Using Structured Headlines to Predict Stock Prices

Masters Comprehensive Exam, Spring 2008
Department of Computer Science
University of Colorado at Boulder

ABSTRACT

Financial professionals have long tried to predict the movement of the stock market. Traditional techniques typically focus on past movement in stock price and ignore potentially valuable new information. In this study, we examine the effect of using natural language processing techniques to analyze business headlines from a novel corpus. The highly regular structure of our corpus is unique to our system and makes it easier to use shallow semantic information in our feature set.

Our results are promising, and we provide a reasonable metric for comparing future work in this area to ours. Given the enormous complexity of the stock market, the predictive value of our model is substantial and can be valuable in a more general predictive framework. We examine the strengths and weaknesses of our approach and provide suggestions for future work.

Word count: 6,521

1. INTRODUCTION

1.1 Problem Statement

For as long as markets have existed, investors have sought the best ways to turn information into profits. Stock markets have existed for nearly two millenia [19] and in many ways they represent an ideal avenue for exchanging information for profit, in that if an investor holds information that has implications for the future value of a stock's shares, he/she can act on that information by investing in the stock or selling its shares. Modern technology has transformed the financial world, making such information increasingly and immediately available and allowing for instantaneous execution of financial transactions. With investors continually searching for better ways to translate this increase in information into tangible profits, it has become apparent that it is impossible for an individual or even a group of individuals to fully digest the information available to today's investors.

1.2 Available Resources

Comprehensive time series data in the form of share price values over time is heavily used and widely available. Not only has the quantity of information regarding the stock market dramatically increased, the diversity of the types of information available to investors has also increased. For instance, with the click of a button (and sometimes even without that much effort), investors now have access to press releases, news articles, analyst reports, government regulatory changes, blogger opinions, court decisions, patent awards, and more. All of these sources of information have the potential to positively or negatively impact a stock's share price. Complicating the matter even further, much of this information is extraneous and has nearly no effect on share price, and it is often extremely difficult to distinguish what information does or does not have an effect on a company's share price.

Although computers have in many ways made this task more complex simply by making the information available, the power of modern computers has increased to the point where it is feasible to apply them to the task of interpreting information and turning it into actionable predictions. Understandably, there has been considerable interest in making use of this computational power and although doing so is by no means trivial, significant progress has been made along this front. There are a number of approaches to tackling this problem.

1.3 Approaches

Quantitative models rely solely on time series data and employ any number of statistical techniques analyzing previous numerical data to aid in the prediction of share price. Such techniques may include: regression analysis, least squares analysis, neural network learning mechanisms, and more. *Qualitative models* attempt to take advantage of outside information that is not present in the stock's share price history that might affect a stock's share price in the future. Such information may include: news articles, press releases, prices and availability of related commodities, and more. *Hybrid models* combine the strengths of quantitative and qualitative models into a single system in the hopes that bigger gains can be had by making use of multiple systems instead of just one or the other.

1.4 Outline

In this paper, we will outline the history of computational techniques for predicting equity performance, briefly cover relevant financial theory and describe the dominant computational techniques currently employed for use in equity price prediction. We will then identify the areas of work that yield additional gains over traditional techniques. Our analysis of past work is not comprehensive; however, there are certain discernable characteristics of existing and current work we can use to gain a better understanding of the task. Given the insights we glean from our initial analysis,

we then present a model for stock price prediction based on shallow semantic analysis and other natural language processing techniques. Given the sparse nature of the literature on applying unstructured qualitative data to the share price prediction task, we believe our model will contribute to the understanding of the task. We have compiled a unique corpus that to our knowledge has not yet been applied to the share price prediction task. By aligning our corpus with historical price movements, we have constructed a model that we will show performs better than a random baseline. We also propose a standard scoring metric for comparing future systems to our own that overcomes several problems and inconsistencies in methods for evaluation of similar systems. Finally, we suggest ways in which our insights can be applied to future work and integrated with existing work in ways that will inevitably yield better results.

2. FINANCIAL BACKGROUND

2.1 Share Price Factors and Motivations for Predictive Models

The stock market's sheer complexity has confounded traders looking for an efficient way to exploit market inefficiency for many decades. Throughout the course of history, traders have attempted to observe and understand any given equity's behavior with the goal of predicting the future value of that equity and subsequently securing profits. These efforts have in large part been thwarted by the numerous factors that affect a stock's performance. These factors include but are not at all limited to: company leadership, earnings, significant events occurring within the company and local, national, or global events, product demand, efficiency, dividend yields and small-scale and large-scale trends in valuation, profitability and contemporary cultural and political situations [2].

