Textual Analysis of Stock Market Prediction Using Financial News Articles

Robert P. Schumaker and Hsinchun Chen
Artificial Intelligence Lab, Department of Management Information Systems
The University of Arizona, Tucson, Arizona 85721, USA
{rschumak, hchen}@eller.arizona.edu

Abstract

This paper examines the role of financial news articles on three different textual representations; Bag of Words, Noun Phrases, and Named Entities and their ability to predict discrete number stock prices twenty minutes after an article release. Using a Support Vector Machine (SVM) derivative, we show that our model had a statistically significant impact on predicting future stock prices compared to linear regression. We further demonstrate that using a Named Entities representation scheme performs better than the de facto standard of Bag of Words.

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 human behavior. 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 analysis have shown a definite, but weak ability to forecast the direction of share prices.

In this paper we experiment using several linguistic textual representations, including Bag of Words, Noun Phrases, and Named Entities approaches. We believe that using textual representations other than the de facto standard Bag of Words will yield improved predictability results.

This paper 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. Finally, Section 7 delivers our experimental conclusions with a brief oratory on 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 into 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, industry 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 the trading philosophies; LeBaron et. al. posited that much can be learned from a simulated stock market with simulated traders (LeBaron, Arthur et al. 1999). 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, Arthur 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 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 (Gidofalvi 2001). 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 because 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 simplest methods is to tokenize and use each word in the document. While this human friendly approach may help users to understand the syntactic structure of the document, machine learning techniques do not require such structural markings. This technique also assigns importance to determiners and prepositions which may not contribute much to the gist of the article. One method of circumventing these problems is 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 ease of use.

Building upon the Bag of Words approach, another tactic is to use certain parts of speech as features. This method addresses issues related to article scaling and can still encompass 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 a semantic lexical hierarchy where nouns and noun phrases can be classified as a person, organization, or location (Sekine and Nobata 2003). This hierarchy operates by analyzing the synonyms of each noun and generalizing their lexical profile across the rest of the noun phrase (McDonald, Chen et al. 2005). Named Entities in effect provide for a more abstract representation than Bag of Words or Noun Phrases.

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 ‘beat earnings’ or ‘unexpected loss’ can lead to predictable outcomes. These outcomes 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 1 illustrates a Stock Market prediction taxonomy of the various machine learning techniques.

Algorithm	Classification	Source Material	Examples
Genetic Algorithm	2 categories	Undisclosed number of chatroom postings	Thomas & Sycara, 2002
Naïve Bayesian	3 categories	Over 5,000 articles borrowed from Lavrenko	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

Table 1. Taxonomy of prior algorithmic research

From Table 1, several items become readily noticeable. The first of which is that a variety of techniques 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 was.

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, Giampapa et al. 2002). Phrase co-occurrence and price directionality is learned from the articles which leads 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., 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 (Fung, Yu et al. 2002). 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' category. 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 also used SVM is his research to find an optimal profit trading engine (Mittermayer 2004). 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 0.2% average return.

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. Reports of an unexpected nature can lead to wildly significant changes in the price of a security. Table 2 illustrates a taxonomy of textual financial data.

Textual Financial Source	Types	Examples	Description
Company Generated Sources	Quarterly & Annual Reports	8K	SEC-mandated report on significant company changes
		10K	SEC-mandated Annual reports
Independently Generated Sources	Analyst Created	Recommendations	Buy/Hold/Sell based on expert assessment
		Stock Alerts	Alerts triggered by barriers such as support/resistance levels
	News Outlets	Financial Times	Provides news stories on company activities
		Wall Street Journal	Provides news stories on company activities
	News Wire Services	PRNewsWire	Provides breaking financial news articles
		Yahoo Finance	Compilation of 45 independent financial news wire sources
	Financial Discussion Boards	The Motley Fool	A forum for investors to share stock-related information

Table 2. Taxonomy of textual financial data

Textual data itself 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, Eklund et al. 2004). This textual wealth of information may not be explicitly shown in the financial ratios but 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 look at 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, Giampapa 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.

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 on a Bag of Words approach and not the more abstract textual representations. From these gaps in the research we form the crux of our research 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.

- What textual representation can best predict future stock prices?

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.

