
Using News Articles to Predict Stock Price Movements

Győző Gidófalvi

Department of Computer Science and Engineering
University of California, San Diego

La Jolla, CA 92037

gvozo@cs.ucsd.edu

2001, June 15, 2001

Abstract

This paper shows that short-term stock price movements can be predicted using financial news articles. Given a stock price time series, for each time interval we classify price movement as "up," "down," or (approximately) "unchanged" relative to the volatility of the stock and the change in a relevant index. Each article in a training set of news articles is then labeled "up," "down," or "unchanged" according to the movement of the associated stock in a time interval surrounding publication of the article. A naïve Bayesian text classifier is trained to predict which movement class an article belongs to. Given a test article, the trained classifier potentially predicts the price movement of the associated stock. However, the efficient markets hypothesis asserts that this classifier cannot have predictive power. In careful experiments we find definite predictive power for the stock price movement in the interval starting 20 minutes before and ending 20 minutes after news articles become publicly available.

1 Introduction

According to the “efficient market hypothesis”, in financial markets profit opportunities are exploited as soon as they arise, hence stock prices follow a random walk and are extremely difficult to predict [5]. However, as was pointed out in [1], [2] and [4], in the financial market setting the task is rather to generate profitable action signals (buy and sell) than to accurately predict future values of a time series. As described in an earlier work [1], a usually less successful technical analysis tries to predict future prices based on past prices, whereas fundamental analysis tries to base predictions on factors in the real economy (e.g. inflation, trading volume, organizational changes in the company, demand for products or services offered by the company). As financial textual data (news articles) became available on the web, a new source of indicators appeared, which potentially could contain useful

information for fundamental analysis. The objective of this project is to analyze and extract such information, and derive numerical indicators from financial text.

2 Task, data, and system overview

Since profit opportunities in the stock market are present for only an extremely short period of time, high frequency information is essential to profitable trading strategies. The data that we have obtained from Lavrenko [1] contains both prices in 10-minute intervals and news articles with timestamps for 127 stocks for the time period ranging from 1999/11/14 to 2000/02/11.

Indicators can be of two types: those derived from textual data (news articles), and those derived from numerical data (stock prices). Since in [1] the extraction, analysis, and use of indicators from numerical data was widely explored, we chose to focus our research on indicators derived from textual data. We obtain indicators derived from textual data by learning a naïve Bayesian text classifier for higher-level, relative price movements of stocks. We then use this trained naïve Bayesian classifier to compute the probability for every new, stock-specific news article that that particular news article belongs to a class representing a particular movement class.

3 Indicators derived from textual data

As was mentioned in the previous section we derive a set of indicators from textual data using a naïve Bayesian text classifier. This derivation can be divided into the following subparts:

- Identification of movement classes,
 - Alignment of news articles
 - Scoring of news articles
 - Labeling of news articles
- Training a naïve Bayesian text classifier for the movement classes, and using posterior probabilities for each news article as indicators

3.1 Identification of movement classes

Unlike in the usual text classification framework in our task news articles are initially unlabeled. Defining classes and obtaining labels for training examples is crucial for any classification task. The following subsections describe in detail our approach to accomplish this task.

3.1.1 Aligning of news articles

According to our initial hypothesis news articles contain information that have an effect on stock prices. To evaluate this possible effect of a news article, for each article we define a time interval that we call the *window of influence*. The window of influence of a particular article d with a timestamp t can be characterized by lower boundary offset and an upper boundary offset from t . An offset is negative if $t + \text{offset}$ is prior to t . As an example figure 2 shows on the left an alignment with $[-20, 30]$ offsets and on the right an alignment with $[20, 40]$ offsets.

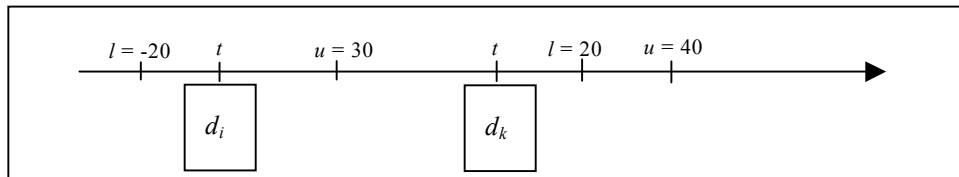


Figure 2: Illustration of alignments with different offsets. l and u are abbreviations for lower and upper boundary offsets and are measured in minutes.