Although large quantities of data exist regarding national and global stock markets, only a small subset of the total sum of existing data provides specific information about the factors that led a share price to a given value. An even smaller subset of that data provides information that is readily apparent to the researcher, as many of the causal factors are rather subtle and difficult to decipher. Still, because uncovering any correlation even in this tiny subset of the total available data has implications for profit opportunities.

2.2 Efficient Markets Hypothesis

Since 1965, financial theory has been heavily influenced by Fama's [4] idea that a stock's current price is the result of the total impact of all of the above noted factors and potentially many other undiscovered factors. The logical conclusion of this concept is that all existing information about the equity at that moment in time is reflected in the share price and therefore its future direction cannot be known until new information becomes available. This notion is referred to as the efficient markets hypothesis. Given the assumption that the efficient markets hypothesis is true, as soon as a piece of information that might lead to profits becomes available, the opportunity is seized and the equity's price reflects the change at the moment the transaction is executed.

Nevertheless, there is an ongoing debate within the financial community (described by Yoo et al [22]) regarding the validity of the efficient markets hypothesis. The driving factor behind this debate is the fact that if the efficient markets hypothesis holds, purely quantitative analysis methods of analyzing past time series data should not yield significant correlations to future performance. Opponents of the efficient markets hypothesis point to quantitative models which have in fact been able to discern limited correlations from past time series data. A possible counterargument in favor of the efficient markets hypothesis holds that the hypothesis has value as an idealized model. Much like the ideal gas model in physics does not perfectly describe the behaviors of gases in reality, imperfections in the efficient markets hypothesis' predictions are evident in real world markets. It is impossible in a real world scenario for all information about an equity to be known by all interested parties at any given moment in time. This point of view would explain why numerous models have been able to identify minor correlations to future performance in past time series data [18].

2.3 Quantitative Models

Models for predicting future stock prices are not new. Indeed, creating such models is one of the primary goals of the field of econometrics, which formally dates to 1910 but has informally been studied since the advent of public stock markets [3]. Because a large amount of data regarding previous price movements is and has been available since that time, it has been an obvious target for analysis and a large body of work has emerged researching the potential for the use of past financial time series data in predicting future performance. This task was performed manually for many years by statisticians, economists, accountants, and other financial professionals attempting to predict future equity performance based on past performance data. Such models are at odds with a widely held notion that stock price functions are a form of random walk, that is, that future values of the price function are determined solely by position and a random variable [6]. These models are motivated by the fact that real world markets are not perfectly efficient.

Pettee's [15] analysis of such models for the time period from 1850-1930 shows that the performance of these manually-constructed models was collectively inconclusive at best and utterly ineffective at worst. Nevertheless, statisticians persisted in attempting to model stock price time series data using conventional statistical techniques including but not limited to regression analysis. Computational models for predicting and/or modeling equity prices emerged in the 1960s and 1970s, as accessibility to computers improved and researchers recognized their capacity to facilitate the same sort of manual models that had been applied to modeling and predicting price fluctuations (see for example [14]). While such models had limited success, it was over a decade later that a major success by White [21] generated renewed interest in the topic.

The continuing evolution of computing power and development of sophisticated artificial intelligence and statistical techniques led to a new generation of equity performance prediction models beginning in the late 1980s and continuing into the present. Increases in computing power made techniques like neural networks, support vector machines, classification trees, and other machine learning techniques tractable. In contrast to previous approaches, mechanisms

such as neural networks allow the computer to automatically detect relationships among different dimensions of the time series data. For example, White discovered that using feed forward neural networks uncovered small correlations in stock price fluctuations that went undiscovered by previous classical statistical techniques. In the years since White's work, number of studies have verified his finding that neural networks can detect correlations in financial time series data. Among the researchers inspired by the notion that neural networks were an effective way to model share price were Lowe and Webb [11] and Yoon [23][24], who demonstrated quantitatively that neural networks yielded results superior to traditional analysis.

The overall conclusion reached by Yoo et al [22] in a broad survey of existing techniques for this task is that quantitative models based on artificial neural networks and other similar techniques need to be combined with qualitative analysis in order to obtain the best results possible. While quantitative methods have been rather well-explored, there is currently much interest in investigating the possibility of using news articles and press releases as qualitative variables potentially influencing a stock's valuation. We will present a brief overview of several of these approaches and then present our own method for dealing with the same problem.

