News and Trading Rules

James D Thomas

CMU-CS-03-123

January 2003

School of Computer Science
Computer Science Department
Graduate School of Industrial Administration
Carnegie Mellon University
Pittsburgh, PA 15213

Thesis Committee:
Katia Sycara, Chair
Andrew Moore
Bryan Routledge
Blake LeBaron

Submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

This research is sponsored by the Office of Naval Research (ONR) under contract N000140210438 and contract N000149611222, by the US Air Force (USAF) under grant F306029820138 and grant F496200110542, and the National Science Foundation (NSF) under grant IIS-9612131 and grant IIS-9712607. The views and conclusions contained in this document are those of the author and should not be interpreted as representing the official policies, either expressed or implied, of the ONR, USAF, NSF or the U.S. government.

This thesis is dedicated to Ednah Thomas, who showed me language; and Victor Yonash, who showed me computers.

Abstract

AI has long been applied to the problem of predicting financial markets. While AI researchers see financial forecasting as a fascinating challenge, predicting markets has powerful implications for financial economics – in particular the study of market efficiency. Recently economists have turned to AI for tools, using genetic algorithms to build trading strategies, and exploring the returns those strategies generate of evidence of market inefficiency.

The primary aim of this thesis is to take this basic approach, and put the artificial intelligence techniques used on a firm footing, in two ways: first, by adapting AI techniques to the stunning amount of noise in financial data; second, by introducing a new source of data untapped by traditional forecasting methods: news.

I start with practitioner-developed technical analysis constructs, systematically examining their ability to generate trading rules profitable on a large universe of stocks. Then, I use these technical analysis constructs as the underlying representation for a simple trading rule leaner, with close attention paid to limiting search and representation to fight overfitting. In addition, I explore the use of ensemble methods to improve performance. Finally, I introduce the use of textual data from internet message boards and news stories, studying their use both in isolation as well as augmenting numerical trading strategies.

Acknowledgments

I have to beg for patience from the reader. I have gone through many twists and turns, false starts and dead ends on my way to writing this thesis – and I have people to thank at every point along the way.

Allow me to start at the beginning. This thesis is dedicated to two people – my grandmother, Ednah Thomas, and my sixth-grade teacher Victor Yonash. My grandmother opened my eyes to language. Mr Yonash showed me my first computer – I can still remember walking into his room after class and seeing the Apple][. Both of them changed my life, and in a very real way, planted the seeds of the research in this thesis.

Much of my academic career has alternated between thinking about language and thinking about artificial intelligence. At MIT, I studied language. When I left to come to the CMU school of computer science, I thought I'd left language behind. It is one of the great graces of my life that I have come full circle and worked on AI and language together.

At Wisconsin I have to think Kyle Johnson for introducing me to real linguistics, and Malcolm Forster for shaping my thinking about the entire enterprise of science.

At MIT, I have to thank Ken Wexler – who saw something in me, God knows what, to want me as his student, and especially Ted Gibson, who took me under his wing when my real research got under way. I also thank my class cohort – Joshua Tennenbaum, Emo Todorov, Jenny Ganger, Cristina Sorrentino, and Stephen Lines. Although I left MIT early, I think of them as my peers and friends more than anyone I've meet since. But there were others at MIT – Zoubin Gharamani, John Houde, John Kim, Flip Sabes, Suzie Johnson – whose friendship and conversation made my days in E-10 far more interesting than they had any right to be.

At CMU, first and foremost I have to thank my advisor, Katia Sycara, who showed incredible patience and support for me over the years. My research sometimes carried me pretty far afield – but she always trusted that I knew what I was doing, and provided perceptive guidance and criticism. I also have to thank Bryan Routledge, who took my naive enthusiasm about economics and shaped it into something useful to the discipline. The other members of my committee, Andrew Moore and Blake LeBaron, kept me honest and provided rigor and invaluable direction.

The list of people who befriended me and helped keep me sane in my over six years at CMU is enormous. My officemates – especially Henry Rowley, Bwolen Yang, Dennis Strelow, Sanjit Seisha, and Ted Wong – not only put up with stress, bad habits, and frequent naps, but also provided advice and much needed company. My landlords, Charlee Brodsky and Mark Kamlet (and, of course, Gillian and Ally), gave me a home for 7 years and delaying long-standing plans for their house for a few extra months as I rushed towards completion. WRCT provided a place to blow off some late night steam. My fellow grad students, especially Matt Glickman, Phoebe Sengers, Frank Dellaert, Matt Siegler, Daniel Teitelbaum, Belinda Thom, Chris Leger, Rosie Jones, Alex Gray, and Joseph O'Sullivan, provided friendship and an often exhilarating intellectual environment. Faculty, like Simon Penny, John Miller, Roni Rosenfeld, took my sometimes odd ideas seriously. I need to reserve special mention for two people: Corey Kosak, for the deep (if occasionally muddy) footprints he left on my psyche, and Jeff Smith, for helping make grad school one of those experiences you talk about or the rest of your life.

From my year-long sojourn in California at the now defunct Codexa, I have to thank David Leinweber, whose passions for exactly the research questions this thesis asks led him to found Codexa, and led me to California. It's hard for me to imagine my time there without Sharon Deprano and Gary Ghazarian. Rahim Ezmailzadeh showed me that life in the private sector was livable, and that research questions in finance were no less interesting for being done by corporations. He served as a de facto advisor away from home, providing real-world guidance that has heavily shaped my thesis research. Back in Madison, Joel Gratz provided assistance invaluable to the completion of this thesis. And Linda McAllister, though far away in body, was always close in spirit, and gave me the advice and support that only one who has just finished wrestling with her own dissertation can.

Finally, I have to thank my family. My grandmother Ednah Thomas, who died a month after I arrived in Pittsburgh – my only sorrow in graduating is that she will not be there to watch. My grandparents Fred and Blanche Zellmer, for their love and support. My sister coincidentally moved to Pittsburgh during the home stretch of my thesis. Her presence – five blocks away from me – was one of the great pleasant surprises of my life, and made my last year in Pittsburgh that much more enjoyable. And finally my parents, William and Kathryn Thomas, for far more than I have space to detail here – they haven't always known where I was going, but they always had faith that I would get there. I'd like to believe I've proved them right.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

> "A Modus–what is that?"
>
> "It is a legend in sporting circles, Dr. Mallory. A Modus is a gambling-system, a secret trick of mathematical Enginery, to defeat the odds-makers. Every thieving clacker wants a Modus, sir. It is their philosopher's stone, a way to conjure gold from empty air!"
>
> "Can that be done? Is such an analysis possible?"
>
> *William Gibson & Bruce Sterling, The Difference Engine [92]*

Predicting financial markets has long been a dream of Artificial Intelligence. Beyond that, almost anyone who's ever looked at the jagged lines of financial data has searched for patterns.

AI researchers have long assumed that the patterns were there, and it was only a matter of applying the right algorithms. Financial economists, on the other hand, have long understood just how tricky the data is, just how easy it is for the noise in financial data to trick and swallow good algorithms whole.

Recently, these two camps have started to inch towards each other: economists have started to draw techniques from AI in their explorations of market efficiency. However, there is still much ground to cover: economists customarily do not have the training to possess the full arsenal of artificial intelligence techniques at their disposal, and artificial intelligence researchers tend to ignore the near-pathological noisiness of financial data.

Into this promising situation comes a new source of data: news. Most examinations of market inefficiency on both sides of the fence have focused on numerical data generated by markets themselves: past prices, trading volume, highs and lows. But what really moves markets is *events in the world*, for which numerical data is a poor proxy.

With the notable exceptions of event studies and some keyword studies, the impact of news stories on market behavior has largely gone unexplored. That is about to change.

The explosive growth of the web has not left financial data behind; financial news is plentiful on the web – both from established suppliers of financial news such as Reuters, and newer, web-based actors such as the Motley Fool. Include the informal but copious text from stock-related discussion boards, and there is a large data set of text financial data that is easily findable, harvestable, and archiveable for free.

In Artificial Intelligence there is a growing research tradition devoted to the automated analysis of text data, most importantly the classification of text data. These techniques allow for computers to take news stories as inputs and output classifications. These classifications can be defined more or less arbitrarily – including operational criteria based on stock prices – and can easily be integrated into quantitative frameworks.

Combine these two developments and it becomes possible to build computer programs that take in news stories and integrate them with traditional quantitative trading rules.

There is no doubt in my mind that this approach will, in time, revolutionize efforts to apply machine learning to the problem of financial forecasting.

I hope that this thesis provides some of the first steps of this path.

1.1 Organization

Although the end goal of this thesis is the integration of news data with more traditional machine learning methods, first, there is much groundwork to be laid, interesting in and of itself.

This thesis is organized into four substantive chapters. In the rest of this chapter, I present the motivation for this work, as well as a description of the high level organization of the thesis.

Chapter 2 sets up a framework for examining technical analysis trading rules and systematically examines their success over a 5-year data set over over 1700 stocks.

Chapter 3 builds on this work, by using the technical analysis statistics explored in chapter 2 as the fundamental representation of a genetic programming based trading rule learner. I concentrate on exploring issues of overfitting through representation and search, and conclude that the problem of learning trading rules over financial markets is so much noisier than traditional domains of application for machine learning that a ground-up re-examination of the role of representation and search are necessary to fight overfitting.

Chapter 4 explores textual data – internet message boards – although purely as a source of numerical data, by measuring the volume of message board traffic.

Chapter 5 is the culmination of this thesis – the integration of news headlines with the trading rule leaner. I propose a comprehensive ontology breaking down the kinds of financially news stories stories into a set of sensible categories. Then, I hand-build classifiers for a large subset of the ontology's categories, allowing for analysis of the market impact of various news categories, and the integration of the news data with the trading rule learner developed in chapter 3.

Finally, chapter 5.10.1 concludes, summarizes and points to future directions.

1.2 Motivation: The Efficient Market Hypothesis

The motivation of this thesis is the efficient market hypothesis. Although there are many ways to approach the efficient market hypothesis, the intuition behind it is simple: markets efficiently process all relevant information into a single price. In principle, past data cannot be used to predict future prices in capital markets.

To computer scientists, this seems implausible: of course there should be patterns in past data that are usable to predict the future – there's useful patterns in all data.

But, there are strong theoretical reasons for believing that markets are efficient. The theoretical justification goes something like this: if there were patterns in past data one could use to predict future prices, someone would use those patterns, predict the future, make a huge amount of money, and in the process of making that huge amount of money make the past pattern invalid (there is a lot of math backing this up, see [30] for a deeper exploration, or Casti's [23] for a more casual explanation).

But, theoretical reasoning aside, whether markets are efficient or not is an empirical question. Over 30 years of examination, the EMH has proven to be surprisingly robust (see Fama's papers [33],[34] for a description of just how robust, more discussion is given in section 1.3.1). However, exploring possible market inefficiencies is a well-traveled road in economics, and recently it's been traveled with some success.

