Analyzing Parts of Speech and their Impact on Stock Price

Robert P. Schumaker
Computer and Information Science Dept.
Cleveland State University
Cleveland, Ohio 44115, USA
rob.schumaker@gmail.com

ABSTRACT

Financial articles can move stock prices. The terms used in an article can be a predictor of both price direction and the magnitude of movement. By investigating the usage of terms in financial news articles and coupling them with a discrete machine-learning algorithm, we can build a model of short-term price movement. From our research, we investigated the terms creating the largest price movements amongst five part of speech textual representations; bag of words, noun phrases, named entities, proper nouns and verbs.

Keywords: Knowledge management, prediction from textual documents, stock market research

INTRODUCTION

Term selection in a document is important. Not only for the author's ability to convey information precisely to an audience, but term selection can also be used to carry an emotional sentiment. This sentiment could reveal a bias in the author's treatment of a subject or uncover word choice tendencies. When applied to financial text documents, biases and word choice tendencies can have a real impact on stock price movement.

The ability to predict stock market behavior has always had a certain appeal to researchers. While numerous attempts to accurately predict price have been made, the difficulty has been incomplete models of human trading behavior, which at the core rely on rational decision-making. These human behavioral patterns are difficult to define and are constantly changing; thus making accurate predictions quite difficult. To further add to the uncertainty, there are two entirely opposed philosophies of stock market research; fundamental and technical analysis techniques (Technical Analysis, 2005). Fundamentalists seek to leverage a security's relative data, ratios and earnings, while technicians analyze charts and modeling techniques based on historical trading volume and pricing. The entire problem thus becomes, *does price history matter?*

As the roles of computers in electronic stock trading has grown, along with the ease of gathering information, it has been possible to not only test both the fundamental and technical trading models, but also to create electronic trading mechanisms without the problem of human bias. Many of these systems have simply followed the trend of automating existing fundamental and/or technical strategies. Their goal is to achieve better returns than human traders by removing the elements of emotion and bias from trading (Jelveh, 2006). The downside is that these systems lack intuition and will continue to buy even after unfavorable news events, such as losing a costly court battle. In order to work effectively, these systems require that news events be translated into numeric data before appropriate decisions can be made. This

information-translation problem introduces serious lag-time into decisions and in some cases human analysts must override trades.

The motivation of this paper is to build and test a financial news article system that investigates those terms that create the most price movement in textual financial news articles. By identifying those terms, researchers and traders alike can further refine existing quantitative models and further the science of price prediction in the stock market.

This paper is arranged as follows. Section 2 provides an overview of literature concerning Stock Market prediction and textual representation techniques. Sections 3 and 4 describe our proposed approaches and the AZFinText system respectively. Section 5 provides an overview of our experimental design. Section 6 details our experimental findings and discusses their impact on price prediction. Section 7 delivers our conclusions.

LITERATURE REVIEW

There are two major market prediction theories; Efficient Market Hypothesis (EMH) and Random Walk Theory. In EMH, the price of a security is a reflection of complete market information and when new information is added, the market instantly adjusts the price of the security (Fama, 1964). EMH can vary the amount of information sharing throughout the market in three distinct levels; the weak form, the semi-strong and the strong form. In weak EMH, only historical data is embedded within the current price. Semi-Strong EMH incorporates both historical and current public information into its prices. Strong EMH includes all pertinent information such as history, current public information and private information, such as insider trading. From EMH theory, it is the belief that markets behave efficiently and instantaneous price corrections make any prediction model useless.

Random Walk Theory is similar to Semi-Strong EMH where all information is contained within the current price and cannot be used in future prediction. This theory slightly differs from EMH by maintaining that short-term price movements are indistinguishable from random noise (Malkiel, 1973).

Under Random Walk Theory, this short-term random movement produces unpredictable prices and makes it impossible to consistently outperform the market.