3. RELATED QUALITATIVE WORK

Fung et al [7] present a layered system, involving both time series analysis and qualitative natural language analysis. The time series data is analyzed using traditional regression techniques in order to identify "up" or "down" trends in the data. Once a model of stock trends is created, they apply a unigram approach to mining articles gathered from a Reuters corpus and aligning them to the trends discovered using their regression and clustering techniques. The corpus they use provides ticker symbols Reuters deemed relevant to the given article. Information content of unigrams is measured using term frequency and inverse document frequency, and that value is used to weight its importance as a feature. The authors use two support vector machine classifiers to determine whether or not an aligned article will trigger a rise or a drop in the share price. An article is deemed to have no effect on the share price if both classifiers come back with a negative result. The system is evaluated using a "buy-and-hold" test and an immediate buy, sell, or hold action as determined by the output of their model. The baseline they compare their work to is their own model, modified to wait a fixed amount of time prior to executing the given action. They ignore broker transaction costs and find that the immediate action method performs better on average than the fixed period transaction execution approach.

A key point to note from the work by Fung et al is that the regression techniques the authors use on the time series data is used only as a method for annotating the training set. The actual prediction mechanism is based only on the support vector machines employed to classify a given article. As such, it provides insight into only the effect qualitative data has on future prices. This kind of research is valuable because it provides analysis of the impact of qualitative data on prices in isolation. In doing so, it demonstrates that automated qualitative analysis can yield correlation to actual price movement, and also demonstrates the efficacy of the particular approach taken by the authors.

Mittermayer [12] presents a system termed NewsCATS, which uses press releases from PRNewswire [1] as the unstructured text input. His system excludes press releases which contain more or less than precisely one ticker symbol, which contain a reference to a non-NYSE and non-NASDAQ stock exchange, and which do not explicitly include a subject code. His system further limits the included press releases according to company size and trading activity, eventually ending up with 6,602 distinct press releases. A non-neutral movement is defined in this system as a 3% increase at any point in a 60 minute span following the release of a press release, with the stipulation that after that 60 minute period a 1% increase must be sustained. The classifier used to determine the category of a given press release is a modified support vector machine. Each press release is preprocessed as a set of unigrams and the 1,000 "most meaningful" (as defined by the term frequency-inverse document frequency) unigrams are selected as features for the classifier.

Mittermayer's system is evaluated using the weighted recall for each of the categories "good news", "bad news", and "no movers", for which he reports values of 58%, 54%, and 60% respectively. He also measures his performance in terms of the average profit per trade given the output of his system as compared to a trader applying random buying decisions when a press release becomes available to the public. To that end, he reports that NewsCATS yields on average 0.11% profit per trade while a random trader yields 0.00% profit per trade on average.

Gidofalvi [8], similar to Fung et al and Mittermayer, uses a unigram approach for feature selection. Like Mittermayer, Gidofalvi makes use of intraday trading data. Unlike Mittermayer, the intraday data he uses consists of trading volume information, which he uses as a proxy for share price. The corpus used is roughly 6,000 newswire articles over a one-year span relating to companies chosen based on the high number of articles written about them. The author reports precision and recall plots that are not statistically significant and attributes the model's failure to the use of trading volume instead of price data. Although this model did not yield significant results, his work is helpful in that it indicates that news articles yield to poor predictions of trading volume.

van Bunningen [20] was the only author we found who investigated the application of techniques beyond unigram analysis in the qualitative phase of the system. The author's system uses shallow semantic templates as a way of encoding information about news articles as features for a support vector machine (SVM) implementation. In addition, he also uses syntactic parses and WordNet [5] synsets in determining the feature set along with the typical TF-IDF[1] unigram approach to feature selection. He found that headlines (as opposed to entire articles) are better sources of information for a predictive model because they concisely encapsulate the main point of an article. He notes that the absence of intraday stock prices hindered the efficacy of his model.

[1]Term Frequency-Inverse Document Frequency. See [17] for an overview

4. EXPERIMENTAL MOTIVATIONS

A number of trends became clear from our search through the literature. The vast majority of attempts at predicting stock prices using natural language use a very simple unigram TF-IDF approach in conjunction with support vector machines. Although the authors make the motivations for this choice clear, the complete lack of diversity in approaches to the problem is a detriment to the advancement of the study of this topic. Our intuition from this observation is that it would be helpful to the field to examine an alternative approach to the problem.

Another trend we observed was that very few of the authors paid special attention to the quality of the natural language input used in their systems. An example of poor quality natural language input is that many news articles simply report price movements that have already happened. Even discounting those articles, there are many news articles that report the same event multiple times, which would throw off TF-IDF based approaches. As we have seen, all of the studies we examined made use of TF-IDF. Mittermayer's system used press releases as the sole natural language input to the system. While a press release can certainly have an immediate effect on a stock's share price, a large number of press releases are simply "marketing fluff" which tend to make a lot of noise about potentially insignificant events.