Understanding real-world market efficiency has tremendous implications for the broader economy. Capital markets have a function: to guide the flow of investment in the economy. If they are functioning inefficiently, then, quite literally, capital being misallocated, creating very real waste in the economy.

One example of this – although the conventional interpretation is by no means uncontroversial – is the recent internet stock bubble. In retrospect, it seems that tremendous amounts of capital were misdirected towards questionable start-ups. I do not mean to make a concrete case for this interpretation here; rather, I offer a plausible example of how market inefficiency might step out of theory and into real life.

In order to understand how this thesis addresses the efficient market hypothesis, two technical points need discussed: first, information sets; second, the role of risk.

1.2.1 Efficient Markets, Information Sets

One of the key questions in the definition of efficient markets is *what kind* of information is relevant. Forecasting stock movements based only on past prices is very different than forecasting stock movements based on insider information about a pending merger. Roberts [84] set up the standard taxonomy of information sets:

- *Weak-form Efficiency*: The information set includes only the history of prices or returns
- *Semistrong-form Efficiency*: The information set includes all information known to all market actors – all publicly available information
- *Strong-form Efficiency*: The information set includes all information known to any market actors – all private information.

Most economics work on efficient markets has focused on past price data, on the idea of weak-form efficiency. Some work, like event studies, has gone beyond price information to study public events such as merger announcements, producing tests of semistrong-form efficiency.

The work in this thesis squarely addresses semistrong-form efficiency: news data is a canonical example of public information that is not prices. Expanding the information set

However, the idea of semistrong-form efficiency is primarily of theoretical interest. The idea that one could have any kind of insider information – regulatory decisions, merger intentions – and *not* profit from it if allowed to strains credulity. In addition, in order to test the semistrong-form, one would have to have access to all relevant private information – no easy task even after the fact.

1.2.2 Measuring Performance and Joint Tests

The basic idea behind using prediction methodologies as an operational test of the efficient market hypothesis: can the predictions be used as part of a trading strategy to produce superior performance to a benchmark?

The first obvious question is: what is the benchmark? The standard comparison is to some measure of market performance – the S&P 500, or, in this thesis, the full Russell 3000 (the return series of the two indexes is virtually identical).

The deeper question is how to measure performance. Simply looking at percentage returns – which strategy makes more money – isn't enough. It is well-understood in finance that returns are dependent on risk – riskier investments earn higher returns. So any measure of performance has to account for risk as well as return. The traditional measure of risk in financial markets is simply the standard deviation of returns. The standard measure of returns adjusted for risk is the Sharpe Ratio, which takes the returns produced by the strategy the returns produced by the strategy, minus the risk free returns (generally a long-term government treasury bond), divided by the standard deviation of returns.

Formally:

- Let r_s be the return produced by the strategy.
- Let r_{rf} be the "risk-free" return.
- Let σ_s be the standard deviation of the returns produced by the strategy.

- Then the Sharpe ratio is: $SR = (r_s - r_{rf})/\sigma_s$

However, there is a deeper point here. Any test of the efficient market hypothesis is unavoidably a test of assumptions of a pricing model, of assumptions about risk. This may not seem obvious at first glance, but imagine the following example. Take the following security: an option on the S&P 500 that pays off in case of a large market drop (one can think of this like insurance; securities like this exist and are widely traded).

What is the return profile of selling this security? A steady stream of returns, *until* there is a market crash. Most of the time, the Sharpe ratio for this security will look exceptionally strong; but the real risk involved in the security is in infrequent events.

Now, this example is a far cry from the trading strategies developed in this thesis, but it illustrates the point that a strong Sharpe ratio does not necessarily reflect the risk inherent in a trading strategy.

Unfortunately, there is no easy solution to this issue – one has to make assumptions. However, most of the recent work in economics using trading rules (discussed below, in section 1.3.1) that inspired this thesis has simply used Sharpe ratios as a measure of risk-adjusted returns. This thesis will share that approach.

1.3 Relation to Other Work

The work presented in this thesis does not easily fit into a single tradition of research; it addresses concerns in both finance and computer science, and even within those broad disciplines, the work draws from diverse threads of research as such as studies of trading rules and event studies in finance, and ensemble methods, genetic programming, and natural language processing in Artificial Intelligence.

This work can be seen primarily as contributing to two separate research traditions, one in economics and one in computer science. I discuss each of these traditions in sections 1.3.1 and 1.3.2 below. In addition, I draw from a wide variety of AI techniques for tools in building the trading rule learner and news classifier that have not traditionally been applied to financial data; I discuss these in section 1.3.2 as well. Each chapter presents a brief previous work section as well, addressing its specific contents.

1.3.1 Economics

The first is the thread of finance research using excess returns produced by trading rules (both derived from practitioners and induced by machine learning techniques) as operational evidence for market inefficiency.

This line of research is only a subset of the broader study of market efficiency, which includes a wide variety of theoretical and empirical work. The current notion of efficient markets really started to take off in the 60s'; one seminal paper was Samuelson's [87] formulation of security prices as random walks; further formalization and summary of contemporary work been provided by Fama in two classic papers [33, 34]; more recently, Lo [66] provides an introduction to more recent developments in the field, which covers some of the research threads related to this thesis that emerged in the 90's. While traditional theoretical economics work has focused on developing the theoretical machinery behind efficient markets, Grossman and Stiglitz [44] published a classic paper reconciling the possibility of above market returns with the cost of information that provides them.

There are strong theoretical reasons to believe that markets are efficient – they all involve some sort of arbitrage argument, which boils down to the following intuition: why isn't everybody doing it? If there are inefficiencies in the market, firms would take advantage of them, thus making them disappear. This is a powerfully persuasive argument, and has no comprehensive competitors – even if one admitted inefficiency, there is no strong theoretical framework to explore it.

On the empirical side, There is a long tradition in economics of empirical work supporting the idea of efficient markets, mostly examining anomalies and technical analysis indicators claimed to be useful by practitioners. Aside from some well-known and fading anomalies such as the January effect and the value effect, the empirical literature has traditionally supported the idea of market efficiency. In particular, the literature has traditionally been derisive of technical analysis rules.

Recently, that has started to change. In the early nineties a new line of research emerged that uncovered promise in technical analysis. This research started in earnest with Brock, Lakonishok, & LeBaron's 1992 paper [22] examining simple technical trading rules – specifically moving average and trading range break, finding statistically significant anomalies in returns. This has expanded in recent years; Carol Osler has looked at head-and-shoulder patterns [74]. Lo and collaborators have explored a wide variety of similar approaches [16], finding some evidence for asset return predictability from patterns inspired by visual analysis.

The next evolution of this line of research was to apply the following logic: if simple trading rules are good, might complex trading rules be better? Economists turned to machine learning to induce complex trading rules using genetic programming; specifically, Allen and Karjalainen [15] and a series of papers by Neely and collaborators [71, 70]. This is still a small line of work; the economists are still largely limited by inexperience with machine learning techniques.

Much of the work in this thesis can be viewed as an extension of this line of research, and I feel that it makes the following specific contributions.

First, A dramatic expansion of the stock universe involved. – previous work with trading rules on the economics side has customarily focused on indexes or small groups of forex series. Although I use a shorter time period – five years for the work on technical analysis and machine learning, only 14 months of data for the work on news – the stock universe I utilize is far greater. The original Brock et al paper [22] operated over the Dow Jones index; Allen & Karjalainen [15] looked at the S&P 500; and some of Neely's work [71] was applied to a small set of less than fifteen foreign exchange rate series. The stock universes I utilize range between around 1300 stocks, more than two orders of magnitude larger than some of the datasets in previous work.

Second, I make a real attempt to adapt machine learning techniques specifically to the domain. Financial data is noisy – so noisy that the theory this line of work struggles against, the efficient market hypothesis, posits that the data is essentially all noise. In any case, it is likely the one of the noisiest datasets machine learning has ever faced. Previous work on the economics side has been content to apply genetic programming methods without much thought to their applicability or susceptibility to overfitting.

Machine learning researchers understand well that noisy data leads to overfitting within the algorithm, and in chapter 3, I attempt to fight overfitting at every level of the algorithm – through sensibly limiting representation and search, and introducing ensemble methods on top of the simple rule learner. Results show just how important this task is – applying the standard genetic programming framework to the problem of inducing trading rules produces poor results. But by limiting representation and search to almost absurd levels of simplicity, performance is dramatically increased, showing just how aberrant financial data is when compared to traditional machine learning techniques have faced.

Another contribution in this vein is a methodology to resist data mining at the algorithm level. Before I discuss this, I need to clear up some potential terminological confusion. To an economist, "data mining" is a bad thing. It represents the practice of examining large numbers of patterns to find one that functions well on a specific dataset – with no regard for its ability to generalize to novel data. In the case of financial data, supremely noisy even in the best of times, this usually ensures that the chosen pattern will correspond more to an accident of noise in the data than underlying regularities in the data. This problem – when phrased in terms of automated search over training sets – is are core concern of machine learning, customarily referred to as overfitting.

In machine learning, "data mining" is a good thing – it means searching for patterns in data that *do* generalize to novel data. Ironically, one of the biggest concerns of "data mining" in the machine learning sense is avoiding "data mining" in the economics sense. Henceforth, I will refer to both "data mining" in the economics sense and the traditional machine learning concept of overfitting by the less semantically loaded term overfitting.

Economists are painfully aware of the the possibility of overfitting trading rules (see [93]). However, it is possible to overfit at the algorithm level as well. The space of sensible ways to learn trading rules is far more robust to overfitting than the space of trading rules itself. The space of machine learning methodologies is constrained, because it consists of techniques that have proven over a wide variety of data sets to produce induced hypotheses that do generalize well to novel data. In other words, machine learning techniques prosper exactly because they are resistant to overfitting.

But, when dealing with near-pathologically noisy data, it is still easy to fine-tune algorithm parameters to improve performance on a specific data set. Therefore, when exploring the effectiveness of machine learning methods on extremely noisy data such as financial markets, the traditional training/test set split is not enough.

I use the following methodology in chapter 3: I split the data into training, test, and holdout sets, and make all the algorithm design decisions based on performance on the train/test data. Only when I commit to a final do I then check the results on a further holdout set. This provides a firm guard against overfitting at the algorithm level, and I would hope that it become standard methodology in all future applications of machine learning to financial data.

Perhaps the most important contribution is to integrate new sources of data – message boards and news – into the trading rule framework. Although I feel this is the largest contribution of the thesis, certainly with regards to the economics literature, it is hard to put it into context with existing economics research – there is almost nothing like it in the literature.

Clearly, the idea of examining market-external events for their implications to the idea of efficient markets is not new – event studies have long been a cornerstone of financial research. However, such event studies require hand classification, and as such both their scope and applicability are limited.

In this sense, the work here represents a profound advance. The fact that news stories are now present on the web and easily gathered, and that they can be classified into categories, serving as excellent proxies for events in the real world, allows for the integration of of real world events with traditional numerical-based trading rule methodologies in a common framework.