The ability to scrutinize human trading decisions and uncover the effects of trading behavior throughout a market exchange is an extremely difficult problem. To lessen this complexity and simultaneously test the impact of fundamental and technical trading strategies, LeBaron created an artificial stock market with simulated traders whose trading decisions can be dissected (LeBaron et al., 1999). LeBaron accomplished this by introducing new pieces of information into the market and then adjusting the amount of time between when an individual trader would receive information and act upon it. He discovered that traders with longer waiting times formed fundamental trading strategies (e.g., relying more heavily on company-specific performance data) while those with shorter waiting times developed technical strategies (e.g., timing a market trade). This study led to a discovery between the lag time that information is introduced and the time when the market returns to equilibrium. This delay in market behavior helped to dismiss the instantaneous correction tenets of EMH and lent support to the idea that within these informational lag-times, markets can be forecast following the introduction of new information. Further research into the limits of this lag-time length led to the discovery of a twenty minute window of opportunity before and after a financial news article is released (Gidofalvi, 2001). Within this window, weak prediction of the direction of a stock price is possible.

FINANCIAL NEWS ARTICLES

Information is introduced into the stock market all the time. This information can take the form of rumors, eavesdropping and scandals and all can have a visible impact on stock market prices. Textual financial news articles are considered to be a more stable and trustworthy source. This stability has caused some to declare that news can be considered another form of commodity (Mowshowitz, 1992) that can have differing values (Raban & Rafaeli, 2006). However, the exact relationship between financial news articles and stock price movement is complex. Even when the information contained in financial news articles can have a visible impact on a security's price (Gidofalvi, 2001; Lavrenko et al., 2000a; Mittermayer, 2004; Wuthrich et al., 1998), textual financial articles are not the sole determinant of price

movement. Sudden price movements can still occur from other sources, such as large unexpected trades (Camerer & Weigelt, 1991).

The first challenge of a textual financial prediction system is to manage the large amounts of textual information that exist for market securities. This material can include required reports such as periodic SEC filings, press releases and financial news articles reporting both unexpected events and routine news alike. These textual documents can then be parsed using Natural Language Processing (NLP) techniques to identify specific article terms or phrases most likely to cause dramatic share price changes, such as "factory exploded" would probably indicate a price plunge in the near future. By automating this process, machines can take advantage of arbitrage opportunities faster than human counterparts by repeatedly forecasting price fluctuations and executing immediate trades.

Obtaining timely financial documents from reputable Web sources is a critical step and there are many financial news aggregation sites to provide this service. One of these sites is Comtex, which offers real-time financial news in a subscription format. Another source is PRNewsWire, which offers free real-time and subscription-based services. Yahoo! Finance is a third such source and is a compilation of 45 different news sources including the Associated Press, Financial Times and PRNewsWire among others. This source provides a variety of perspectives and timely news stories regarding financial markets.

TEXTUAL REPRESENTATION

Once financial news articles have been gathered, we need a way to represent their important features in machine-friendly form. One technique is a Bag of Words approach which has been extensively used in textual financial research (Gidofalvi, 2001; Lavrenko et al., 2000a). This process involves removing semantically meaningless stopwords such as conjunctions and declaratives from the text and using what remains as the textual representation. While the Bag of Words method has been popular in linguistic research, it suffers from noise issues associated with seldom-used terms and problems of scalability, where immense computational power is required for large datasets. An improved representational system is Noun Phrases. This representation retains only the nouns and noun phrases from a document and can adequately represent the important article concepts (Tolle & Chen, 2000). As a result, this technique uses