Finally, consistent evaluation of work in the area is difficult. Some authors evaluated their systems in terms of profit performance [12][7], others in terms of precision and recall [8], and still others using different methods. This unfortunate fact is an inevitable result of the complex and interdisciplinary nature of the task. Because insights from economics, statistics, finance, computer science, linguistics, and psychology are required for adequate performance on the task, the customary means for evaluating work within each discipline are applied according to the dominant background of the author. Furthermore, predictions of share price movements in much of the existing work in this area do not take into account larger scale price movements in the broader stock market indices. Our work is intended to address these issues.

5. METHODOLOGY

The primary intent and motivation for our experiment was to expand the scope of work in the task of using natural language qualitative data for share price prediction. To that end, we gathered a novel corpus that sidesteps many of the problems encountered by other authors in their search for a reliable source of qualitative information pertaining to stock movement. For preprocessing our corpus, we employed simple semantic analysis techniques [10] to emulate the strengths of the approach employed by [20] while also taking advantage of unigrams as the other authors did. Nevertheless, because we wanted to avoid the problems encountered by a TF-IDF approach to unigram selection, we attempt a selection process that differs from the other approaches using unigrams. Furthermore, we wish to standardize and streamline the process of evaluating systems such as ours, so we propose a metric for doing so that can be employed by other researchers. Our system architecture for this task appears in Figure 1.

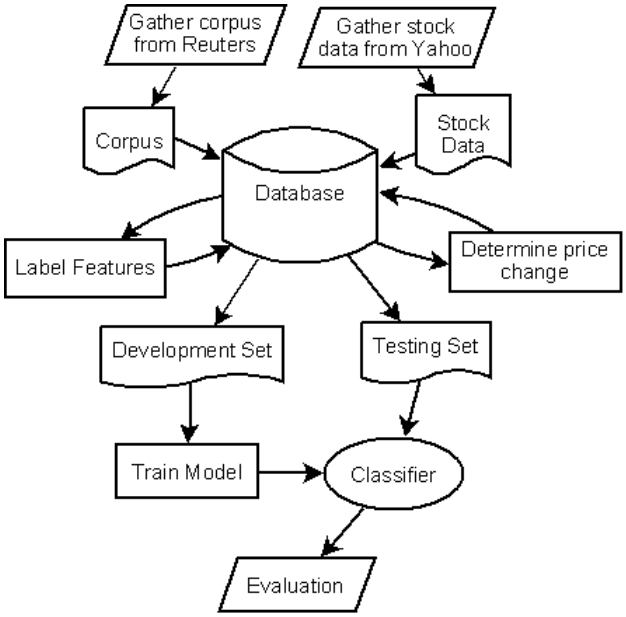

Figure 1: System architecture

We gathered a corpus comprising roughly 16,000 articles obtained from a service called Reuters Key Developments (RKD) [16], spanning the year between April 23, 2006 and April 23, 2007. RKD is a service provided by the media company Reuters with the stated intention of including only those articles that might have a significant impact on a given company. The editors filter out marketing fluff press releases and articles reporting only already observed stock movements. An added advantage is that because the news comes from a single aggregated source, there are no repeated articles. Inspired by van Bunningen, who observed that the headline of an article quite succinctly captures the entire article's primary sentiment, we looked only at the headlines of these articles. The articles we selected were those relating to the members of the Russell 1000 stock index during the aforementioned time period. We excluded those companies that were dropped from the index during that time period and those that did not have any news articles published about them. Of the approximately 15,000 articles remaining in our corpus after reduction based on the above mentioned selection criteria, we used a randomized training set of approximately 10,000 headlines and a randomized test set of approximately 5,000 headlines.

Another reason we chose to use RKD as our corpus is that the editors format the information content of the headlines in a highly regular fashion and using a consistent vocabulary, which is valuable for a machine learning approach. For example, all articles relating to company predictions as compared to analysts' expectations follow the form:

Company X issues Y guidance {*above,below,in line with*} *analysts' expectations*

where Y is the type of guidance being issued. The nature of our corpus is naturally suited to avoiding the problems we noted other authors have faced in their natural language

data. Manual verification of the predicate argument structure of this construction generated by the semantic processing package we used [10] confirms it generally yields accurate results and is therefore useful as a feature in our model. Similar examples are used for many of the 32 semantic features in our model. The semantic information used by our system is similar but simpler than that used by [20]. We believe that the simplifications still capture the semantic essence of the sentence without the need for added complexity. Figure 2 illustrates the structure and format of an example semantic template feature we used in creating our system. A critical aspect of this approach to preprocessing the natural language data is the fact that a unigram TF-IDF approach will not identify the acting agent in a transaction. For example, consider the following two sentences.