Since the numerical data for the stocks contains price information only from 9:34 to 16:44 in ten-minute intervals for official trading days, we disregarded news articles from our data set that could have an ambiguous effect. In other words, we have disregarded news articles that were posted after closing hours, during weekends, or on holidays. Furthermore, a particular alignment puts additional limits on the articles that we include in our set. For example, the alignment shown on the right in figure 2 only allows articles to be included in our set if the posting time is between 9:14 and 16:04. For the 12 stocks selected for the three-month period this preprocessing reduced the total number of articles from 15,000 to approximately 5,500 depending on the particular alignment.

3.1.2 Scoring of news articles

One commonly used method to evaluate the performance of a particular stock is based on the volatility of the stock, which is known as the β -value. In short, this β -value describes the behavior or movement of the stock relative to some index, and is calculated using a linear regression on the data points $(\Delta index-price, \Delta stock-price)$. Hence a stock with a β -value of 1 has the characteristic that whenever the percent change for the index price is δ the percent change for the stock price is expected to be δ as well. Similarly a stock with a β -value of 2 has the characteristic that whenever the percent change in the index price is δ the percent change in the stock price is expected to be 2δ . Stocks with a β -value greater than 1 are relatively volatile, while stocks with a β -value less than 1 are more stable.

Since our numerical data does not include prices for the NASDAQ index, we approximate this value by the arithmetic average of the selected stock prices. Even though this approximation does not directly take into account relative volume information for weighting prices according to importance, weighting is indirectly present in the stock prices. In other words, we make the assumption that the relative difference in volume between two stocks is equivalent to the relative difference in price between the two stocks.

To eliminate the effects of the exponential change of stock prices we calculate the change on a log scale according to the following formula:

$$\Delta price(u, v) = \ln \frac{price(v)}{price(u)}.$$

We define m , the *movement* of a stock in a time interval $[u, v]$, as follows:

$$m(u, v) = \frac{\Delta sp(u, v)}{\beta} - \Delta ip(u, v),$$

where $\Delta sp(u, v)$ and $\Delta ip(u, v)$ represent change in price during the time interval $[u, v]$ for the stock and the index respectively. Using this movement measure a news article d with a timestamp t when aligned using offsets $[l, u]$ receives a score $m(t+l, t+u)$.

3.1.3 Labeling news articles

A movement of zero at a particular time and for a particular window of influence means that the movement of the stock at that time is as predicted or expected.

Similarly, a movement greater than or less than zero means that the movement of the price of the stock at that time is *respectively* better or worse than expected. We emphasize the word *respectively*, since according to our scoring method a news article may receive a positive score even if during the window of influence of the news article the change in the stock price is negative. Similarly a news article may receive a negative score even if the change in the stock price is positive. Our measure of movement is a relative one, which is not only based on the change in stock price, but is also based the change in the index price and our expectation of the stock's behavior to this change. We define three movement classes: upward movement (UP), downward movement (DOWN), and expected movement (EXP), according to the following rules:

$$mc(m) = \begin{cases} UP & m > \rho_{positive} \\ DOWN & m < \rho_{negative} \\ EXP & otherwise \end{cases}$$

Although this rule for labeling news articles may seem simple, the task at hand is highly non-trivial. As we pointed out earlier assigning correct labels to documents is essential to our classification task. To demonstrate the importance and difficulty of the labeling process consider a case depicted in figure 3. Let us assume that the true distribution of the news articles in the three classes is as shown above the axis. Here we assume that each class has an associated language usage, which is distinct from the others. In particular we would expect that words like *lost*, *shortfall*, or *bankruptcy* would occur more frequently in news articles that discuss downward movements in price, whereas words like *incept*, *propel*, or *peak* would occur more frequently in news articles that discuss upward movements in price. Setting our threshold values $\rho_{negative}$ and $\rho_{positive}$ to the values shown, in figure 3, results in a non-optimal or incorrect labeling. Some news articles that discuss downward movement in price of a certain stock are labeled as *EXP* in our model, and some articles that discuss expected movement in price are labeled as *UP* in our model.