As far as stock-related message boards go, they are a relatively new phenomenon, and serious economics work on them is only beginning, with a trail blazed by Wysocki [99]. The analysis I provide follows the pattern he has set closely; the additional contribution I make consists of enlarging the stock universe and time scale, and exploring ways to use the message board signals

1.3.2 AI

The second research tradition is the vast body of work applying AI to financial forecasting. This isn't so much a coherent research tradition as a shared dream; while there have been a huge number of papers written on AI and the markets, they rarely follow a common framework, or build on a shared tradition of work. Rather, sometimes it seems that every single AI technique ever developed gets thrown at the problem of financial forecasting sooner or later.

The contributions on the AI side should definitely be considered in the spirit of applying AI to an interesting domain. The work here does not advance the theoretical underpinnings of any of the techniques utilized, except to increase awareness of their functioning under conditions of near-pathologically noisy data.

The amount of work on AI and finance is too great to completely summarize here; it is likely that every AI technique ever developed has been thrown at financial markets. The traditional approach has been to take an AI technique and apply it straightforwardly to the task of mapping past prices onto buy and sell signals. Unfortunately, this approach, which ignores the challenges financial data poses, is usually unproductive.

Here, I focus on a selection of work that I feel represents intelligent attempts to adapt AI technologies to financial markets.

Hellström and Holmström [48] present an interesting overview of the quirks of the financial domain from the machine learning perspective, addressing issues such as returns vs. prediction accuracy. In addition, it gives an overview of neural network and memory based method techniques for forecasting.

By and large, the most interesting papers that apply AI to problems in the finance domain are those that forgo the vanilla approach of naively using an AI algorithm to pick stocks.

A particularly interesting example of a novel approach to the problem is that taken by Jefferies, Lamper, Johnson, and their colleagues: instead of the standard machine learning approach of learning to predict via function approximation past data, they use agent-based market simulations to make predictions [53, 55]. They base their work on the basic framework of the minority game, introduced and developed by Challet and Zhang [26, 103, 27]. While the idea of studying markets via multiagent simulations is certainly not new, the strong claim that agent-based simulations can make predictions about future market behavior is, and, if true, is an important watershed.

Another excellent example, and one strongly related to the conclusions reached in chapter 3, is a series of papers by Hellström ([46, 47]). Hellström uses simple techniques – straightforward regression – to predict the *rank* of returns for a given stock out of a universe of 80 from the Swedish stock exchange. The results are quite promising, and the strategies eventually induced by the learner in chapter 3 are very similar.

A promising new area of work, and one that the work in this thesis squarely addresses, is the use of text as a new data source for financial forecasting. Real research here has only recently begun, producing only a few papers. Lavenko et al [95] identify short term trends in intraday stock prices and use naive Bayesian classifiers to predict these trends based on news stories. This work is extremely promising, linking news story classification directly to the prediction of price movements. This approach builds on previous wok by Fawcett and Provost who adapt a fraud-detection system developed in [36] to forecast price spikes based on news stories [35].

Another pioneer of this research has been Wüthrich and his students [98, 94], who have built keyword-based systems to predict major market indexes with considerable success. This work differs from that presented here in several ways. First, while I explore trading rules that operate over a vast array of stocks, Wüthrich and his colleagues are focused on major market indexes (specifically, the Hang Seng Index). As a result, the relevant set of news data is different: when predicting major indexes, nearly *all* financial news is likely relevant to some degree. On the other hand, when focus is narrowed to an individual stock, the amount of news relevant to that stock drops dramatically.

Also, Wüthrich built a purely operational system, automatically learning which keywords are most important to forecasting. In my approach, the work in chapter 5 combines text processing with an intelligent breakdown of the kinds of financial news; while it is my belief that a system like this potentially provides substantial operational advantages, it also allows for integration with existing economics work of how markets react to specific kinds of news events, such as mergers and earnings.

Contributing AI Techniques

In addition to the contributions described above, this work draws from a wide variety of other AI technologies for use as tools for application in the financial domain. These techniques are genetic programming, ensemble methods, and text classification. The work in this thesis should be considered in the spirit of adapting these techniques to interesting domains; there is little in this thesis that advances the state of the art of the techniques themselves.

One of the main conclusions of chapter 3 is that an overemphasis on specific search methodology is misguided, and, although this work uses genetic programming as a technique, many different search techniques would likely work as well. But the starting point for much of the work here is genetic programming, and so a brief discussion of it here is necessary.

Genetic programming is an extension of genetic search methods into learning functions. The first work in genetic programming was done by Koza (see [57] for a comprehensive overview), drawing on the foundational work in genetic algorithms done by Holland [51, 50] (see also Goldberg [42] for an excellent overview of this school of research) in the US. Although they comprise separate strains of research, similar techniques were developed in parallel by Fogel in the USA [37], called Evolutionary Programming, and Rechenberg and Schwefel in Europe with Evolutionsstrategie [80, 81, 88].

The core intuition behind all of these techniques is the idea of search through mimicking the natural process of evolution – and although that metaphor has stretched considerably over the years, it is still a useful crutch for understanding. Genetic search methods all start with a population of candidate solutions – where a solution can be a function, a set of parameter values, a set of portfolio weights, depending on the specific problem.

Fundamentally, there is nothing magical about genetic programming vis-a-vis other genetic search methods; at its core, genetic programming is simply standard genetic search methodology with a clever tree-structure representations that allows for searching over spaces of arbitrary functions. Traditional genetic search methods had focused on optimizing fixed structures of parameters; the core innovation of genetic programming was an open-ended representation that could encompass relatively unconstrained functions.

Ensemble methods are a rapidly growing area of interest in machine learning. They are a meta-technique – a technique that sits on top of an existing machine learning technique and makes it better. The basic idea is simple: instead of using the function induced by an individual learner, ensemble methods apply the learner multiple times on the same problem under slightly tweaked conditions to produce a set of slightly different learned functions. The outputs of these functions are then combined together to produce a single output. The intuition behind this is that having many different functions learned over slightly different conditions will smooth out the effects of noise in the data or learning process. Given the severely noisy nature of financial data, this is an especially promising prospect.

Ensemble methods all share this basic framework; they differ in how they tweak the conditions under which the learner runs. The best known methods tweak the training set under repeated iterations. Bagging, developed by Breiman ([21]), generates novel training sets through bootstrap replication, while boosting, a later development developed by Shapire (nicely summarized in [91]) perturbs the training set in response to the learner's performance on each point in the training set. There are, of course, other methods (see Dietterich [28] for a comprehensive introduction).

Ensemble methods have demonstrated convincing effectiveness when applied a broad spread of machine learning techniques, and have become a significant thread of machine learning research.

Perhaps the the research tradition most important to the core of this thesis is that of text classification. Although the work addressing news data in chapter 5, does not explicitly use any of the traditional statistical-based techniques such as k-nearest neighbors or decision trees, the research is clearly performed in the spirit of the text classification.

Research text classification originated with interest in automated document retrieval in the seventies (see Cleverdon [24] or Salton [85] for an introduction to the state of the art in the 1970s). The primary thrust of this early work was in managing large document collections, automating their retrieval a task somewhat analogous to that performed by search engines today.

This early work laid the foundation of automating document handling through statistical measurements over key terms, and established the framework of evaluating document classification performance in terms of precision and recall [63, 25].

These early methods depended on straightforward notions of the relative frequency of terms within documents. Since the 90s, there has been an explosion in the application of more sophisticated techniques to document classification, including Bayesian classifiers [64], decision trees [17, 64], maximum entropy [72] and support vector machines [54]. This body of research is massive; both Faloutsos & Oard [32] and Mladenic [69] provide excellent surveys.

The approach I develop in chapter 5 does not use any of these statistical techniques; rather, I draw inspiration (if not actual algorithms) from classification techniques derived from information extraction [83]. Information extraction is a technology designed not primarily for classification, but rather for extraction key information from documents. One standard approach is to build a set of pre-defined sentence templates that identify contexts where the key information appears; the textual information that fits in the open slots of the template is then extracted.

This approach can be straightforwardly extended to classification [83] by simply think of the key sentence templates as features; once this conceptual leap is done, the full strength of the statistical methods described above can be brought to bear on this feature space.

As I mentioned before, my approach in chapter 5 is not nearly this sophisticated, but it does depend on a notion of recognizing key combinations of nouns and verbs, and a notion of logically combining the resulting features together.

1.3.3 Summary of Contributions

To summarize the primary contributions of this thesis:

- A cataloging of technical analysis indicators and their effectiveness in a trading strategy framework.
- The use of a large stock universe – literally two orders of magnitude larger than the time series used in some comparable work – for testing the usefulness of technical analysis, machine learning, and news data.
- A deeper appreciation for just how noisy financial data is compared to traditional machine learning datasets, and an understanding of the use of representation and search to fight noise.
- The introduction of ensemble methods to the problem of learning trading rules.
- A methodology for fighting overfitting at the algorithm level in applying machine learning to financial data.
- The introduction of a framework to start research in earnest of
- A productive example of integration between numerical market data and text data in a machine learning trading rule system.

Chapter 2

Technical Analysis

2.1 Introduction

Technical analysis has long been used by a dedicated core of practitioners in the financial world. But it has also long been a disreputable subject in the world of academic finance; intentionally cloaking itself in arcane technology, flaunting the idea of efficient markets and fundamental financial analysis, technical analysis has long been dismissed by both economists and many mainstream investment managers a waste of time or even a self-deceptive form of snake oil.

But in the early nineties, a few economists have dipped their toes into the water and started to rigorously examine whether technical analysis has anything to offer, and that thread of research has grown steadily over time.

This chapter builds on that tradition. It asks a straightforward question: can technical analysis techniques in common use by practitioners be used to produce statistically significant returns that beat a reasonable benchmark strategy?

While this question has been asked before – on innumerable occasions by practitioners, more recently by a growing number of academic economists, whose efforts are summarized in section 2.2 – I believe this work represents a real advance, for the following reasons.

- Data Set: I use over 5 years of data from over 1750 stocks. Traditionally, economists analyzing technical analysis have examined single indexes or a small class of price series.

- Comprehensiveness: I test every technical analysis indicator I could easily formalize, along with multiple methods for transforming the indicators into actual portfolios. Traditionally, papers have examined one or two techniques; this work attempts to catalog a wide variety of techniques.

2.1.1 Goals of this Chapter

This thesis is structured so that each chapter builds on the progress of the last. This chapter is no exception. It has two primary goals:

First, I test the technical analysis indicators themselves for evidence of predictive power (measured by Sharpe ratios of trading strategies utilizing the indicators). To this end, I list a large catalog of technical analysis techniques and test their effectiveness.

But just as important, I study the technical analysis indicators as representations for suitable for introducing machine learning in chapter 3 and integrating message board and text data in chapters 4 and 5. To this end, I use representations that will prove suitable to parameterization and search, and I pay attention to understanding interesting problems that can be tackled with machine learning and the application of textual sources of data.