Company A acquires company B.
Company B acquires company A.

These sentences contain *exactly* the same unigrams; however, their meanings are clearly quite different. Our semantic processing techniques allow us to make these kind of fine-grained distinctions. To determine if the head entity is the company of interest, we compare its edit distance with the other tagged NPs in the headline against a database mapping ticker symbols to company names. Although it is possible that this method will not yield a correct answer, we find that it works well in most cases.

For modeling our data and for the purpose of classifying instances of the data into one of our three classes, we used an implementation of the Maximum Entropy (MaxEnt) [9] algorithm. Because nearly all of the authors in our survey of the literature made use of support vector machines as classifiers for the natural language input, we wanted to investigate the efficacy of using MaxEnt on this task. In previous work, Nigam et al [13] suggest that MaxEnt is a good choice for text classification. Data for MaxEnt consists of an instance, with N binary features. In other fields, MaxEnt is also known as multinomial logistic regression.

We attempted the problem multiple times, and we outline our progression as the system improved.

5.1 First attempt

Our first implementation broke down stock movement into the five classes seen in Table 1. The stock prices were not considered in relation to the overall movement of the Russell 1000 index. We trained the model on a non-normalized dataset that had the distribution shown in Figure 3. We used 50 features, of which 32 were shallow semantic structures. We used ten features to check for unigram combinations involving patents, as [20] noted that patent decisions and lawsuits can have a highly predictive effect on a company's share price. We also checked for the presence of guidance below, above, or in-line with analysts' estimates, as discussed above. Anecdotally, we found that how well a company's guidance matched analysts' expectations had a significant impact on short term stock prices. The net result of this approach was that our classifier learned to simply label everything as *no change*, so the accuracy was equivalent to the percentage of instances that were labeled *no change*. Obviously, this is a meaningless result, but it gave us insight into some of the problems with our system and inspired our second attempt.

Table 1: Original stock price movement classifications

Classification	Threshold
big gain	$\delta \geq 0.05$
small gain	$0.05 > \delta \geq 0.01$
no change	$0.01 > \delta > -0.01$
small loss	$-0.01 \geq \delta > -0.05$
big loss	$-0.05 \geq \delta$

5.2 Second attempt

Our second attempt resolved a number of issues present in our first implementation. We normalized the training data, ensuring that there were an equal number of instances of each class.[2] In doing so, we eliminated the possibility that our model would predict in a manner similar to the way it did on our first attempt because there is no bias to one class or another in the training set. In addition, we decided we should simplify the classification task by collapsing the five classes of Table 1 into three: *gain, no change, loss.* Our rationale for doing this was that any factors correlating a headline to a stock's share price is unlikely to be fine-grained enough to justify the decision to have five separate classes.

We also realized that by classifying absolute stock price movements, we were ignoring the influence of larger factors affecting an individual stock's price change. For example, if the entire market moved in a positive direction on a day when a headline corresponding to a negative share price movement appeared, the net effect might still be positive. To compensate, we computed the relative stock movement δ as follows:

$$\Delta\rho = \frac{\rho_{close} - \rho_{prev}}{\rho_{close}} \tag{1}$$

$$\delta = \Delta\rho_s - \Delta\rho_r \tag{2}$$

where $\Delta\rho$ is the fractional change of the current day's closing price, ρ_{close}, as compared to the previous day's closing price, ρ_{prev}, and where $\Delta\rho_s$ and $\Delta\rho_r$ are the results of evaluating Eq (1) for this stock and the Russell 1000 index, respectively. We empirically determined that considering $|\delta| < 0.037$ as the threshold for classifying a relative stock movement as *no change* yielded roughly equal quantities of articles classified in each of the three classes.

Our second attempt produced results representing a big improvement over our first attempt and gave us our first meaningful results (discussed below in the Results section). Because the feature set used was the same as in our first attempt, we were dissatisfied with its quality and decided to revise it, leading to our final system implementation.

