An Analysis of Verbs in Financial News Articles and their Impact on Stock Price

Robert P. Schumaker
Iona College
715 North Ave
New Rochelle, NY 10801, USA
rob.schumaker@gmail.com

Abstract

Article terms can move stock prices. By analyzing verbs in financial news articles and coupling their usage with a discrete machine learning algorithm tied to stock price movement, we can build a model of price movement based upon the verbs used, to not only identify those terms that can move a stock price the most, but also whether they move the predicted price up or down.

1 Introduction

Predicting market movements is a difficult problem that deals mostly with trying to model human behavior. However, with the advent of quantitative trading systems its now easier to dissect their trading decisions. These systems are nearly instantaneous in their ability to make trades, but their Achilles heel is a reliance on human counterparts to translate relevant news into numeric data. This introduces a serious lag-time in trading decisions.

2 Literature Review

Information is fed into the market all the time. While some information sources can move a stock price, e.g., rumors and scandals; financial news articles are considered more stable and a form of its own commodity (Mowshowitz, 1992).

The first challenge of a textual financial prediction system is to manage the large amounts of textual information that exist for securities such as periodic SEC filings, press releases and financial news articles. These textual documents can then be parsed using Natural Language Processing (NLP) techniques to identify specific article terms most likely to cause share price changes. By automating this process, machines can take advantage of arbitrage opportunities faster than human counterparts by repeatedly forecasting price fluctuations and executing immediate trades.

Once financial news articles have been gathered, we need to represent their important features in machine-friendly form. We chose to implement a verb representation scheme which was found to be most predictive for financial news articles.

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 (Joachims, 1998). This technique not only eliminates noise from lesser used terms, but also reduces the number of features to represent. Once scalability issues have been addressed, the data needs to be prepared in a more machine-friendly manner. One popular method is to represent article terms in binary where the term is either present or not in a given article. 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 have been represented, learning algorithms can then begin to identify patterns of predictable behavior. One ac-

cepted method, Support Vector Regression (SVR), is a regression equivalent of Support Vector Machines (SVM) but without the aspect of classification. 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 System Design

To analyze our data, we constructed the AZFinText system. The numeric component gathers price data in one minute increments from a stock price database. The textual piece gathers financial news articles from Yahoo! Finance and represents them by their verbs.

For the machine learning algorithm we chose to implement the SVR Sequential Minimal Optimization function through Weka. This function allows discrete numeric prediction instead of classification. We selected a linear kernel and ten-fold cross-validation.

4 Experimental Design

For the experiment, we selected a consecutive five week period of time to serve as our experimental baseline. This period of research from Oct. 26, 2005 to Nov. 28, 2005 was selected because it did not have unusual market conditions and was a good testbed for our evaluation. We further 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. 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, and 10,259,042 stock quotations.

The first task is to extract financial news articles. The entire corpus of financial news articles are represented by their verbs in binary. If a particular verb is present in the article, that feature is given a 1, else a 0 and then stored in the database. To build a model, we first pair together the representational verb and stock quotation at the time the article was released, for each financial news article. This data is then passed to the SVR algorithm where a multi-dimensional price prediction model is constructed. This weighted model can then be dissected to determine the most relevant factors that can influence price movement.

5 Results and Discussion

From the trained AZFinText system, it was unsurprising that a majority of weight was placed on the stock price at the time the article was released and is consistent with prior observation where the article terms were found to be important and were used to fine-tune price prediction. Of the verbs, 211 were used by the system as support vectors. An abbreviated multi-dimensional price prediction model is as follows. The constants represent the weight given by the SVR algorithm and the verbs are binary, representing their existence within the financial news article.

$$0.9997\text{Intial\_Price} + 0.0045\textit{planted} + 0.004\textit{announcing} + 0.003\textit{front} + 0.0029\textit{smaller} + 0.0028\textit{crude} - 0.0029\textit{hereto} - 0.002\textit{comparable} - 0.0018\textit{charge} - 0.0015\textit{summit} - 0.0015\textit{green}$$

The five verbs with highest negative impact on stock price are *hereto*, *comparable*, *charge*, *summit* and green. If the verb *hereto* were to appear in a financial article, AZFinText would discount the price by $0.0029. While this movement may not appear to be much, the continued usage of negative verbs is additive.

The five verbs with the highest positive impact on stock prices are *planted*, *announcing*, *front*, *smaller* and *crude*.

References

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

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.

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

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