2.1.2 Chapter Organization

The structure of this chapter is as follows. First, in section 2.2, I cover related work, both by practitioners (little of this work can be counted as rigorous, but it does lay the groundwork) and by academic economists.

Then, in sections 2.3 and 2.4 I present the general methodology of the enterprise – data sets, assumptions, the mechanics return calculations, tests of statistical significance and benchmark calculations. In section 2.5 I present a brief discussion of technical analysis indicators in general, with an example, and a discussion of how I transform the numbers spit out by the indicators into portfolios.

The core of the chapter is a series of sections covering individual technical analysis indicators, from section 2.6 to section 2.12.4. In each of these sections, I present the formalization of the indicator with a brief discussion, and then the results of the indicator applied to the dataset, along with discussion of interesting results.

In section 2.13, I take and summarize which indicators (along with which methods of interpretation) prove to be successful, and examine some characteristics of that set of indicator. Section 2.14 explores connections between indicators, including discussion of the correlations in daily returns between the successful rules. Section 2.15 explores potential problems in implementing these trading strategies, including liquidity constraints and transaction costs, and finally, section 2.16 concludes and summarizes.

2.2 Previous Work on Technical Analysis

Much has been written about technical analysis, split into two camps: practitioners and economists. These two camps have very different styles and goals. Practitioner discussions appear on popular TV shows, magazines, and financial websites; the work of economists appears largely in academic journals. While practitioners are theoretically driven by a profit motive – either by trading, or by salaries as media figures; economists – in theory – are motivated by the advancement of science. Practitioners – at least those willing to talk in public – have no tradition of rigorous testing – their job is to get others excited about the techniques. The goal of economists, on the other hand, is largely to debunk spurious claims of market efficiency implicit in claims of the effectiveness of technical analysis indicators.

Therefore, I consider the two traditions separately.

2.2.1 Practitioners

The practitioner literature is difficult to approach for two reasons. First, the tradition of technical analysis shies away from rigorous analysis. In addition, practitioners have a strong disincentives from sharing knowledge – if they know something works, and they are positioned to financially benefit from it, then they have an incredible incentive *not* to share useful information with the general public. Those who do share with the public may not be the best equipped to real discuss what works.

However, there are a number of sources that semi-formally describe well-known technical analysis techniques. One excellent introduction is Pring's "Technical Analysis Explained" [76]; in addition, there are several websites, such as those found at "Technical Analysis from A to Z" [13] `http://www.equis.com/free/taaz/` and

Stockchart.com's chart school [49] `http://www.stockcharts.com/education/` that present a practitioner's view.

Sadly, discussions of technical analysis by practitioners are rarely accompanied by any sense of rigorous evaluation; the usual commentary is that technical analysis techniques should not be relied upon by themselves, but should rather be one factor in a trading decision made by a human being. Perhaps this is true – but it certainly rules out any statistically sensible notion of testing the effectiveness of technical analysis.

2.2.2 Economists

Fortunately there has been an increasing amount of interest on the part of economists on technical analysis. Traditionally, economists have focused on showing that technical analysis is all smoke and mirrors. However, recently, recently, new investigations have hinted that some technical analysis claims may stand up to rigorous analysis.

One school of exploration started with Brock, Lakonishok & LeBaron [22]. They took simple moving average trading rules and measured the excess returns they produced over a buy-and-hold strategy on almost 100 years of Dow Jones data. They found that the simple moving average rules consistently produced higher returns with lower risk than a buy-and-hold strategy. Their real contribution was to link this observation to a null-model generated by bootstrap statistics; for the first time, economists really had statistics they could take seriously.

This basic idea of using the excess profits (excess with respect to a passive buy and hold strategy) generated by trading rules as a diagnostic for market inefficiency has been spun out into asking more specific questions about market efficiency. For example, LeBaron [59] took this basic technique of using excess returns generated by simple trading rules as a measure of inefficiency and used it to examine the efficiency implications of central bank intervention in foreign exchange rates.

Carol Osler has also produced interesting work showing that the head and shoulders formation [74] produces excess returns on some currency markets. More recently, she has examined the concepts of support and resistance. Osler [73] examined support and resistance levels provided by investment banks and found that foreign exchange rates "bounced" off of the support and resistance levels far more often than chance would indicate. Lo, Mamaysky, and Wong [16] have produced perhaps the most extensive treatment of technical analysis to date in this vein.

2.2.3 Charges of Data Mining

One of the most effective counter arguments to the economics work done above are criticisms of data mining, made most notably by White in papers such as [93]. The criticism is that given that there are a wide variety of potential technical analysis strategies to evaluate, if you keep looking hard enough eventually you will find one that meets a test of statistical significance. Thus, the statistical significance of a strategy cannot be evaluated in isolation; it must be evaluated in the context of the universe of technical analysis trading rules.

There is no easy answer to this critique. I am examining technical analysis indicators carefully selected over the years out of a huge universe of possible technical analysis indicators; the idea that someone designing an indicator looked at fifty similar constructs and happened to pick one that worked well on the data is not implausible. In response to this critique, I make two points:

- All of the techniques examined here were developed long before the start of the time period used for my data set; therefore, if they were selected for their coincidental good performance on data, it was not data present in these experiments.

- I fix the algorithms I use to transform the indicators into specific trading strategies with what seem like a priori reasonable parameters. I do not attempt to engineer or fine-tune these strategies to the mix of indicator and data, even in a situation where a computer scientist, working with less problematic data, would find the temptation unbearable. This reduces the possibility that I am presenting the best results found across many possible formalizations.

2.3 Data

I use the following dataset: stocks present in the Russell 3000 as of June, 2001, which have a continual return series from January 1, 1995 to July 31, 2001. This represents 1682 stocks, and 1409 trading days. Stock prices – closes, opens, highs, and lows, as well as trading volume – were gathered from the Yahoo! Finance quote server at `http://finance.yahoo.com` [9]. These prices are both split- and dividend adjusted.

2.4 Methodology

The methodology of this chapter is simple: take technical analysis indicators, formalize simple trading strategies that generate portfolio holdings from the indicator scores, and measure the returns produced by each strategy. Note that the kinds of technical analysis indicators at issue here do not automatically translate into buy and sell signals or portfolio holdings; I have to provide those mappings in the form of various strategies (these strategies are discussed in depth in section 2.5.2. Then, the Sharpe ratios of those indicator/strategy combinations are tested for statistical significance, both against an appropriate buy-and-hold benchmark (the Russell 3000), as well as varying null hypotheses using bootstrap statistics. This echoes the basic methodology established by Brock, Lakonishok, and LeBaron [22]. To be more specific, the procedure for measuring the returns is:

- On each day, the technical analysis indicator spits out a number.
- A trading strategy translates that number into a long, neutral, or short trading signal (these strategies are described below, in section 2.5.2)
- A long portfolio is constructed, equally weighted among all stocks with long trading signals.
- A short portfolio is constructed, equally weighted among all stocks with short trading signals.
- The total portfolio is market neutral – the aggregate long and short positions are equal.
- The returns from the long and short portfolios are calculated based on the closing price, and a .1% one-way transaction is applied where appropriate.

A transaction cost of .1% is on the low side for investigations of equity trading. I explore the effects of higher transaction costs below in section 2.15.2. my intent here is to cast a wide net and capture cases that might show borderline effects, and then investigate to see how performance holds up under increasing transaction costs.

Formally, our daily portfolio returns are computed as follows:

- Let $r_{i,t}$ be the daily log return for each stock i at time t

- Let $h_{i,t}$ be our daily holdings of stock i at time t (this is positive for a long position, negative for a short position, and zero for a neutral position)
- Let tc be the level of one-way transaction cost (assumed to be .1%)
- total daily log return: $r_{strategy,t} = \sum_{i=1}^{n} r_{i,t} \cdot h_{i,t} - (h_{i,t} - h_{i-1,t}) \cdot tc$

To evaluate each indicator/strategy combination, I present the following summary statistics:

- Sharpe ratio (this is the standard measure of risk adjusted returns; it's the annualized excess return over the annual risk free rate divided by the annualized volatility (traditionally measured by the standard deviation of returns). Formally: $SR = (AR_{strategy} - R_{risk_free})/\sigma_{strategy}$
- Annualized total return
- Annualized standard deviation of returns (this is the canonical academic finance measure of volatility). This was computed using the following formula.
- Biggest peak-to-trough drop.
- Percentage of long and short positions taken.

I computed the annualized return and standard deviation over the log returns, then converted the resulting figures to percentages before computing the Sharpe ratio. Letting r_t be the daily log return, and T be the total number of trading days in the sample, and 252 be the number of trading days in a year, then the formulas are as follows:

- Mean daily log returns $(\mu_r) = (1/T) \cdot \sum_{t=1}^{T} r_t$
- Standard deviation of daily log returns $(\sigma_r) = \sqrt{(1/T) * \sum_{t=1}^{T} (r_t - \mu_r)^2}$
- Percentage annualized returns: $e^{252 \cdot \cdot \mu_r} - 1$
- Percentage annualized standard deviation of returns: $e^{\sqrt{252} \cdot \sigma_r} - 1$

For the risk free rate, I use the rate on the 90-day Treasury Bill. Although the only measure I use for statistical comparison is the Sharpe ratio, the other measures are intended to give a broader picture of the strategy – the peak-to-trough drop figure can expose risk not apparent in the standard deviation statistic, and the long/short holdings figure indicate if a strategy is only rarely making bets, or is almost always taking a position.

Many of the Sharpe ratios presented in this chapter and in chapters 3 and 5 will seem high by the standards of the literature (although in line with similar results produced with similar trading strategies by Hellström [47, 46]). Although I do take into account transaction costs, real world issues would undoubtedly render it impossible to acheive the full Sharpe ratios found here.

Also, the issue of measuring risk itself is problematic; although my standard measure of success is adjusted for a measure of risk, it is very possible that these trading strategies invoke forms of systematic risk not accounted for by the standardized deviation of daily returns. Innovative analyses of risk such as Hsieh's analysis of hedge fund strategies as options [41] could be applied in the future to get a better picture of the true risk involved.

In addition, for strategies with interesting properties, I plot their daily net asset value (NAV) graphs for examination.

2.4.1 Benchmark

The appropriate benchmark for both qualitative and quantitative comparison is the return series of the Russell 3000, a broad index tracking the performance of 3000 American equities comprising roughly 98% of the value of the American market. The better-known S&P 500 could have been chosen as well, but since the stock universe used in these experiments was drawn from the Russell 3000, it seems a more appropriate benchmark. In any case, as the Russell 3000 tracks the S&P 500 closely, results would be essentially identical.