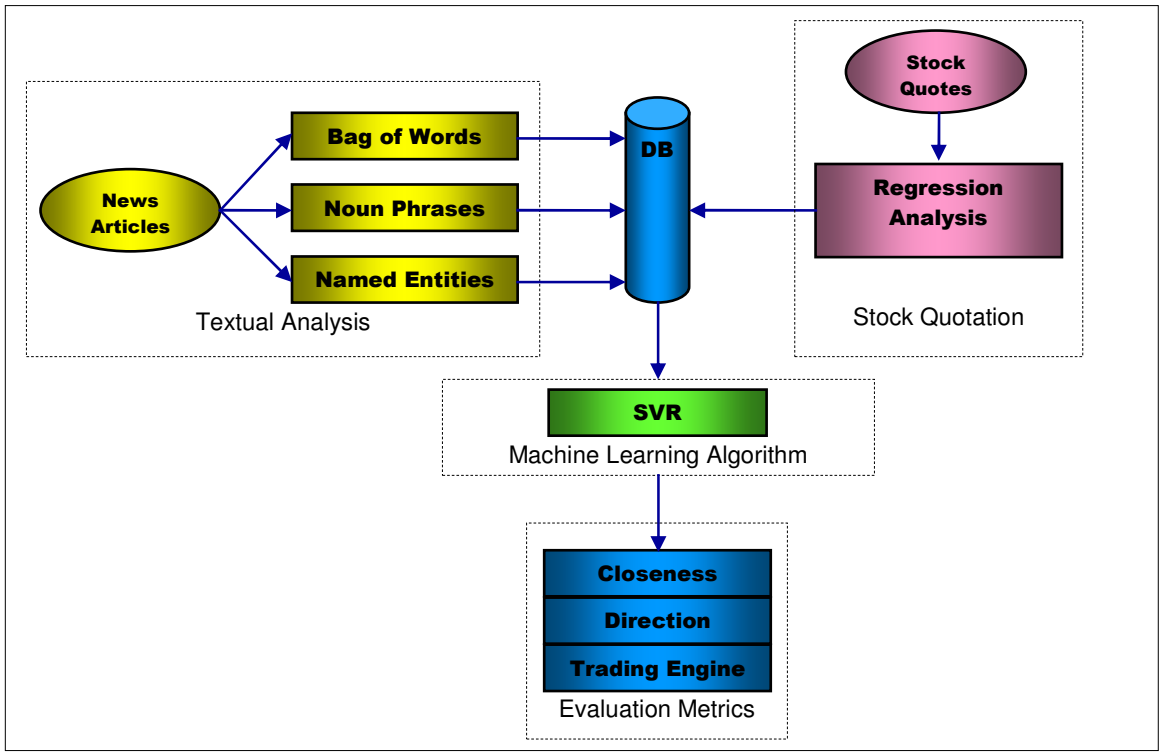

Figure 1. AZFinText system design

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).

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 60 minutes prior to article release and extrapolate what the stock price should be 20 minutes in the future. Any less time would yield a questionable estimation and any more time would severely limit the number of articles used.

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&P500 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. Articles were further constrained to a time frame of one hour after the stock market opened to twenty minutes before the market closes. 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. The filtering process resulted in the following breakdown:

- Bag of Words used 5,285 terms from 2,853 articles
- Noun Phrases used 5,293 terms from 2,861 articles
- Named Entities used 2,858 terms from 2,760 articles

Article and stock quote data was then processed by a Support Vector Machine derivative, Sequential Minimal Optimization style of regression (Platt 1999), which can handle discrete number analysis. We chose a linear kernel and performed 10-fold cross validation.

Following training, we chose three evaluation metrics; Closeness, Directional Accuracy, and a Simulated Training 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 to maximize short-term trading profit (Mittermayer 2004). 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.

To better understand the process employed, we selected an article and display the three textual representations for the benefit of the reader.

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

From this article, the following representative terms were identified and are displayed in Table 3.

Bag of Words	Noun Phrases	Named Entities
fined	Reuters	Reuters
fourth quarter	NYSE	fourth quarter
the NYSE	fourth quarter	Schwab
Schwab	profit	
profit	Schwab	
fell		

Table 3. Representative terms

As the reader will note, the Named Entities representation used the least number of terms of 3 unique terms, while Bag of Words used the most at 6 unique terms. Applying these terms to the weights obtained by the SVM model, returns the following +20 minute prediction; Bag of Words $15.645, Noun Phrases $15.629, and Named Entities $15.642. For this particular article the stock price at the time of article release was $15.65 and dropped to $15.59 in +20 minutes. While all three representations correctly predicted the downward direction of price movement, Noun Phrases was closer to the +20 minute price. As for the Simulated Trading Engine, none of these three representations predicted a value that would have triggered a trade.