5.3 Final system

Our final implementation is essentially the same as our second, with an additional 50 features. Although the semantic features used in our original feature set provided good information where they were applicable and gave our model

[2] roughly 3,200 each

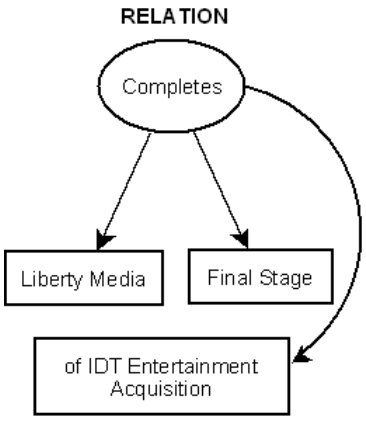

Figure 2: An example shallow semantic relation used by our system

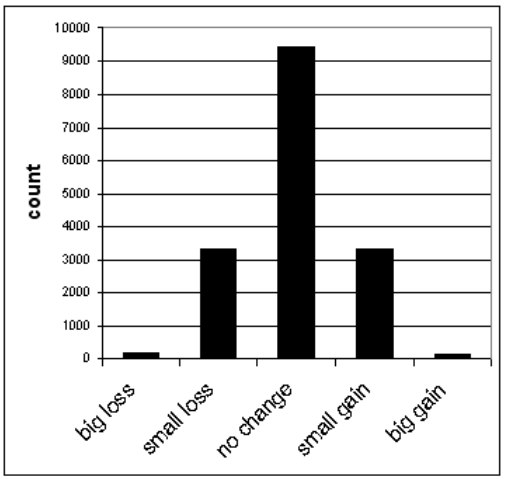

Figure 3: Distribution of price change classes for attempt 1

results better than random, a large number of very different training instances had identical feature vectors. In an effort to utilize unigrams (which other authors found helpful), we decided to examine the effect of using unigrams selected from the top 500 unigrams by frequency as features. We discarded unigrams with little content information such as determiners and prepositions and those that were part of company names. Using domain knowledge, we selected the unigrams that we felt would have the largest impact on stock price. This approach is in contrast to TF-IDF approaches in that it pays attention only to term frequency. The motivation for this approach is the assumption that the RKD team did a good job of identifying those articles which had the greatest impact on share price, and therefore that the most common terms will be the ones with the most information content.

6. EVALUATION

We measured the performance of our first two systems only in terms of precision (accuracy) over the entire instance set:

$$accuracy = \frac{|correct|}{|instances|} * 100\% \quad (3)$$

We observed artificially high accuracy values (∼58%) for our first system (as noted earlier) simply because the system labeled all instances as *no change*. Our second system used a normalized training set and therefore produced a meaningful accuracy value; however, we were not satisfied with the accuracy metric because it ignores the significance of recall. The metric we used to evaluate our final system remedies this deficiency, but because we were immediately aware of the flaws of our first two systems, we did not attempt to measure their performance according to the final metric. The results for our final implementation appear in Table 2. The baseline measurement we and other authors use is random selection of a target class. van Bunningen provides the following example scenario under which it is problematic to use precision and recall to evaluate a system compared to the random baseline: Consider target classifications C_1, C_2, and C_3 with 100 instances each. If C_3, accounts for 3 of the instances, a model precision of .30 is a significant improvement over a random baseline, which would achieve a precision of roughly 0.09, assuming it labels 33 of the 100

Table 2: Final implementation results

Measure	Value
α	1.074
N	10
|devset|	10,882
|testset|	5,442
f_{gain}	0.318
f_{loss}	0.376
$f_{nochange}$	0.369
$\overline{f}$	0.354

instances as C_3:

$$precision = \frac{|correct|}{|correct| + |incorrect|} \quad (4)$$

Substituting our example:

$$P_{C_3} = \frac{3}{3+30} \approx 0.09$$

We derived a method for circumventing the issues discussed by van Bunningen to make precision and recall meaningful in the context of this problem. By performing N runs using random classifications and recording the precision and recall for the random runs, it is possible to normalize the f-measures produced by the model and the random runs to produce an *improvement ratio* α as follows:

To calculate the f-measure, we need Eq (4) and recall:

$$recall = \frac{|found|}{|available|} \quad (5)$$

Then the f-measure is the harmonic mean of Eq (4) and Eq (5):

$$f = \frac{2 * (precision * recall)}{(precision + recall)} \quad (6)$$

Averaging f-measures over all target classifications:

$$\overline{f} = \left(\sum_{i=0}^{|C|} f_{C_i} \right) \left(\frac{1}{|C|} \right) \quad (7)$$

Table 3: Confusion matrix for final implementation

	Predicted		
Actual	*gain*	*loss*	*no change*
gain	540	673	792
loss	483	707	679
no change	371	520	687

Then averaging over N runs:

$$\overline{\overline{f}} = \left(\sum_{j=0}^{N} \overline{f}_j\right)\left(\frac{1}{N}\right) \tag{8}$$

Finally:

$$\alpha = \frac{\overline{\overline{f}}_{model}}{\overline{\overline{f}}_{random}} \tag{9}$$

Given $N = 10$, our final model produced $\alpha \approx 1.074$, indicating that our model performs approximately 7.4% better than random on average on an unknown test set. For reference, the confusion matrix appears in Table 3. It seems that the model had the most difficulty in correctly determining headlines in the class *no change*. This fact intuitively makes some sense: it seems less likely that a headline with no impact on stock price would have more distinguishing features than one with a large impact on a stock's price. Still, this is just conjecture and a more rigorous analysis of our system might yield other reasons why our confusion matrix takes the shape it does. Our final system is an improvement over our previous two attempts in a number of qualitative respects. Still, it is not at all the final word in research in this area and it is only intended to provide a solid starting point for future researchers intending to expand the scope and variety of work in the share price prediction task. We discuss the implications of our work and possibilities for future work in the next sections.