The daily closes for the Russell 3000 were harvested from the Yahoo.com quote server. Note that exact comparisons to the Russell 3000 are impossible without precise dividend data – by holding the Russell 3000, one would also accumulate dividends, which are not accounted for in the price of the index. If these dividends were included, they would slightly increase the returns on the Russell 3000. To account for this, I take a conservative approach and add daily return equal to an annualized return of 1.5% to the returns of the Russell 3000 when testing for statistical significance. The 1.5% is not an exact figure; according to the Motley Fool's index center, as of September, 2001, the dividend yield on for the latest Russell 2000 is 1.38%, for the S&P 500 is is 1.18% [7] (note while the dividend yield for the stock universe used in these experiments has not been precisely calculated, there is no reason to expect it to deviate significantly from these figures) – so picking a number slightly higher than this is a safely conservative assumption.

I make no attempt to account for dividend payments in the trading strategies themselves (although the price series is dividend-adjusted to handle ex-date effects); thus, it is possible that the results for the trading strategy are underestimated with regards to the benchmark. However, this effect would certainly be small, and not enough to change any of the qualitative conclusions reached in this chapter.

The table below presents the standard summary statistics for the Russell 3000, and the Russell 3000 with a 1.5% dividend component added; figure 2.4.1 presents the normalized NAVs of the Russell 3000 over time, plotted on a logarithmic scale. Investigating the plot, one sees a fairly steady rise until spring of 2000, when the index begins a steady decline that lasts until the end of the sample period.

Strategy:	Sharpe	Returns	Std Dev	Decline	Long/Short
Russell 3000	.33	11.95%	20.50%	-28.77%	100%/0%
Russell 3000 + 1.5%	.40	13.34%	20.50%	-27.97%	100%/0%

2.4.2 Testing Statistical Significance

I test statistical significance in two ways. The first is to compare yearly and quarterly Sharpe ratios of the trading strategies against the benchmark in paired two-tailed t-tests, under the null hypothesis that the difference in Sharpe ratios is less than zero.

Note that testing returns or standard deviation of returns in isolation is misleading – the trade-off between returns and volatility is well established in finance. Taking on additional risk (through leverage, or choosing riskier securities) will increase absolute returns – so comparing returns without understanding the risk profile of the of the returns is potentially misleading. Only Sharpe ratios, which balance the two, provide the full picture, and so Sharpe ratios will be the primary comparison metric.

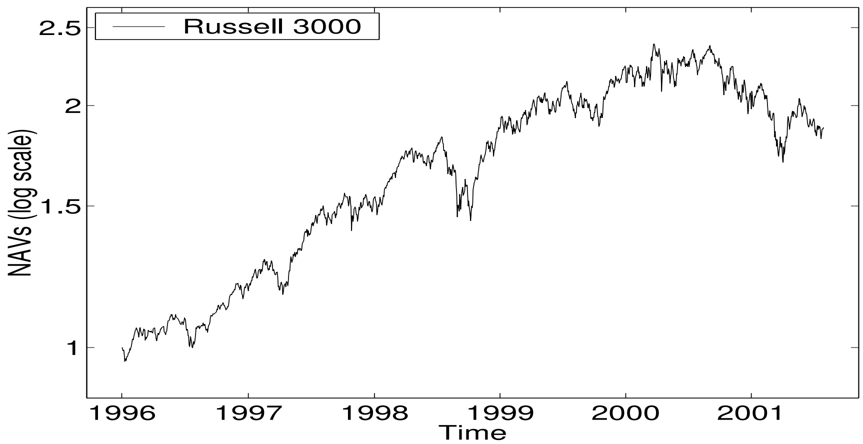

Figure 2.1: Normalized NAVs of the Russell 3000

A second statistical test will be performed using bootstrap statistics (for an excellent introduction to bootstrap statistics, see [31]). The general idea behind bootstrap hypothesis testing is to test against a null hypothesis by comparing the results at issue against results generated over large numbers of artificial datasets whose statistical properties conform to the null hypothesis. The basic procedure is as follows:

- Define the null hypothesis.
- Generate a large number of bootstrap datasets with statistical properties identical to that of the null hypothesis.
- Calculate the appropriate results of the trading strategy for each bootstrap dataset
- Calculate the proportion of bootstrap datasets that produce Sharpe ratios that exceed the Sharpe ratio of the strategy on the real data.
- This percentage corresponds to a p-value.

For the purposes of the following experiments, I use the null hypothesis that log returns follow a random walk (with a general upward trend). I realize that this is somewhat simplistic; some previous work has been careful to examine numerous null hypotheses, including AR, ARCH, and GARCH processes. But given that these additional null hypotheses have not had much of an impact on previous work, the practical considerations in computing such series for over 1600 stocks, as well as the added assurance provided by the other statistical testing of differences in Sharpe ratios, I feel that using random walk in log returns is sufficient.

Therefore, I generate bootstrap datasets as follows. For each stock I take the daily log returns, percentage difference between the high and the close, the difference between the low and the close, and the trading volume. These numbers are then scrambled as a set by date. A new series of closing prices is computed based on the new log returns series, and the high and low prices are computed using the percentage differences from the high and close, and low and close, respectively. The bootstrap trading volume is simply a scrambled series of trading volume.

Note that for each bootstrap iteration, each stock's scrambled time series is scrambled in exactly the same way, and that the scrambled date is the same for the high, the low, the close, and the trading volume: they move as a unit.

It is important to note the difference in intention between the two statistical tests. The yearly and quarterly Sharpe ratio paired t-test is testing for a difference in performance between the technical analysis strategies and the benchmark.

The bootstrap tests do not reference the benchmark at all; they are designed to see if the performance of the trading strategies are an artifact of their statistical distributions of the stock return series.

This work tests statistical significance of many, many kinds of technical trading rules. This leads to a potential problem: when testing twenty different strategies, one would expect one strategy to show statistical significance by chance. The more approaches tested, the more likely one or more will hit at the traditional 5% level just at random.

In the economics literature, the standard way to handle statistical significance in cases like these would be to apply the Bonferroni correction [68], which corrects the p-values required to show a given level of statistical significance by a factor appropriate to the number of hypotheses being tested.

However, to apply the Bonferroni correction, given the number of rules I test, I would have to run prohibitively large numbers of bootstrap simulation. Instead I have to plead for the reader to interpret claims about specific strategies carefully (although claims of the failure of canonical technical analysis methodology stand without lessening from this charge). The real test will come in the next chapter, 3, when the strategies that show promise here are used as raw representation for machine learning algorithms and tested on novel data.

2.5 Technical Analysis Indicators

The sections that follows contain descriptions and formalizations of the technical analysis indicators themselves, as well as the results generated by associated trading strategies. The general format of each section is as follows:

- Present the indicator itself, descriptively and formally.
- Evaluate each trading strategy on the full data set and present the results in summary form, and for interesting strategies, as graphs of Net Asset Value (NAV).
- Examine the statistical significance of the results of interesting trading strategies.
- Summarize the conclusions.

Where interesting questions present themselves, I pursue them. For example, some technical analysis indicators have more interesting side techniques whose usefulness needs to be evaluated.

As I describe each indicator, I will attempt to illustrate the intuition behind each one. However, these explanations should be taken in the proper spirit – the world of technical analysis does not consist of rigorous derivations from first principles.

2.5.1 What do Technical Analysis Indicators Look Like?

Before I present the technical analysis indicators themselves, I here attempt to give a feel for what technical analysis indicators look like in practice. As an example, I present the simple moving average indicator (this indicator will be analyzed in more detail in section 2.6 applied to the Russell 3000. Figure 2.2 attempts to give some intuition; the left graph plots the Russell 3000 along with its 150-day simple moving average, and the right graph plots the ratio of the Russell 3000 to its 150-day moving average.

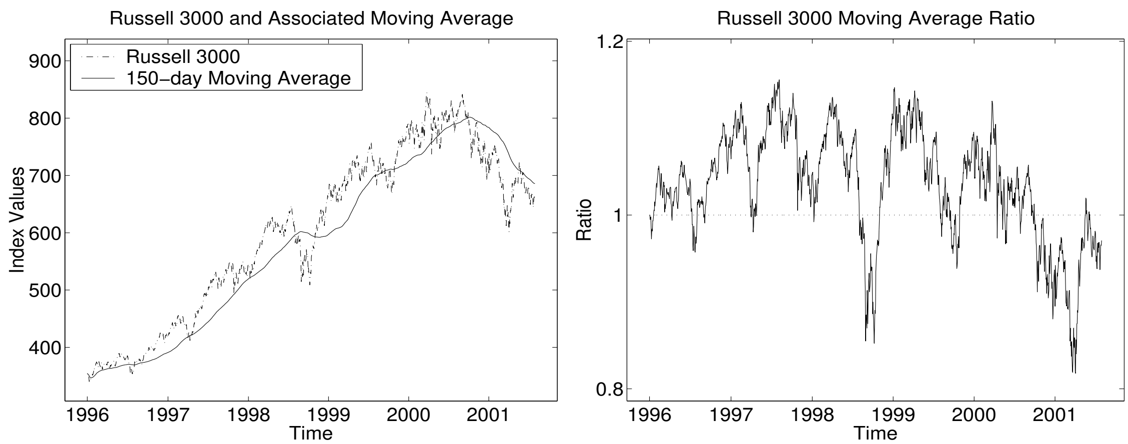

Figure 2.2: Moving average example on the Russell 3000

Technical analysis has traditionally been dependent on visual interpretation; the canonical way to interpret moving average indicators is to look at crossovers – where the price crosses its n-day moving average on the way up or on the way down. The traditional interpretation of moving average crossovers is that crossing above the centerline a good buy signal; crossing below the centerline is a good sell signal. The plots in figure 2.2 present an interesting mix of situations. The downcrossing towards the end of 2000 looks like a good sell signal; clearly, the Russell 3000 drops steadily afterward. However, there are two "whipsaws" in late 1998 and late 1999, where the Russell 3000 drops below its moving average but then rapidly crosses back above. Visual inspection of the graph certainly holds out tantalizing possibilities, although it is not clear how to capitalize on them.

2.5.2 From Indicators to Actual Portfolios

Technical analysis indicators transform price and volume series into a corresponding numerical series. But how to appropriately transform that number into a concrete set of portfolio weights is not always obvious. Details in the practitioner literature are sketchy. Discussions often warn that these indicators should not be used as stand alone trading rules, and – aside from a few vague generalizations – command the reader to interpret the signals on their own.

But of course, in order to test the usefulness of the indicators, I must provide a reasonable formalization of a complete trading rule. That formalization consists of three parts:

- Indicator: Technical analysis indicator that takes price and volume time series and transforms them into a corresponding time series.
- Strategy: A rule that maps the indicator time series onto a corresponding series of portfolio weights.
- Direction: The traditional interpretation of technical analysis indicators is that high values signal bullish positions and that low values signal bearish positions; but for completeness I test the opposite direction.

The indicators will be discussed below, each in its individual section. I discuss the strategies common to all indicators below.

