa, J., and Sycara, K. 2002. Text classification for intelligent portfolio management. Tech rep. Robotics Institute, Carnegie Mellon University.

Tay, F. and 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. http://www.traders.com/documentation/RESource_docs/glossary/glossary.html.

Thomas, J. D. and Sycara, K. 2002. Integrating genetic algorithms and text learning for financial prediction. In *Proceedings of the Genetic and Evolutionary Computation Conference (GECCO)*.

Tolle, K. M. and Chen, H. 2000. Comparing noun phrasing techniques for use with medical digital library tools. *J. Amer. Soc. Inform. Sci.* 51, 4, 352–370.

Vanschoenwinkel, B. 2003. A discrete kernel approach to support vector machine learning in language independent named entity recognition. Tech. rep. Computational Modeling Lab, Vrije Universiteit, Brussels.

Received May 2006; revised May 2008; accepted September 2008