7. CONCLUSIONS

In summarizing our work, we return to the key observation put forward by Yoo [22] we described in the introduction. Being that quantitative share price prediction systems based on neural networks are better understood and more well-developed than systems making use of qualitative data including natural language event data, the next major developments in this area will come from the integration of the two approaches in the form of hybrid models. Although it would likely be profitable to immediately combine our model with a quantitative model, we believe that more work needs to be done in the qualitative area alone. Due to the complexity of coming up with a qualitative model with good predictive capabilities, it is important to isolate and standardize work in this area before moving on to the development of hybrid models.

With that goal in mind, we believe our results represent a significant step in the right direction. We have demonstrated that a qualitative system can perform significantly better than random. We have proposed a standard scoring metric we hope will be employed by future researchers. We have also made a number of improvements over previous work, and we have contributed to the literature by employing techniques not previously applied to this task. As noted before, it is difficult to compare our results with the other relevant work in the field. van Bunningen's system most closely resembles ours and it examines only pharmaceutical and biotech companies. A detailed comparison would therefore be meaningless, given that it targets only a particular subdomain of the problem we tackled. Furthermore, van Bunningen provides his results only in the form of confusion matrices, making them difficult to directly compare to our results or to a random baseline. Other authors' systems differ from our own in even more ways.

Despite the difficulty in comparing our work to other authors' work, our system addresses a number of potential deficiencies of previous systems and we invite comparisons of future systems to our own. The system we developed handles a task more general than the work by [20], and eliminates sources of variability not considered by the other authors (for example, the movement of the entire market on the same day a headline is released). We show our model to have predictive power substantially better than random, and given the complexity of the stock market, we are quite pleased with this result. Still, we are not at all convinced that we have even approached the upper bound of the potential for using headlines as stock price predictors. By introducing a single relevant metric, α, we hope to invite future comparisons to our work.

8. FUTURE WORK

We have only minimally investigated the quality of the features we used. An obvious next step for future work would be to empirically determine those features which give the most information to the model. An unsupervised feature discovery algorithm may lend itself well to this task; however, given the highly regular structure of our corpus, we believe it is possible to build a superior model using supervised techniques. In our desire to investigate the possibility of using an approach other than TF-IDF to unigram selection, we may have also overlooked some of the strengths of using that approach. Because the intent of TF-IDF is to uncover the terms that yield the most informational content, it may well be useful to include its features in future models.

It would also be useful to re-examine this topic in collaboration with someone with a financial background. Although we find that the use of standard natural language processing techniques is sufficient to perform better than random on this task, in-depth knowledge of the financial world might be of significant benefit to the model. Furthermore, we would hope future systems would involve other qualitative information specific to a given domain. For example, in evaluating the future performance of an airline stock, new developments in the price of oil can be considered significant. By combining the strengths of analysis of financial markets using neural networks and other domain-specific knowledge, it ought to be possible to create a model with very significant predictive power.

Another obvious extension to our work would be to introduce intraday stock prices to our model. It is possible that a news article may have its greatest effect on a stock's price within a small time frame. For example, if a news article is released in the middle of a trading day, by the time the

market closes, the effect of the news item may not be apparent in the stock's price. Fine-grain intraday data would enable us to capture these trends which currently go undetected in our model. The exact amount of time during which a news article impacts a stock is unknown. In our model, we assume this amount of time is one day or less, and this is a rather arbitrary assumption. Where intraday data would reveal the impact of a news article on a very short time scale, it is also possible that a news article may affect a stock's price over a longer period than that examined in this paper. Another possibility for future research is to empirically determine the optimum timeframe to monitor a stock's price for effects from a news article, starting with short time frames and intraday data, and expanding the window to several days. Integrating the results of such research into the model would inevitably yield a performance increase.

Our model simultaneously addressed multiple potential shortcomings of previous work, including but not limited to: the lack of semantic information, normalization against larger trends, and the lack of a well-structured corpus of natural language data. It would be informative to conduct a formal, controlled analysis of the individual components of our model to determine the optimal combination. The outcome of such an experiment would surely yield additional insights and foster growth in multiple areas related to the topic of this paper.