Strategies

I propose four specific strategies for transforming the technical analysis indicators into specific buy and sell signals:

- Comprehensive: Go {long,short} every stock whose indicator is {above,below} its centerline.
- Crossover: For each {upcrossing,downcrossing} of the technical analysis indicator centerline, go {long,short} for 10 days.
- Selective: For each stock a technical analysis indicator in the {top,bottom} 5% of all stocks on a given day, go {long,short}.
- Change: For each stock whose overnight change in the the technical analysis indicator is in the {top,bottom} 5% of all stocks on a given day, go {long,short}. I calculate overnight change by calculating the difference between today's value and yesterday's value, and normalize by dividing by the standard deviation of the last 30 days of the time period.

Of these four strategies, the comprehensive and crossover strategies correspond strongest to traditional interpretations of technical analysis indicators. The crossover strategy accepts upcrossings as bullish, and buys and holds the stock to a fixed time period (here, 10 days), with corresponding short positions after downcrossings.

The comprehensive strategy corresponds to taking this strategy to its full extension, holding the stock after and upcrossing until the next downcrossing.

Most popular discussions of technical analysis explicitly discourage formalization, so there is no firm tradition of exactly how to interpret a bullish or bearish signal; I feel that these two interpretations provide a reasonable coverage.

The selective strategy and change strategies do not correspond to commonly used constructs in technical analysis; they are largely present for thoroughness. It is likely that the change strategy, especially in simple moving average indicators, simply takes advantage of the well-known tendency of stock prices to mean-revert; it will be interpreted with an appropriate grain of salt.

Of course, some of these strategies require discretionary parameters – the post-signal holding period for the crossover strategy, the percentage cutoffs used for the selective and change strategies. I set these at what I believed to be a priori reasonable values.

Direction

There is one more axis along which these strategies can vary. The strategies above assume the traditional technical analysis indicator interpretation that values above the centerline are bullish signals, and values below the centerline bearish signals. For completeness, I will always check the opposite interpretation, where moving average signals above one are bearish signals, and moving average signals below one are bullish. I will hereafter refer to the two approaches as follows:

- "High scores = bullish": Interpret high indicator values as bullish, low indicator values as bearish.
- "High scores = bearish": 'Reverse polarity' of the above strategies; interpret high indicator values as bearish, low indicator values as bullish.

2.6 Moving Averages

Moving averages are probably the most used indicator in technical analysis. There are two kinds of moving averages: the simple moving average (SMA), and the exponential moving average (EMA). The simple moving average is exactly what one expects: the mean of the last n days of data. The exponential moving average is similar, but based on exponential decay. The exact formulas are as follows (let t index time, and n index the desired moving average window length – for this chapter, I set $n = 150$, which is a traditional long-term moving average).

- $SMA_{t,n} = (1/n) \cdot \sum_{i=1}^{n-1} p_{t-i}$
- $EMA_{t,n} = K * EMA_{t-1,n} + (1 - K) * p_t$
- (Where $K = 2/(n + 1)$)

Moving averages are the most often used and discussed construct in technical analysis; they are the building blocks of many of the more complicated techniques. But for stand alone use, the moving average ratio is most often used: simply dividing the current price by the n day moving average. From now on, when I refer to the simple moving average or exponential moving average statistic, I will be referring to this ratio.

2.6.1 Questions

Since this section will serve as a template for all the following technical analysis indicators, let me be clear what the questions I'm trying to answer are.

- Question 1: Can trading strategies using the moving average statistics produce statistically significant Sharpe ratios superior to the benchmark?
- Question 2: If so, what do these strategies look like?
- Question 3: Is there a difference in performance between the simple moving average and the exponential moving average?

2.6.2 Exponential vs. Simple Moving Averages

The differences between the exponential and simple moving averages are subtle. Figure 2.3 takes the 150-day exponential and simple moving averages of the Russell 3000

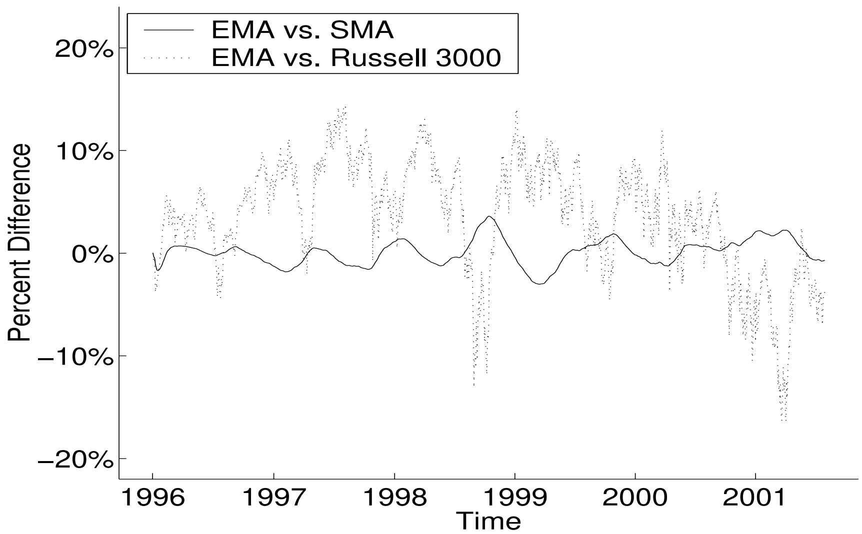

Figure 2.3: percentage differences between moving average types, compared to percentage differences between moving averages and the Russell 3000

time series, and plots the percentage differences both between the two types of moving averages, and between the exponential moving average and the Russell 3000 time series itself. The difference between the moving averages is significantly smaller than the differences between the the exponential moving average and the time series itself (the mean absolute difference between the two moving averages is 1%, compared to a 5.66% mean absolute difference between the exponential moving average and the time series itself).

2.6.3 Empirical Results

The table below contains the results for each of these four strategies, using both simple and exponential moving averages. I present Sharpe ratios, annualized absolute returns, annualized standard deviations, and the average daily percentage of stocks held long or short (out of the whole universe).

Moving Average Results (High Values Bullish interpretation)					
Strategy:	Sharpe	Returns	Std Dev	Decline	Long/Short
Comprehensive (SMA)	-1.71	-16.19%	12.46%	-66.62%	59%/41%
Comprehensive (EMA)	-1.95	-19.63%	12.70%	-72.90%	59%/41%
Crossover (SMA)	-5.55	-51.99%	10.28%	-98.36%	12%/11%
Crossover (EMA)	-5.55	-53.24%	10.53%	-98.61%	13%/12%
Selective (SMA)	-1.56	-51.80%	36.43%	-98.60%	5%/5%
Selective (EMA)	-1.61	-54.11%	36.87%	-99.00%	5%/5%
Change (SMA)	-4.83	-89.30%	19.53%	-100%	5%/5%
Change (EMA)	-4.84	-89.17%	19.47%	-100%	5%/5%

These results are uniformly abysmal. It's hard to beat a 98+% peak to trough decline for utter financial ruin, as evinced by the selective, crossover, and change strategies using both simple and exponential moving averages. In fact, the results are so bad for those strategies one suspects that they are far *worse* than a random strategy would be.

But if you're wrong all of the time, that's just as good as being right all the time. I re-run the strategies while flipping the interpretations of the signals: now, being above the moving average is a bearish signal – indicating a short position; conversely, being below the moving average is a bullish signal, indicating a long position. The results are presented in the table below (single asterisks indicate a bootstrap p-value of less than .005):

Moving Average Results (High Values Bearish Direction)					
Strategy:	Sharpe	Returns	Std Dev	Decline	Long/Short
Comprehensive (SMA)	-.23	2.37%	12.31%	-39.23%	41%/59%
Comprehensive (EMA)	.02	5.38%	12.53%	-37.49%	41%/59%
Crossover (SMA)	1.76 *	23.01%	10.15%	-7.70%	11%/12%
Crossover (EMA)	1.98 *	25.64%	10.34%	-8.08%	12%/13%
Selective (SMA)	1.33 *	52.55%	35.54%	-65.36%	5%/5%
Selective (EMA)	1.43 *	56.68%	36.04%	-67.40%	5%/5%
Change (SMA)	2.40 *	50.44%	18.83%	-19.39%	5%/5%
Change (EMA)	2.28 *	47.96%	18.80%	-20.94%	5%/5%

These results are far more positive. Although the comprehensive produces only a borderline positive Sharpe ratio, the other three strategies produce large Sharpe ratios in comparison with our benchmark – the Sharpe ratio of the Russell 3000 with estimated dividends was .40, and three strategies here produce Sharpe ratios of over 1.3 using either simple or exponential moving averages, with high levels of bootstrap statistical significance.

There is an intriguing distinction between the crossover strategy and the selective and change strategies; the crossover strategy achieves its high Sharpe ratio with (relatively) modest returns and low volatility, and a very small peak-to-trough decline. Although the selective and change strategies return more, they are considerably more volatile; the selective strategy has a troublingly large maximum peak-to-trough decline of 65-67%, with the change strategy falling somewhere in the middle.

The following table presents the mean difference in yearly and quarterly Sharpe ratios between the strategies and the benchmark, with corresponding paired t-test p-values in parenthesis.

Sharpe Ratio Differences over Benchmark		
Strategy:	Yearly	Quarterly
Crossover (SMA)	1.74 (.054)	1.41 (.035)
Crossover (EMA)	2.00 (.017)	1.55 (.012)
Selective (SMA)	1.78 (.192)	2.55 (.049)
Selective (EMA)	1.77 (.136)	2.42 (.015)
Change (SMA)	2.60 (.050)	2.80 (.022)
Change (EMA)	2.42 (.071)	2.66 (.043)

Every quarterly result is statistically significant; the yearly strategies are for the EMA crossover and SMA change strategies are statistically significant as well.

To gain intuition on these points from the actual return series itself, I present the normalized log daily NAVs of the two strategies, in both simple and exponential moving average versions, in figure 2.4. The top plot presents the crossover strategy, the bottom the change strategy.

The first point to notice is that in all cases, the performance of the individual strategies does not vary much between the simple moving average and the exponential moving average version; this point will be examined in more detail below in section 2.6.2. In the NAV plots, the EMA and SMA cases track each other closely and are often hard to distinguish.

Secondly, in both plots the strategies produce return profiles superior to the benchmark. The crossover strategy doesn't produce hugely superior returns than the Russell 3000, but it is significantly less volatile; the change strategy's performance is very impressive – far outpacing the Russell 3000 in returns while maintaining low volatility. The selective strategy produces strong returns until it takes a huge loss in early 2000, corresponding to post-Microsoft anti-trust case peak of the market bubble, before resuming an upward trend, albeit with increased volatility.

2.6.4 Exponential vs. Simple Moving Averages, Revisited

One of the questions posed at the beginning of this section is whether or not exponential moving averages are superior to simple moving averages. Examining the graphs in figure 2.4 produces no clear conclusions, other than that the EMA and SMA indicators produce near identical results.

I tested the difference in yearly and quarterly Sharpe ratios between the strategy using the exponential moving average versus the simple moving average. The results, with paired t-test p-values in parentheses, are presented in the table below.