far fewer terms and can handle article scaling better than Bag of Words. A third representational technique is Named Entities, which is an extension of Noun Phrases. This technique selects the proper nouns of an article that fall within the purview of several well-defined categories. This process uses a semantic lexical hierarchy (Sekine & Nobata, 2004) as well as a syntactic/semantic tagging process (McDonald et al., 2005) to assign candidate terms to categories. The exact categorical definitions are described in the Message Understanding Conference (MUC-7) Information Retrieval task and encompass the entities of date, location, money, organization, percentage, person and time. The Named Entities representation allows for better generalization of previously unseen terms and does not possess the scalability problems associated with a semantics-only approach. A fourth representational technique is Proper Nouns. This method functions as an intermediary between Noun Phrases and Named Entities where it exists as a subset of Noun Phrases by selecting specific nouns, while at the same time is a superset of Named Entities without the constraint of pre-defined categories. This representation removes the ambiguity associated with proper nouns that can either be represented by more than one named entity category or fall outside one of the seven pre-defined categories. In a comparison study using these four representational techniques, it was found that the Proper Noun representation was much more effective in representing textual financial news articles (Schumaker & Chen, 2006). Another representational technique is to use only the verbs of the article. It is thought that the choice of verbs in an article can convey more information to the reader. In a study of using verbs in textual financial research, it was found that verbs can adequately convey the meaning of the article and are conducive to textual financial prediction (Schumaker & Chen, 2009).

Assigning a representational mechanism is not sufficient to address scalability issues associated with large datasets. A common solution is to introduce a term frequency threshold that uses a term frequency cut-off to represent article terms that appear more frequently (Joachims, 1998). This technique not only eliminates noise from lesser-used terms, but also reduces the number of features to represent. Once scalability issues are addressed, the data needs to be prepared in a more machine-friendly manner. Machine learning algorithms are unable to process raw article terms and require an additional layer of

representation. One popular method is to represent article terms in binary where the term is either present or not in a given article (Joachims, 1998). This solution leads to large but sparse matrices where the number of represented terms throughout the dataset will greatly outnumber the terms used in an individual article.

Once financial news articles are represented in machine form, learning algorithms can then begin to identify patterns of predictable behavior. One accepted method, Support Vector Regression (SVR), is a regression equivalent of Support Vector Machines (SVM) but without the aspect of classification (Vapnik, 1995). Like SVM, SVR attempts to minimize its fitting error while maximizing its goal function by fitting a regression estimate through a multi-dimensional hyperplane. This method is also well-suited to handling textual input as binary representations and has been used in similar financial news studies (Schumaker & Chen, 2006; Tay & Cao, 2001).

RESEARCH QUESTIONS

From this look at textual representation schemes for financial news articles, we have formulated several research questions. The first of which is:

- What terms create the most movement in a price prediction model?

Certain terms are expected to move stock prices more than others. However, these terms are expected to differ based upon the textual representation method used. Further, since some of the textual representation schemes are closely related (e.g., Noun Phrases and Named Entities may share similar terms), we further ask:

- What influential terms appear across similar textual representations?

The answer to this question will reveal those terms that exhibit the most influence in price prediction.

SYSTEM DESIGN

In order to evaluate our research questions, we designed the AZFinText system. Figure 1 illustrates the AZFinText system design.