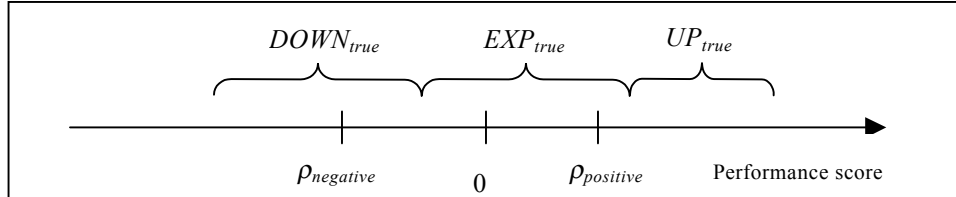


Figure 3: Implications of non-optimal labeling of news articles.

3.2 Learning of naïve Bayesian text classifier and extracting indicators

After we have defined our movement classes and assigned news articles to these classes our learning task can be phrased as follows. Given a news article d , we would like to predict the probability that a movement class c will follow. Using Bayes rule we can calculate this conditional probability as:

$$P(M = c | d) = \frac{P(d | M = c)P(M = c)}{P(d)}$$

After assuming the conditional independence of the words within a document, given a movement class, this can be rewritten as:

$$P(M = c | d) = \prod_{w \in d} \frac{P(w | M = c)P(M = c)}{P(w)}$$

This is the probability modeled by a naïve Bayesian classifier. We have chosen to use the Rainbow naïve Bayesian classifier package [3] for our text classification task. We state the exact options used for the classification in section 4.

After having trained a classifier for the movement classes, for each new news article d we calculate the posterior probabilities $P(M = c|d)$ for every movement class c from the set $\{UP, DOWN, EXP\}$ and use them as possible indicators. Hence for each news article we obtain three numerical indicators: $P(UP|d)$, $P(DOWN|d)$, and $P(EXP|d)$.

4 Evaluation and results

To evaluate our model for predicting movement in stock prices using indicators extracted from news articles, in the following sections first we describe the experimental setup, then we evaluate each of the processing steps performed in our classification task.

4.1 Experimental setup

We identify 12 stocks, which are from the same index (NASDAQ), and for which we have largest amount of news articles available to us. The ticker symbols for the selected stocks are: CSCO, SUNW, MSFT, ORCL, WCOM, YHOO, RHAT, DELL, AMZN, LU, EBAY, and INTC. For all experiments we use the price information and the news articles for the 12 stocks selected. We divide our data that contained information between 1999/11/14 and 2000/02/11 into a training and a test set, such that news articles time-stamped before 2000/1/10 9:34AM belong to the training set and news articles after that date belong to the test set. As a result of this division depending on the particular alignment used, we obtain 4,300-4,650 news articles in the training set and 1,300-1,650 news articles in the test set. Experimental results reported are all results of classifications performed on the test set.

After having tried several settings and options for the Rainbow text classifier we decide to use the Wittenbell smoothing method, assume uniform prior distribution of documents across the classes, remove stop words, use stemming, and use only the first 1,000 words with highest mutual information for the classification. Unless otherwise stated we apply these settings to all tests we perform.

4.2 Evaluation of movement measure and β -values

Since labeling of news articles partially depends on the scores assigned to them, the correctness of the volatility of stocks calculated based on an approximation of the index price is essential to our classification. In table 1 we show β -values derived using linear regression and the r^2 values for the linear regressions we perform.

| Stock | AMZN | CSCO | DELL | EBAY | INTC | LU | MSFT | ORCL | RHAT | SUNW | WCOM | YHOO |
|------------|------|------|------|------|------|------|------|------|------|------|------|------|
| Beta-value | 1.01 | 0.52 | 0.45 | 0.83 | 0.55 | 0.24 | 0.41 | 0.16 | 1.91 | 1.38 | 0.47 | 1.13 |
| r-square | 0.22 | 0.27 | 0.16 | 0.83 | 0.25 | 0.02 | 0.2 | 0.15 | 0.28 | 0.22 | 0.06 | 0.48 |

Table 1: β -values values calculated for individual stocks using linear regression and r^2 values for the regressions as a measure of fit.