Sharpe Ratio Differences between EMA and SMA		
Strategy:	Yearly	Quarterly
Comprehensive	.203 (.065)	.197 (.250)
Crossover	.259 (.213)	.142 (.640)
Selective	-.012 (.959)	-.124 (.826)
Change	-.1990(.094)	-.136 (.503)

These results confirm the intuitions provided by the plots and the statistics presented in the tables above. The comprehensive and crossover strategies perform better using the exponential moving average, while the change strategy performs better with the simple moving average, and the selective strategy is mixed. However, none of these differences are statistically significant.

Another view on this question can be found in the correlation matrix between the log return series of the strategies under discussion here:

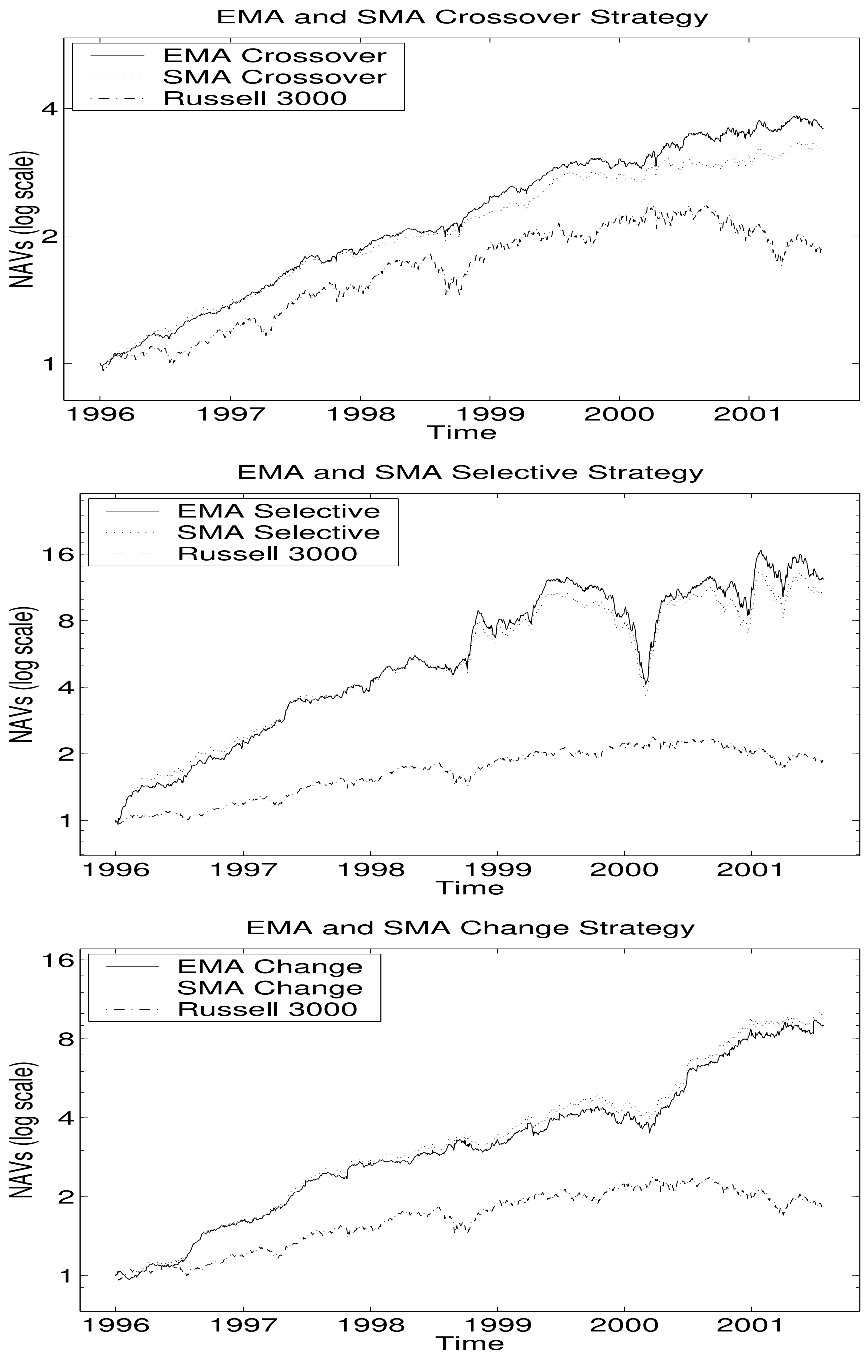

Figure 2.4: Moving Average Strategies NAV Plots

Strategy:	Russ 3000	Cross EMA	Cross SMA	Selec EMA	Selec SMA	Change EMA	Change SMA
Russell 3000	1	-.025	-.015	-.004	-.006	-.057	-.048
Crossover (EMA)	-.025	1	.864	.355	.321	.438	.430
Crossover (SMA)	-.015	.864	1	.360	.321	.445	.444
Selective (SMA)	-.004	.355	.360	1	.977	.121	.112
Selective (EMA)	-.006	.321	.322	.977	1	.109	.099
Change (SMA)	-.057	.438	.445	.121	.109	1	.975
Change (EMA)	-.048	.430	.444	.112	.099	.975	1

First note that the correlation between the simple moving average and exponential moving average versions of the appropriate strategies is extremely high; for the .977 for the selective strategy, .975 for the change strategy, and somewhat less at .864 for the crossover strategy. For the selective and change strategies, the correlation is so high as to make the return profile almost identical; while the correlation between crossover strategies, at .864, is also extremely high, it is not at the same level. Unfortunately, the difference in correlation is only partially born out in the Sharpe ratio data – while the absolute difference in full-period Sharpe ratios between EMA and SMA crossover strategies is higher (.16) than that of the selective (.11) or change (.14) strategies, and the trend seems to hold true for the most part, in the yearly/quarterly data, with differences for the crossover strategy being as larger than differences for the selective and change strategies.

Another interesting observation of this data is that all of the moving average based strategies are almost entirely uncorrelated with the Russell 3000 itself, with the strongest correlation coefficient of -.057.

In summary, the evidence for preferring exponential moving averages to simple moving averages is suggestive at best. But, lacking a clear empirical mandate, I will err on the side of conventional wisdom and henceforth use exponential moving averages in all subsidiary calculations.

2.6.5 Conclusions

- Trading strategies using moving average strategies can produce quarterly Sharpe ratios that are in excess of those produced by our benchmark, a buy-and-hold strategy on the Russell 3000.

- Contrary to conventional wisdom, the traditional interpretation of moving average signals – high moving average ratios and upcrossings are bullish, low moving average ratios and downcrossings are bearish – appears to be inverted.
- Three nearly completely unrelated strategies employing moving average statistics produce statistically significant results under the high indicator values are bearish interpretation – the standard crossover strategy, a strategy based on day-to-day changes in moving average ratios, and a strategy looking at the the highest and lowest daily ratios.
- The evidence for preferring exponential moving averages to simple moving averages, while potentially suggestive, is not convincing.

2.6.6 A Speculative Note

One intriguing possibility is to note that moving average statistics can be thought of as digital filters. When thought of this way, an obvious line of inquiry presents itself: replace moving averages with true digital filters – explore the space of bandpass filters – and see if any improvement presents itself. As far as I know, this possibility has not been explored publicly.

2.7 Moving Average Convergence Divergence

The Moving Average Convergence Divergence (MACD) statistic is a simple difference between two moving averages – a short windowed one and a longer windowed one. The traditional values are 12-day and 26-day.

Signals are generated in multiple ways. Since the MACD is an oscillator, simple upcrossings across the zero line are considered bullish; correspondingly, downcrossings are considered bearish. In addition, the MACD statistic is often plotted against its own 9 day exponential moving average. Upcrossings of the MACD against its moving average are considered bullish; downcrossings are bearish.

- $MACD_t = EMA_{t,12} - EMA_{t,26}$

Since the MACD is an oscillator, I apply the same four strategies – comprehensive, crossover, selective, and change – used for the moving average ratios described above in section 2.5.2. The table below presents the results for the four strategies, interpreting MACD values greater than zero as bullish.

MACD Results (High Values Bullish Direction)					
Strategy:	Sharpe	Returns	Std Dev	Decline	Long/Short
Comprehensive	-1.60	-12.96%	11.26%	-58.85%	56%/44%
Crossover	-2.44	-24.68%	12.22%	-80.24%	15%/14%
Selective	-1.46	-44.58%	34.16%	-96.56%	5%/5%
Change	-4.18	-79.30%	20.18%	-99.99%	7%/5%

These results are not promising – none of them produce positive Sharpe ratios. As with the moving average results, I re-run the experiments while flipping the interpretation of the signals – positive MACD values are now bearish signals; negative MACD values are bullish. The following table presents the results:

MACD Results (High Values Bearish Direction)					
Strategy:	Sharpe	Returns	Std Dev	Decline	Long/Short
Comprehensive	-.31	1.66%	11.14%	-40.90%	44%/56%
Crossover	-.90	-5.84%	12.21%	-44.54%	14%/15%
Selective	1.13*	42.67%	33.28%	-45.93%	5%/5%
Change	3.37*	71.51%	19.66%	-19.73%	5%/7%

These results are an improvement across the board; both the selective and change strategies produce strong positive Sharpe ratios, both statistically significant with a p-value of less than .005 on our bootstrap test. The test of yearly and quarterly Sharpe ratio differences are presented below with two-tailed p-values in parentheses:

Sharpe Ratio Differences over Benchmark		
Strategy:	Yearly	Quarterly
Selective	0.99 (.214)	2.00 (.021)
Change	3.67 (.019)	3.95 (.003)

The quarterly differences are statistically significant, confirming the results above. For a deeper understanding, the NAV series of the selective strategy is plotted in figure 2.5, along with NAVs of the Russell 3000 and the exponential moving average selective strategy for comparison. Casual examination shows that the NAV series produced

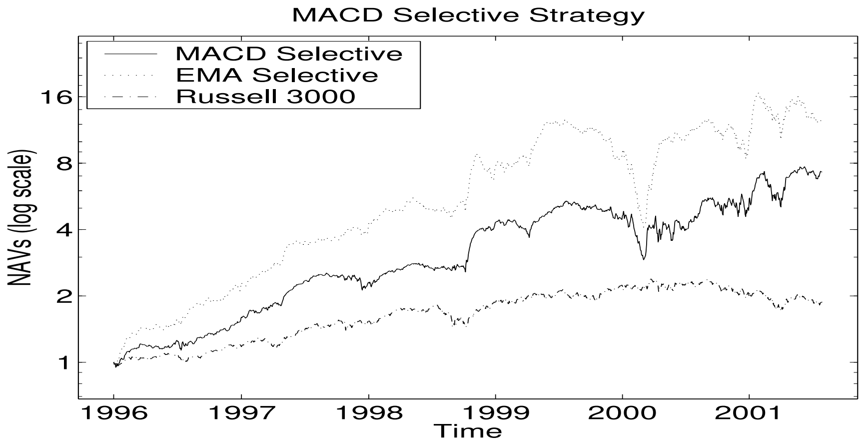

Figure 2.5: MACD Selective Strategy NAV Plot

by the selective strategy on the MACD statistic is qualitatively similar to that of the selective strategy on the exponential moving average, matching the sharp decline found in early 2000 and subsequent increased volatility. The correlation between the log returns produced by the two is .669.