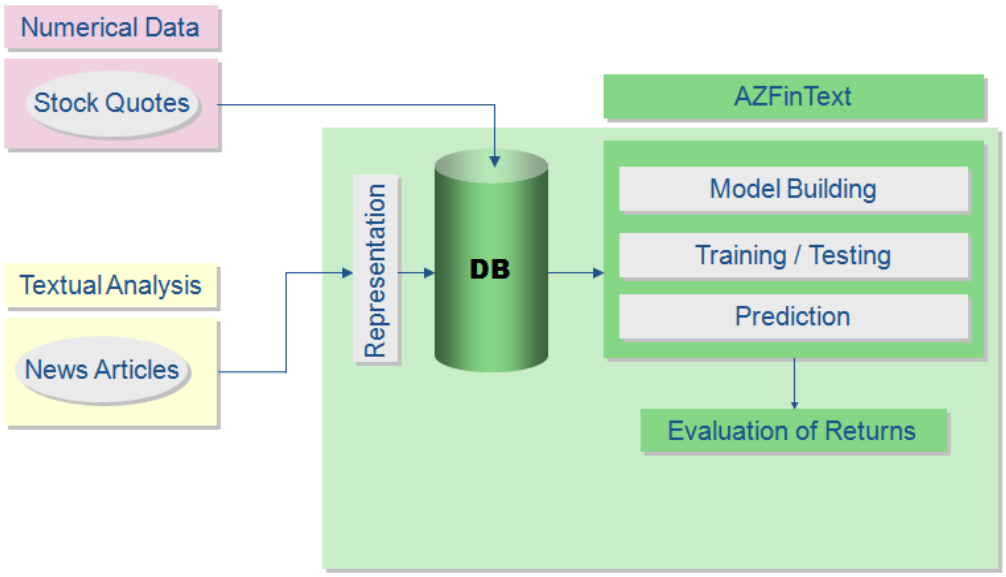

Figure 1. The AZFinText System

From the AZFinText system design in Figure 1, there are several major components to describe in detail. The first component is Numerical Data that gathers stock price data in one-minute increments from a commercially available stock price database. The second component is Textual Analysis. This component gathers financial news articles from Yahoo! Finance and represents them using the five textual representations; bag of words, noun phrases, named entities, proper nouns and verbs. This module further limits extracted features to three or more occurrences in any document, which cuts down the noise from rarely used terms (Joachims, 1998).

Once the data is gathered, AZFinText builds the appropriate textual models and trains the SVR algorithm on the price of the stock at the time the article was released as well as the terms used in the particular article, based on the representation scheme.

For the machine learning algorithm we chose to implement the SVR Sequential Minimal Optimization (Platt, 1999) function through Weka (Witten & Eibe, 2005). This function allows discrete numeric prediction instead of classification. We selected a linear kernel and ten-fold cross-validation. A similar prediction method was employed in the forecasting of futures contracts (Tay & Cao, 2001).

AZFinText is then trained on the data and issues price predictions for each financial news article encountered. Evaluations are then made regarding the effect of stock returns in terms of the models generated.

EXPERIMENTAL DESIGN

For the experiment, we selected a consecutive five week period of time to serve as our experimental baseline. This period of research was from Oct. 26, 2005 to Nov. 28, 2005 and incorporates twenty-three trading days. The five-week period of study was selected because it gathered a comparable number of articles in comparison to prior studies: 6,602 for Mittermayer (Mittermayer, 2004) and 5,500 for Gidofalvi (Gidofalvi, 2001). We also observe that the five-week period chosen did not have unusual market conditions and was a good testbed for our evaluation. In order to identify the companies with the most likelihood of having quality financial news, we limited our scope of activity to only those companies listed in the S&P 500 as of Oct. 3, 2005. Articles gathered during this period were restricted to occur between the hours of 10:30am and 3:40pm. Even though trading starts at 9:30am, we felt it important to reduce the impact of overnight news on stock prices and selected a period of one-hour to allow these prices to adjust. The 3:40pm cut-off was selected to disallow any +20 minute stock predictions to occur after market hours. A further constraint to reduce the effects of confounding variables was introduced where two articles on the same company cannot exist within twenty minutes of each other or both will be discarded. The above processes filtered the 9,211 candidate news articles gathered during this period to 2,802, where the majority of discarded articles occurred outside of market hours. Similarly, 10,259,042 per-minute stock quotations were gathered during this period. This large testbed of time-tagged articles and fine-grain stock quotations allow us to perform a systematic evaluation.

The first task is to extract financial news articles. The entire corpus of financial news articles is represented by each of the five textual representations in binary. If a particular feature is present in the article, that feature is given a 1, else a 0 and then stored in the database. Similarly, stock quotations gathered on a per minute basis and stored. To build a model, we first pair together the financial article's

representation and stock price at the time the article was released, for each financial news article. Then, depending upon the particular model that is tested, data is aggregated and passed to the machine-learning component for training and testing. Stock price predictions are then made for each financial news article and each term's role in the prediction model is analyzed.

EXPERIMENTAL RESULTS

To answer our first research question of *what terms create the most movement in a price prediction model*, we tested the five textual representation models and extracted those terms that are creating the most price movement. The statistics from these five representations are presented in Table 1.