9. ACKNOWLEDGEMENTS

We extend our sincerest thanks to Arthur van Bunningen for his valuable insights and for giving us access to the innards of his system. The knowledge we gained from analyzing his system greatly assisted us in completing this work.

10. REFERENCES

[1] Pr newswire: News distribution, targeting, and monitoring. More information available at: www.prnewswire.com.

[2] R. S. Bower and D. H. Bower. Risk and the valuation of common stock. *Journal of Political Economy*, 77(3):349–62, May/June 1969.

[3] J. Eatwell. *Economic Development: The New Palgrave*. W.W. Norton, New York, NY, USA, 1995.

[4] E. Fama. Random walks in stock market prices, 1965.

[5] C. Fellbaum. *Wordnet: An Electronic Lexical Database*. Bradford Books, 1 edition, 1998.

[6] W. Feller. *An Introduction to Probability Theory and Its Applications*, volume 1. John Wiley and Sons, New York, NY, USA, 3 edition, 1968.

[7] P. C. Fung, G. X. Yu, and W. Lam. Stock prediction: Integrating text mining approach using real-time news. In *Proceedings of the IEEE International Conference on Computational Intelligence for Financial Engineering*, pages 395–402, 2003.

[8] G. Gidofalvi. Using news articles to predict stock price movements. Technical report, University of California, San Diego, 2001.

[9] Z. Le. Maximum entropy modeling toolkit for python and c++, 2003. Available at: homepages.inf.ed.ac.uk/s0450736/maxent_toolkit.html.

[10] H. Liu. Montylingua: An end-to-end natural language processor with common sense, 2004. Available at: web.media.mit.edu/∼hugo/montylingua.

[11] D. Lowe and A. Webb. Time series prediction by adaptive networks: a dynamical systems perspective. *Radar and Signal Processing, IEE Proceedings F*, 138(1):17–24, Feb 1991.

[12] M.-A. Mittermayer. Forecasting intraday stock price trends with text mining techniques. In *HICSS '04: Proceedings of the Proceedings of the 37th Annual Hawaii International Conference on System Sciences (HICSS'04) - Track 3*, page 30064.2, Washington, DC, USA, 2004. IEEE Computer Society.

[13] K. Nigam, J. Lafferty, and A. McCallum. Using maximum entropy for text classification, 1999.

[14] Nimrod, Vance L. and Bower, Richard S. Commodities and computers. *The Journal of Financial and Quantitative Analysis*, 2(1):61–73, mar 1967.

[15] E. W. Pettee. Long-term commodity price forecasting-1850 to 1930. i. *The Journal of Business of the University of Chicago*, 9(2):95–113, apr 1936.

[16] Reuters. Key developments. More information available at: http://www.reuters.com.

[17] G. Salton and M. J. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, Inc., New York, NY, USA, 1986.

[18] M. Schumann and T. Lohrbach. Comparing artificial neural networks with statistical methods within the field of stock market prediction. *System Sciences, 1993, Proceeding of the Twenty-Sixth Hawaii International Conference on*, iv:597–606 vol.4, 5-8 Jan 1993.

[19] B. M. Smith. *A History of the Global Stock Market: From Ancient Rome to Silicon Valley*. University of Chicago Press, Chicago, IL, USA, 2004.

[20] A. van Bunningen. Augmented trading: From news articles to stock price predictions using syntactic analysis. Master's thesis, University of Twente, Enschede, Netherlands, 2004.

[21] H. White. Economic prediction using neural networks: the case of ibm daily stock returns. In *Neural Networks, 1988., IEEE International Conference on*, pages 451–458 vol.2, 1988.

[22] P. D. Yoo, M. H. Kim, and T. Jan. Machine learning techniques and use of event information for stock market prediction: A survey and evaluation. In *CIMCA '05: Proceedings of the International Conference on Computational Intelligence for Modelling, Control and Automation and International Conference on Intelligent Agents, Web Technologies and Internet Commerce Vol-2 (CIMCA-IAWTIC'06)*, pages 835–841, Washington, DC, USA, 2005. IEEE Computer Society.

[23] Y. Yoon and G. Swales. Predicting stock price performance: a neural network approach. *System Sciences, 1991. Proceedings of the Twenty-Fourth Annual Hawaii International Conference on*, iv:156–162 vol.4, 8-11 Jan 1991.

[24] Yoon, Youngohc, Swales, George, Jr., and Margavio, Thomas M. A comparison of discriminant analysis versus artificial neural networks. *The Journal of the Operational Research Society*, 44(1):51–60, jan 1993.