We observe that some of the results from the linear regressions have very low r^2 values, hence the predicted value does not correctly model the actual change in

stock price. To visualize this we show the regression lines and that data for the regressions with lowest and highest r^2 values in figure 4.

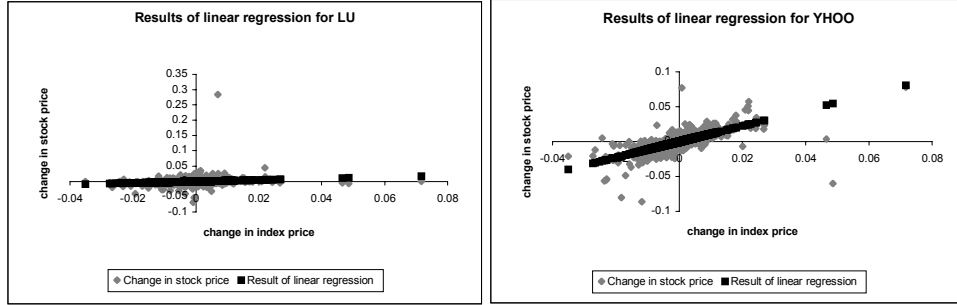


Figure 4: Charts show the data used to calculate the volatility of stock LU (left) and YAHOO (right) and their regression lines.

4.3 Evaluation of labeling

As we have discussed, an optimal labeling of news articles is essential to our classification task. We conduct several experiments with different labeling threshold values. In these experiments for various values of $\rho_{negative}$ and $\rho_{positive}$, and a fixed alignment with $[0, 20]$ offsets we evaluate classification results both in terms of accuracy of prediction and statistical significance. We evaluate the accuracy of our model, as defined in [6], and compare it against the best random predictor that places each of the news articles in the class, which has the highest class prior. To account for the disadvantage of our model due to the assumption of uniform prior distribution of news articles we conduct all experiment both assuming non-uniform prior and assuming uniform prior distribution of documents in our model. We also test the results of the classification against the null hypothesis, the hypothesis that the results of the classification are the result of random guessing. Results of the Chi-square hypothesis tests against the null hypothesis are shown in terms of z-scores. The results of the experiments are summarized on the charts of figure 5 and 6. We find that the best classification accuracy relative to the random predictor's prediction accuracy, and the highest statistical significance of our classification is at values $\rho_{negative} = -0.002$ and $\rho_{positive} = 0.002$. We fix our thresholds at these values for further experiments.

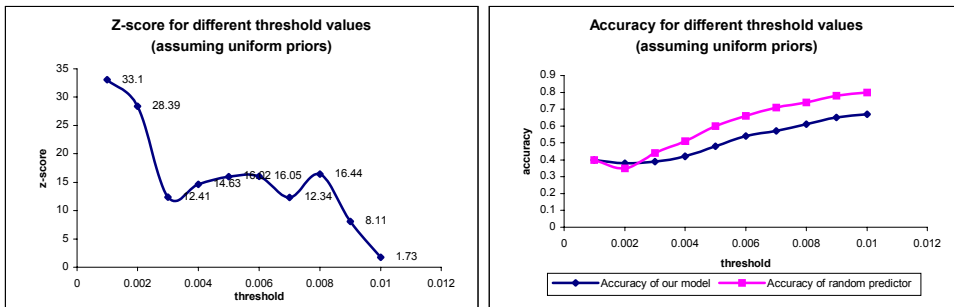


Figure 5: Results assuming uniform prior distribution. The chart on the left shows the z-scores of the Chi-square test against the null hypothesis for different threshold values. The chart on the right shows prediction accuracy of our model against the random predictor's accuracy for different threshold values.