	Bag of Words	Noun Phrases	Named Entities	Proper Nouns	Verbs
# Supp. Vect.	1,847	2,302	927	1,435	197
# Positive	821	1,026	418	699	90
# Zeros	159	119	40	72	8
# Negatives	867	1,157	469	664	99
StkZero	0.9994	0.9985	0.9997	0.9993	1.0001
Constant	0.0004	0.0008	0.0000	0.0001	0.0002
Term Weight	0.0002	0.0007	0.0003	0.0006	-0.0003

Table 1. Representation statistics

The first variable, number of support vectors, is critical in understanding the SVM algorithm. While hundreds or thousands of terms may comprise a particular financial news article, the SVR algorithm is tasked with trying to maximally divide the multi-dimensional data by creating a mathematical regression through the hyperplane. As a result of this division, certain terms close to the divide are used in that mathematical regression and are hence referred to as support vectors. They are further broken down into positive, negative and zeros components. Positive support vectors are those article terms that exhibit a positive influence on the price of the stock. Negative support vectors are as the name implies, those terms that create a negative impact on stock price. There are also support vectors that have neither a positive nor negative influence on the stock price. While these support vectors are not useful for our study of terms and price influence, they are important to the SVR regression calculation and are presented here as

a courtesy to the reader. As shown in Table 1, Bag of Words used the most support vectors and the impact of terms on price was generally more negative than positive.

Within this regression price estimate, we also included the StkZero variable. This is the price of the stock at the time the financial news article was released. The value presented, 0.9994 in the case of Bag of Words, refers to the weight assigned to this variable. Within the regression equation, there is also a constant and the weights of each support vector. As an abbreviated example using the Bag of Words representation, an SVR regression price estimate would be as follows:

$$0.9994\text{StkZero} + 0.0004 + 0.0062\text{dedicated} + 0.0048\text{refining} + 0.0038\text{schlumberger} + 0.0037\text{front} + 0.0035\text{planted} - 0.0024\text{regions} - 0.0025\text{aetna} - 0.0025\text{mid} - 0.0032\text{aep} - 0.0041\text{simmons}.$$

In this example, if the term *dedicated* exists within the news article, it is assigned a one, otherwise a zero, and so on for each financial news article. Consistent with Joachims, we used a binary representation, independent of the number of times the term appears within the article. The *term weight* variable in Table 1 refers to the aggregate weight that the article terms have on the price of the stock. In the Bag of Words representation, all article terms have a 0.0002 combined weight as it relates to predicted price. While this weight may appear insignificant, it was found that this weight is important in fine-tuning the price prediction system and readers are referred to (Schumaker, 2006) for further information.

From this regression equation, prices can be predicted for each stock. However, the scope of this paper is more interested in the terms that have the greatest influence on price. Tables 2 and 3 present the top 5 positive and negative article terms and their weights, respectively.

Bag of Words	Noun Phrases	Named Entities	Proper Nouns	Verbs
0.0062 dedicated	0.0053 deductions	0.0061 deepwater	0.0055 alliance	0.0043 planted
0.0048 refining	0.0045 schlumberger	0.0044 monsanto	0.0043 deepwater	0.0022 announcing
0.0038 schlumberger	0.0040 monsanto	0.0032 senate	0.0039 senate	0.0021 smaller
0.0037 front	0.0038 EOG Resources Inc	0.0032 XL Capital Ltd	0.0037 monsanto	0.0020 switched
0.0035 planted	0.0038 refining capacity	0.0031 medicare	0.0036 schlumberger	0.0016 earned

Table 2. Positive article term weights

Bag of Words	Noun Phrases	Named Entities	Proper Nouns	Verbs
-0.0041 simmons	-0.0047 Simmons Company	-0.0057 anadarko	-0.0032 chiron	-0.0022 hereto
-0.0032 aep	-0.0034 chiron	-0.0051 Simmons Company	-0.0032 800 million	-0.0016 reinvested
-0.0025 mid	-0.0032 medco	-0.0034 a year earlier	-0.0031 a year earlier	-0.0015 insures
-0.0025 aetna	-0.0032 brightpoint	-0.0034 800 million	-0.0028 phase	-0.0015 approx
-0.0024 regions	-0.0029 profit	-0.0029 chalmette	-0.0028 io	-0.0014 due

Table 3. Negative article term weights

These tables neatly segway into our second research question, *what influential terms appear across representations*? First, recall that the bag of words representation is simply a collection of terms minus the meaningless stopwords. So terms appearing in the more restrictive textual representations, also have a chance to appear in the bag of words representation if enough weight is assigned to them. From this, we note that *schlumberger* and *planted* appear prominently across textual representations as does *simmons* for the negative weights. This means that articles within our corpora containing the terms *schlumberger* and *planted* will experience a modest price increase while articles with the term *simmons* will experience price decreases.