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 model against a regression-based predictor

using the three dimensions of analysis; measures of Closeness, Directional Accuracy, and a Simulated Trading Engine. Table 4 shows the results of the Closeness measures, Table 5 illustrates Directional Accuracy, and Table 6 displays the results of the Simulated Trading Engine.

MSE Analysis	Regression	Our Model
Bag of Words	0.07253	0.04713
Noun Phrases	0.07257	0.05826
Named Entities	0.07244	0.03346

Table 4. MSE analysis of the data models

Directional Accuracy	Regression	Our Model
Bag of Words	47.6%	49.3%
Noun Phrases	47.6%	50.7%
Named Entities	47.4%	49.2%

Table 5. Direction Accuracy of the data models

Simulated Trading	Regression	Our Model
Bag of Words	-$1,809	$5,111
Noun Phrases	-$1,809	$6,353
Named Entities	-$1,879	$3,893

Table 6. Simulated trading engine on the data models

6.1 Our model predicts stock prices significantly better than regression

For our first research question, *how effective is the prediction of discrete stock price values using textual financial news articles*, we compare the regression estimate against our model using the three textual representations to yield a Closeness measure of MSE which is shown in Table 4. From this table, our model had statistically lower MSEs than the linear regression counterparts for each textual representation (p-value < 0.01); Bag of Words was 0.04713 to 0.07523, Noun Phrases was 0.05826 to 0.07257, and Named Entities was 0.03346 to 0.07244 respectively. This significantly lower MSE score across all three textual representations

means that our trained system was significantly closer to the actual +20 minute stock price than the regression estimate.

In looking at the Directional Accuracy metric of Table 5, we can further see that our predictive model was again consistently better in all three textual representations (p-value < 0.01); Bag of Words was 49.3% to 47.6%, Noun Phrases was 50.7% to 47.6%, and Named Entities was 49.2% to 47.4% respectively, where the percentage refers to the number of times the predicted value was in the correct direction as the +20 minute stock price. This result would further imply that our model was better able to use the textual financial news articles in a predictive capacity.

The third test was a Simulated Trading Engine. Table 6 shows that our predictive model again performed better than regression. Bag of Words gained $5,111 using our predictive model versus losing $1,809 for regression. Noun Phrases had a similar performance gaining $6,353 versus losing $1,809. Named Entities also gained $3,893 in our model versus losing $1,879 using regression.

From these three metrics it is quite clear that our machine learning method using article terms and the stock price at the time of article release, performed much better at predicting the +20 minute stock price than linear regression. While Random Walk theory may explain the failures of regression, it would appear that our model was better able to predict future stock prices.

6.2 Named Entities was the better textual representation

To answer our second research question, *what textual representation can best predict future stock prices*, we again look at Tables 4, 5, and 6, while scrutinizing the differences between various textual representations.

For the Closeness measures of Table 4, Named Entities had the lowest MSE score of 0.03346. Bag of Words was next lowest at 0.04713 and Noun Phrases had the highest MSE of the textual representations, at 0.05826 (p-values < 0.01).

In the Directional Accuracy measures of Table 5, quite the opposite effect was observed. Named Entities had the lowest directional accuracy at 49.2%, followed by Bag of Words at 49.3%, and Noun Phrases performed the best by predicting the stock price direction 50.7% of the time (p-values < 0.01).

As for the Simulated Trading Engine results from Table 6, Noun Phrases had the highest net profit of $6,353 followed by Bag of Words at $5,111 and Named Entities at $3,893.

These seemingly confusing results, where some textual representations perform better than others on particular metrics do not fully explain the observed effects. To answer this problem, we look deeper into the outlay of cash and percentage returns of the Simulated Trading Engine yielding Table 7.

Outlay	**Our Model**
Bag of Words	$228,000
Noun Phrases	$295,000
Named Entities	$108,000
Percentage Return	
Bag of Words	2.24%
Noun Phrases	2.15%
Named Entities	3.60%

Table 7. Simulated trading engine showing outlay and percentage return

From this table, Noun Phrases invested the most money, $295,000, and consequently also had the lowest percentage return of 2.15% of the three representations. Whereas Named Entities had the lowest investment amount of $108,000 and the highest return of the three at 3.60%.