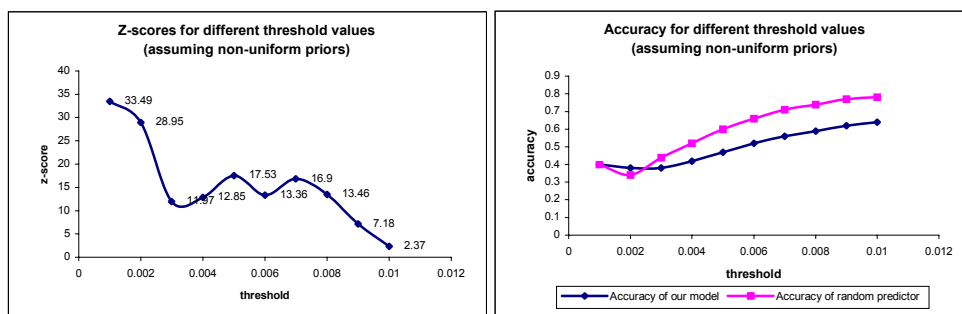


Figure 6: Results assuming non-uniform prior distribution. The chart on the left shows the z-scores of the Chi-square test against the null hypothesis for different threshold values. The chart on the right shows prediction accuracy of our model against the random predictor's accuracy for different threshold values.

4.4 Evaluation of alignments

To evaluate the effects of different alignments, we consider two basic types of alignments: alignments that assume that information contained in a news article influence the stock price before the time the news article was publicly available, and alignments that assume that a news article affects stock price only after it becomes publicly available. Figures 7 and 8 show the results of the experiments for a set of alignments for both alignment types. For all the charts in the figure the x-axis shows the window boundary offsets in minutes for the alignments and should be interpreted as follows. For negative window boundaries we implicitly assume that the upper boundary offset is set to zero and the negative boundary represents the lower boundary offset. Hence a data point at -30 represents a $[-30, 0]$ alignment. Conversely for the positive values of window boundaries we implicitly assume zero values for the lower boundary offsets. Hence a data point at +30 represents a $[0, 30]$ alignment. As in section 4.3 we test the both for classification accuracy and statistical significance for different alignments, and show the results in figure 7. The charts in figure 8 show precision and recall values as percentages for different alignments.

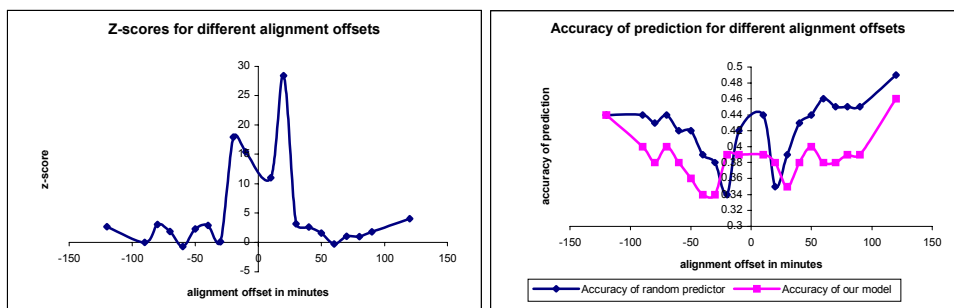


Figure 7: Effects of different alignments of news articles. The chart on the left shows z-scores of different classifications resulting from different alignments when tested against the null hypothesis. The chart on the right shows the prediction accuracy of our model against the prediction accuracy of a random predictor.

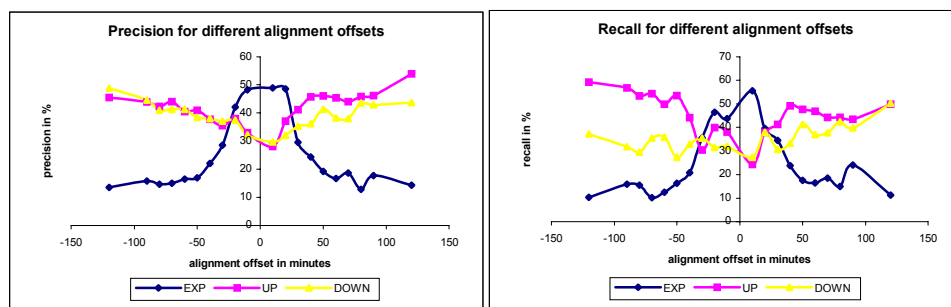


Figure 8: Effects of different alignments of news articles. The chart on the left shows precision values in percent, while the chart on the right shows recall values in percent for each movement class for the different alignments.

