

Analyzing Representational Schemes of Financial News Articles

Robert P. Schumaker

Information Systems Dept.

Iona College, New Rochelle, New York 10801, USA

rschumaker@iona.edu

Word Count: 2460

Abstract

The research presented here examines a predictive machine learning approach for financial news articles analysis using several different textual representations: Bag of Words, Noun Phrases, Named Entities, Proper Nouns and Verbs. Using the Arizona Financial Text System (AZFinText), which is tailored for discrete numeric prediction, we found that both Proper Nouns and Verbs had the highest predictive power with Closeness scores of 0.0341 and 0.0356 respectively. Named Entities had the best Directional Accuracy measure of 51.4% and Verbs had the highest trading return with 3.36%. By comparison, the standard textual representation of Bag of Words was found to perform the worst overall.

1 Introduction

Making accurate predictions of stock market behavior has always had a certain appeal to researchers. The difficulty has been the inability of a system to adequately model human trading behavior. To get around this limitation, systems either need to adapt to human trading tendencies which is addressed in quantitative trading, or systems need to trade in the absence, or limited presence, of human involvement. While this second point may seem implausible, computer systems have the ability to act and react faster than human counterparts. By building a system that can execute trades well in advance of human participation, behaviors are easier to model and hence are more predictable. While any serious quantitative system can model security fundamentals, these systems all lack one important component, the ability to adapt to changing conditions. By developing a system that can read financial news articles and make trading decisions on them well in advance of their human counterparts, the possibility exists to create a new type of financial trader that can adapt and outperform its colleagues.

In this paper we test several ways in which to represent financial news articles by their parts of speech and their impact on price prediction using the Arizona Financial Text System (AZFinText). We believe that combining certain textual representations with historic stock prices will yield improved predictability.

This paper is arranged as follows. Section 2 is an overview of literature concerning Stock Market prediction and textual representation techniques. Section 3 is 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 finishes with our conclusions and future directions for this area of research.

2 Literature Review

There are several theories regarding security forecasting. The first of which is the Efficient Market Hypothesis (EMH), where the price of a security is a direct reflection of all information

available and everyone has access to this information. In EMH, it is believed that markets are efficient and react instantaneously to new information by immediately adjusting the share price. The second major theory is Random Walk Theory where prices are expected to fluctuate randomly in the short-term. This theory is similar to EMH as both theories believe that all public information is available to everyone and that consistently outperforming the market is an impossibility.

From these theories, two distinct trading philosophies emerged; fundamentalists and technicians. In fundamentalist trading the price of a security is derived from the security's financial numbers and ratios. Numbers such as inflation, return on equity and price to earnings ratios can all play a part in determining the price of a stock. In technical trading it is believed that price movements are not random. However, technical analysis is considered to be an art form and as such is subject to interpretation. Researchers further believe that there is a window of opportunity where weak prediction exists before the market corrects itself to equilibrium (Gidofalvi, 2001). Using this small window of opportunity (in hours or minutes) 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

Financial news articles by themselves cannot easily be used by computer systems. We need to represent their important features in a machine-friendly form. One technique is the Bag of Words approach which has been extensively used in textual financial research (Gidofalvi, 2001; Lavrenko et al., 2000a). This process removes the meaningless stopwords such as conjunctions and declaratives from the text and uses the remaining terms as the textual representation. While this method is popular and easy to implement, it suffers from noise-related issues from seldom-used terms and problems with scalability. An improvement to this approach is to use only the Noun Phrases from the document which can adequately represent the important article concepts (Tolle & Chen, 2000). As a result, this technique uses fewer terms and can handle article scaling better than Bag of Words. A third representational technique is Named Entities, which extends Noun Phrases by selecting the proper nouns of an article that fall within well-defined categories. This process uses a semantic lexical hierarchy (Sekine & Nobata, 2004) as well as a syntactic/semantic tagging process (McDonald 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. This scheme allows better generalization of previously unseen terms and does not possess scalability problems associated with a semantics-only approach. A fourth representational technique is Proper Nouns. This approach functions as an intermediary between Noun Phrases and Named Entities. Proper Nouns can be defined as a subset of Noun Phrases by selecting specific nouns and also a superset of Named Entities without the constraint of pre-defined categories. This representation removes the ambiguity where certain proper nouns that could be represented by more than one named entity category or fall outside one of the seven defined categories. A fifth representational approach is to use the verbs and adverbs of the document. Certain information can be captured and retained by verbs that noun-only methods may miss.