Putting all of this information together; Noun Phrases had the best directional accuracy at 50.7%, but its investment strategy yielded the lowest return through poor stock picks. We

believe that the level of abstraction used by this textual representation was insufficient to capture the essence of the article for prediction purposes. This assertion can be backed up by the worst Closeness score of the three representations at 0.05826. While this representation was conservative enough to predict the direction of future stock prices, its wild investment strategy led to too many bad choices and thus reduced its appeal as a representation scheme for financial news articles.

By contrast, Named Entities had the lowest directional accuracy at 49.2%, but its more conservative investment approach coupled with its superior closeness score of 0.03346, led it to the best investment return (3.60%) of the three representations. We believe that the success of Named Entities is directly attributable to its ability to abstract away many of the semantically lesser important terms in an article, and generate a minimally representative essence of the article, suitable for near-term prediction.

7 Conclusions and Future Directions

From our research we found that our machine learning model using article terms and the stock price at the time of article release performed much better predicting the +20 minute stock price than linear regression. These results were consistent throughout the three evaluation metrics of Closeness, Directional Accuracy, and Simulated Trading across all three textual representations.

We further found that Named Entities performed best of the three representations tested. Named Entities conservative investment style steered it to the best investment return (3.60%) as well as the best Closeness score of 0.03346. While it did not have the best Directional Accuracy, we feel that there is some room for futher optimization. We believe that this representations success arises from its ability to abstract the article in a minimally representative way.

Future research includes using other machine learning techniques such as Relevance Vector Regression, which promises to have better accuracy and fewer vectors in classification. It would also be worthwhile to pursue expanding the selection of stocks outside of the S&P500. While the S&P500 is a fairly stable set of companies, perhaps more volatile and less tracked companies may provide interesting results. Lastly, while we trained our system on the entire S&P500, 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, and other unexpected surprises. However, datasets of several years duration could be imaginably unwieldy for this type of research.

References

1. Cho, V. (1999). Knowledge Discovery from Distributed and Textual Data. Computer Science. Hong Kong, The Hong Kong University of Science and Technology.

2. Fama, E. (1964). The Behavior of Stock Market Prices. Graduate School of Business, University of Chicago.

3. Fung, G. P. C., J. X. Yu, et al. (2002). News Sensitive Stock Trend Prediction. Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD), Taipei, Taiwan.

4. Gidofalvi, G. (2001). Using News Articles to Predict Stock Price Movements. Department of Computer Science and Engineering, University of California, San Diego.

5. Joachims, T. (1998). Text Categorization with Suport Vector Machines: Learning with Many Relevant Features. Proceedings of the 10th European Conference on Machine Learning, Springer-Verlag**:** 137-142.

6. Kloptchenko, A., T. Eklund, et al. (2004). "Combining Data and Text Mining Techniques for Analysing Financial Reports." Intelligent Systems in Accounting, Finance & Management 12(1): 29-41.

7. LeBaron, B., W. B. Arthur, et al. (1999). "Time Series Properties of an Artificial Stock Market." Journal of Economic Dynamics and Control 23(9-10): 1487-1516.

8. Malkiel, B. G. (1973). A Random Walk Down Wall Street. New York, W.W. Norton & Company Ltd.

9. McDonald, D. M., H. Chen, et al. (2005). Transforming Open-Source Documents to Terror Networks: The Arizona TerrorNet. American Association for Artificial Intelligence Conference, Stanford, CA.

10. Mittermayer, M.-A. (2004). Forecasting Intraday Stock Price Trends with Text Mining Techniques. Proceedings of the 37th Hawaii International Conference on Social Systems, Hawaii.

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

12. Sekine, S. and C. Nobata (2003). Definition, dictionaries and tagger for Extended Named Entity Hierarchy. Proceedings of the LREC.

13. Seo, Y.-W., J. Giampapa, et al. (2002). Text Classification for Intelligent Portfolio Management. Robotics Institute, Canegie Mellon University.

14. Technical-Analysis (2005). The Trader's Glossary of Technical Terms and Topics, http://www.traders.com. 2005.

15. Thomas, J. D. and K. Sycara (2002). Integrating Genetic Algorithms and Text Learning for Financial Prediction. Genetic and Evolutionary Computation Conference (GECCO), Las Vegas, NV.

16. Tolle, K. M. and H. Chen (2000). "Comparing Noun Phrasing Techniques for Use with Medical Digital Library Tools." Journal of the American Society for Information Science 51(4): 352-370.