CONCLUSIONS

From our investigation, we found that certain article terms can lead to positive or negative influences on stock price. Some of these terms were strong enough to appear in several different representations, implying that articles containing these terms were susceptible to price movement.

There are some limitations to this study that merit some discussion. First, the dataset used was during a compressed period of time and the results provided are indicative of the patterns observed within this dataset. To generalize these findings to other periods of time would provide a clearer picture of the term usage and their impact on stock prices. However, for the purposes of this paper, the period of time used was found to be sufficient. Second, because the period of time used was relatively stable, it would be worthy of future studies to investigate more tumultuous market activity to see how the results may differ.

REFERENCES

Camerer, C. & K. Weigelt 1991. Information Mirages in Experimental Asset Markets. *Journal of Business* 64(4): 463-493.

Fama, E. 1964. The Behavior of Stock Market Prices. *The Journal of Business* 38(1): 34-106.

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

Jelveh, Z. 2006. How a Computer Knows What Many Managers Don't. The New York Times.

Joachims, T. 1998. Text Categorization with Support Vector Machines: Learning with Many Relevant Features. *European Conference on Machine Learning*, Chemnitz, Germany.

Lavrenko, V., M. Schmill, et al. 2000a. Language Models for Financial News Recommendation. *International Conference on Information and Knowledge Management*, Washington, DC.

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.

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

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

Mittermayer, M. 2004. Forecasting Intraday Stock Price Trends with Text Mining Techniques. *Hawaii International Conference on System Sciences*, Kailua-Kona, HI.

Mowshowitz, A. 1992. On the Market Value of Information Commodities. The Nature of Information and Information Commodities. *Journal of the American Society for Information Science* 43(3): 225-232.

Platt, J. C. 1999. Fast Training of Support Vector Machines using Sequential Minimal Optimization. In *Advances in Kernel Methods: Support Vector Learning*, B. Scholkopf, C. Burges & A. Smola. MIT Press, Cambridge, MA, 185-208.

Raban, D. & S. Rafaeli 2006. The Effect of Source Nature and Status on the Subjective Value of Information. *Journal of the American Society for Information Science and Technology* 57(3): 321-329.

Schumaker, R. P. & H. Chen 2006. Textual Analysis of Stock Market Prediction Using Financial News Articles. *Americas Conference on Information Systems*, Acapulco, Mexico.

Schumaker, R. P. & H. Chen 2009. Textual Analysis of Stock Market Prediction Using Breaking Financial News: The AZFinText System. *ACM Transactions on Information Systems* 27(2).

Sekine, S. & C. Nobata 2004. Definition, Dictionaries and Tagger for Extended Named Entity Hierarchy. *Language Resources and Evaluation Conference*, Lisbon, Portugal.

Tay, F. & L. Cao 2001. Application of Support Vector Machines in Financial Time Series Forecasting. *Omega* 29: 309-317.

Technical Analysis 2005. The Trader's Glossary of Technical Terms and Topics. Retrieved Mar. 15, 2005.

Tolle, K. M. & H. Chen 2000. Comparing Noun Phrasing Techniques for Use with Medical Digital Library Tools. *JASIS* 51(4): 352-370.

Vapnik, V. 1995. *The Nature of Statistical Learning Theory*. Springer, New York.

Witten, I. H. & F. Eibe 2005. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, San Francisco.

Wuthrich, B., V. Cho, et al. 1998. Daily Stock Market Forecast from Textual Web Data. *IEEE International Conference on Systems, Man, and Cybernetics*, San Diego, CA.