We find that the most significant classification results and best relative prediction accuracy occur for alignments $[-20, 0]$, and $[0, 20]$. Furthermore we find that as the window of influence is extended, classification results become less and less significant. This suggests that we have found a strong correlation between news articles and the behavior of stock prices from 20 minutes prior and to 20 minutes after news articles become publicly available. Our results disagree with the hypothesis in [5], which states that “predictable profit opportunities in asset markets are exploited as soon as they arise,” since we show a strong correlation between news articles and the behavior of stock prices from 20 minutes prior and to 20 minutes after news articles become publicly available.

5 Discussion of results

Results show that even though classification results are significant for the $[-20, 0]$, and $[0, 20]$ alignments the predictive power of the classifier is low. From the results we present in this work one can see that the predictive power of the classification is low. In the next two paragraphs we conjecture two possible reasons for this low predictive power.

Our analysis of using β -values used to determine the relative movement of a stock price to the movement in the index shows that the volatility measures obtained from an approximated index price does not in all cases model the relative movement of a stock correctly. The movement measure we use in this research is not necessarily an incorrect concept for measuring stock price behavior, but our analysis shows that the more realistic index price information is needed.

Yet another reason for low predictive power, as pointed out by Elkan, could be that important news is reported repeatedly in different news articles by different agencies, but presumably only the first article has an influence on the stock prices.

6 Conclusions

We present an approach for extracting numerical indicators for stock price behavior from financial text using two sources of information: stock prices and news articles about the companies whose stocks are under consideration. We align news articles to stock prices and score them based on the relative movement of the stock price during the window of influence. We identify three movement classes and train a naïve Bayesian text classifier for the movement classes. Even though we find that the predictive power of the classifier is low, we find a strong correlation between news articles and the behavior of stock prices from 20 minutes prior and to 20 minutes after news articles become publicly available. This result disagrees with the

efficient market hypothesis, and suggests that indicators for the prediction of future stock price behavior are present during only a short time interval, but possibly can be used in a profitable way.

7 Future work

We plan to further investigate the possible reasons for the low predictive observed in this project. Once we obtain a model with higher predictive power we plan to exploit the strong correlation found between news articles and movement types using a simple genetic algorithmic approach introduced in [1]. In this approach we plan to evolve simple trading strategies that are disjunctions of rules of the following form:

if stock indicator relation threshold then action stock,

where *indicator* is of type we derive in this work—i.e., posterior probability of a movement class given a news article, *relation* is taken from the set $\{<, >\}$, *threshold* is taken from the interval $[0.0, 1.0]$ with 0.01 increments, and *action* is taken from the set {buy, sell}. We plan to assign fitness to individual trading rules based on the profit generated by the rule.

Acknowledgments

We would like to thank for Victor Lavrenko and all the other members of his research group for providing both the numerical and textual data used in this project. Similarly, I would like to express my appreciation to Charles Elkan who gave helpful ideas and suggestions during the development of the project.

References

- [1] Elkan, C., (1999). Notes on discovering trading strategies.
- [2] Lavrenko, V., Schmill, M., Lawrie, D., Ogilvie, P., Jensen, D., & Allan, J. (2000) Mining of Concurrent Text and Time Series. KDD-2000: Workshop on Text Mining.
- [3] McCallum, A.K. (1996) "Bow: A toolkit for statistical language modeling, text retrieval, classification and clustering." <http://www.cs.cmu.edu/~mccallum/bow>.
- [4] Thomas, J.D. & Sycara, K. (2000) Integrating genetic algorithms and text learning for financial prediction. In A. A. Freitas, W. Hart, N. Krasnogor and J. Smith (eds.), *Data Mining with Evolutionary Algorithms*, pp. 72-75. Technical Report WS-99-06.
- [5] White, H., (1988) Economic prediction using neural networks: the case of IBM daily stock returns. In Proceedings of the 2nd Annual IEEE Conference on Neural Networks II, pp. 451—458
- [6] Yang, Y., (1999) An Evaluation of Statistical Approaches to Text Categorization. In *Informational Retrieval*, vol.1, pp.69-90