Simply assigning a representational mechanism is not sufficient to deal with the scalability issues associated with large datasets. A common solution is to introduce a term frequency threshold that only represents those features appearing 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 cannot 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, 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).

3 Research Questions

Discrete prediction from textual financial news article is possible. Prior research has indicated that certain keywords can have a direct impact on the movement of stock prices and representational methods have almost exclusively used a Bag of Words approach. While Bag of Words has been promising, other textual representation schemes may provide better predictive ability, leading us to the research question: *what combination of textual representations are best suited to stock price prediction?*

4 System Design

For the experiment, we chose to use the AZFinText system which is composed of several major components. The first component tackles stock price by gathering price data in one minute increments from a commercially available stock price database. The second component deals with news articles by gathering them from Yahoo! Finance and representing them by their parts of speech; 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 data is gathered, AZFinText makes +20min price predictions for each financial news article through a 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 +20min price predictions for each financial news article encountered. Evaluations are then made regarding the effect of stock returns in terms of the textual representations used.

5 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. The process 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.

AZFinText’s predictions are evaluated using measures of Closeness, Directional Accuracy and a simple Trading Engine. Closeness, or how close AZFinText’s predicted +20min value was to the actual +20min price, is measured in terms of Mean Squared Error (MSE) where $MSE = (1/n)\Sigma(\text{Predicted} - \text{Actual})^2$ (Cho et al., 1999). Directional Accuracy is simply how often AZFinText was correct in predicting the price direction of the +20min stock (Gidofalvi, 2001). For a Trading Engine, AZFinText utilized a modified version of Lavrenko’s Trading Engine (Lavrenko et al., 2000a) that examines the percentage return of the stock. When a stock demonstrates an expected movement exceeding 1%, then \$1,000 worth of that stock is either bought or shorted and then disposed of after twenty minutes. This modified version differs from Lavrenko’s original design in regards to the dollar amount of stock bought. We further assume zero transaction costs, consistent with Lavrenko.

6 Experimental Findings and Discussion

Examining our research question of *what combination of textual representations are best suited to stock price prediction* led to the results in Table 1.

	Bag of Words	Noun Phrases	Proper Nouns	Named Entities	Verbs
Closeness	0.0442	0.0489	0.0443	0.0341	0.0356
Direction	50.1%	51.1%	51.4%	49.4%	48.4%
Trading	1.59%	2.57%	2.84%	2.02%	3.36%

Table 1. Textual Representation Results

From this table there are several interesting pieces to note. Named Entities had the lowest Closeness value of 0.0341 for any of the textual representation methods. The second lowest was Verbs at 0.0356. However these two were statistically equivalent, all others had p-values < 0.05. It would appear that AZFinText was able to make the most accurate price predictions given either the verbs of the document or the categorized nouns. In looking at the Directional Accuracy of AZFinText’s predictions compared to the market price direction, Proper Nouns had the highest accuracy rating of 51.4% followed by Noun Phrases at 51.1%, p-values < 0.05. While not as accurate in predicting precise price movements, the proper noun representation was more successful in determining price direction than chance alone. By contrast, both Named Entities and Verbs performed worse than chance. In the Trading Engine metric, Verbs performed best with a 3.36% return on investment versus 2.84% for the next highest representation, Proper Nouns, p-values < 0.05. If we were to analyze these results and focus on the relative performances of each representation, we would find that both Proper Nouns and Verbs performed the best while the de facto standard of Bag of Words had the worst overall performance.

7 Conclusions and Future Directions

The success of both Proper Nouns and Verbs in predicting stock prices is rather interesting. These two completely different parts of speech representations would be expected to have differing performances. The fact that they weren't indicates that both textual representations are strong in isolating the important prediction factors within the financial news article. Clearly more research is needed to isolate the extent of predictive ability that these representations have.

References

- Bishop, C. M. & M. E. Tipping 2003. *Bayesian Regression and Classification*. IOS Press, Amsterdam.
- Cho, V., B. Wuthrich, et al. 1999. Text Processing for Classification. *Journal of Computational Intelligence in Finance* 7(2).
- Gidofalvi, G. 2001. Using News Articles to Predict Stock Price Movements. 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. *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.
- 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.
- 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.
- Schumaker, R. P. & H. Chen 2006. Textual Analysis of Stock Market Prediction Using Financial News Articles. *Americas Conference on Information Systems*, Acapulco, Mexico.
- 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.
- 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.