The MACD ratio is also commonly used in comparison to its own 9-day exponential moving average, as a ratio, much like the moving average statistics. To investigate this variation, I took the ratio composed of the MACD divided by its own 9-day moving average and tested it with the standard strategies. I present the results below, for both the high values bullish and high values bearish interpretations:

MACD w/ 9-day EMA Results (High Values Bullish Direction)					
Strategy:	Sharpe	Returns	Std Dev	Decline	Long/Short
Comprehensive	-3.32	-11.77%	5.10%	-50.54%	51%/49%
Selective	-4.83	-37.18%	8.76%	-92.63%	5%/5%
Crossover	-3.84	-15.67%	5.41%	-61.56%	37%/36%
Change	-2.99	.52%	1.54%	-3.89%	2%/0%
MACD w/ 9-day EMA Results (High Values Bearish Direction)					
Strategy:	Sharpe	Returns	Std Dev	Decline	Long/Short
Comprehensive	-3.26	-11.45%	5.10%	-51.86%	49%/51%
Selective	-5.44	-42.45%	8.76%	-95.48%	5%/5%
Crossover	-3.69	-14.91%	5.43%	-60.76%	36%/37%
Change	-3.74	-.57%	1.53%	-7.19%	0%/2%

These results are uniformly negative; none of the strategies produce anything close to a positive Sharpe ratio, and none of them have other interesting characteristics. The change strategy exhibits strange behavior – the MACD divided by it's 9-day EMA often produced numbers of huge magnitude, throwing off the percentage change calculations.

2.7.1 Conclusions

- The MACD indicator produces a positive, statistically significant Sharpe ratio with both the selective and change strategies, under the interpretation that high values of the indicator are bearish signals.
- Applying strategies to the common statistic of the ratio of the MACD with its 9-day exponential moving average produces no evidence of predictive power.

2.8 RSI: Relative Strength Index

The relative strength index (RSI) is an oscillator introduced by Wilder and discussed in "New Concepts in Technical Trading Systems" [97]. Its core is a measure called relative strength (RS), which is a ratio of the average gain (the simple moving average of the gains in the price series) divided by the average loss (the simple moving average of the losses in the price series). This ratio will be greater than 1 if there has been a net gain over the time period, but the interesting property is that high volatility in the stock compresses the RS figure. This measure is then put into a formula intended to produce values from 0 to 100. This is the relative strength index. Formally:

- Average Gain: $AG_{t,n} = \sum_{i=0}^{n-1} max(p_{t-i} - p_{t-1-i}, 0)$
- Average Loss: $AL_{t,n} = \sum_{i=0}^{n-1} min(p_{t-i} - p_{t-1-i}, 0)$
- $RS_{t,n} = AG_{t,n}/AL_{t,n}$
- $RSI_{t,n} = 100 - (100/(1 + RS_{t,n}))$

Following conventions introduced by Wilder, I use a 14-day window to calculate the average gains and average losses.

2.8.1 New Strategies

The RSI has two conventional interpretations in the practitioner literature; an overbought/oversold interpretation and a divergence interpretation. These interpretations apply to several of the subsequent indicators. I formalize them here.

In the overbought/oversold interpretation a high RSI level that declines indicates that the stock is "overbought", or ready to decline, and low RSI levels rising hint that the stock is "oversold", or ready to rise. The traditional trigger points are 70 for overbought and 30 for oversold, respectively. This leads to the following strategy:

- Overbought: Go long when the RSI upcrosses against 30, and short when the RSI downcrosses against 70, and hold these positions for 10 days. (This is similar to the crossover strategy, save for the difference in the triggering event).

The other interpretation is the divergence interpretation: if the RSI and the underlying price series diverge, the price series is expected to follow the RSI eventually. So if the RSI is rising, and the underlying price series is steady or falling, it is expected that the price series will rise soon; conversely, if the RSI falls, and the underlying price series is rising, it is expected that the price series will fall soon.

Formalizing this concept of divergence is non-obvious, and not discussed in the popular literature. The first step is to quantify the notion of "rising" and falling" – simply identifying rising or falling time series is straightforward, but if one wants a quantitative idea that some prices or indicators are rising or falling faster than others, than a measure must be found. Given that I will be using this measure to compare various technical analysis indicators with price series, it must be scale and translation invariant.

Therefore, I use a measure of the normalized n-day rate of change (hereafter, called the NROC) by taking today's value, subtracting the value of the time series n-days ago, and dividing for the standard deviation for the time period. Formally:

- $NROC_{t,n} = (p_t - p_{t-n})/\sigma(p_{t-n}...p_t)$

Given that I am dividing by the standard deviation of the windowed time series, two time series that show equal change will have different NROC scores – the one with the "flatter" profile will score higher, since its standard deviation over the appropriate time period is lower. Although this property deviates from what is strictly desired for this measure, I do not feel it cripples the approach.

Once I have a measure of rate of change, than I can build a concept of divergence. I propose the following two formalizations:

- Divergence I: Take the difference between the n-day NROC of the indicator (here, the RSI) and the underlying price series n-day NROC; the {top,bottom} 5% of values indicate a {long,short} position. This finds the stocks with the maximum "absolute" divergence between the price series and the indicator.

- Divergence II: Go long each stock whose n-day indicator NROC is in the top 10% of all stocks, and whose price series n-day NROC is negative; conversely, short each stock whose n-day indicator NROC is in the bottom 10% of all stocks, and whose price series n-day NROC is positive. This strategy finds the stocks whose indicators are rapidly moving in one direction while their corresponding price series are stable or moving in the other direction.

A key decision to be made is the window over which the NROC is measured; since the literature is devoid of any attempts, I must arbitrarily pick a reasonable number: 50 trading days.

The table below presents the results, run on the four standard strategies described in the section 2.5.2 as well as the overbought/oversold strategy and the the two divergence strategies, under the interpretation that high or rising indicator values are bullish signals:

RSI Results (High Values Bullish Direction)					
Strategy:	Sharpe	Returns	Std Dev	Decline	Long/Short
Comprehensive	-2.63	-20.83%	9.87%	-73.97%	53%/46%
Crossover	-3.63	-25.88%	8.53%	-81.53%	30%/29%
Selective	-2.94	-58.72%	21.74%	-99.33%	6%/5%
Change	-5.10	-68.97%	14.55%	-99.86%	7%/5%
Overbought	-1.77	-17.61%	12.88%	-69.93%	22%/26%
Divergence I:	-1.30	-21.23%	20.30%	-75.17%	9%/5%
Divergence II:	-1.64	-35.48%	24.79%	-91.85%	7%/4%

Note that the four original strategies (comprehensive, crossover, selective, and change) produce the worst Sharpe ratios; the Sharpe ratios of the RSI-specific strategies hover below -2, indicating smaller – although still catastrophic by any measure – losses.

As with previous indicators, I re-ran the strategies, flipping the long/short signal direction. Here are the results:

RSI Results (High Values Bearish Direction)					
Strategy:	Sharpe	Returns	Std Dev	Decline	Long/Short
Comprehensive	-1.55	-10.03%	9.78%	-48.51%	46%/53%
Crossover	-2.85	-19.01%	8.48%	-69.82%	30%/29%
Selective	1.14 (*)	29.73%	21.48%	-32.64%	5%/6%
Change	-3.63	-47.41%	14.47%	-87.26%	5%/7%
Overbought	-.76	-4.71%	12.95%	-38.75%	26%/22%
Divergence I:	-2.24	-40.66%	20.49%	-95.06%	5%/9%
Divergence II:	-1.79	-39.31%	24.87%	-94.87%	4%/7%

The selective strategy is the only one that produces a positive Sharpe ratio, with a bootstrap p-value of less than .005. The overbought/oversold strategy comes in second with a -.79, and the two divergence strategies perform worse than in the high values bullish interpretation.

The yearly and quarterly differences in Sharpe ratio for the selective strategy against the benchmark are presented below, with p-values in parenthesis. The differences are statistically significant across the board.

Sharpe Ratio Differences over Benchmark		
Strategy:	Yearly	Quarterly
Selective	1.530 (.040)	1.477 (.017)

The NAV series for the selective strategy applied to the RSI is plotted in figure 2.6, together with the Russell 3000 and the selective strategy applied the exponential moving average for comparison Examining the graph, the RSI does well until early 2000, when it suffers losses; after this period of losses, it fails to show a clear trend up or down. This qualitative interpretation is disappointing, as it appears possible that the selective strategy applied to the RSI indicator stopped working entirely after the start of 2000.

While the results for the selective strategy are promising, the results for the conventional interpretations of the RSI – the overbought/oversold and divergence strategies – are disappointing. However, care must be taken to interpret these results; there

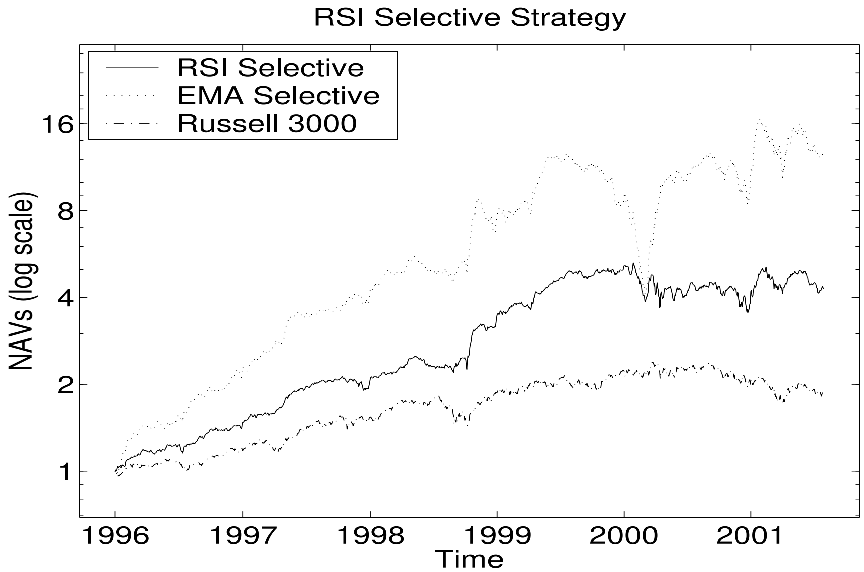

Figure 2.6: RSI Selective Strategy NAV Plot

are many possible alternate formalizations of the concepts of overbought/oversold and divergence, and just because one fails doesn't mean that they all will. And certainly, practitioners go out of their way to emphasize that the conventional wisdom of technical analysis indicator interpretation is meant to supplement human decision making processes, not replace it. Still, the fact that a relatively simple strategy can produce positive Sharpe ratios while reasonable formalizations of the ideas of overbought/oversold and divergence cannot come close to positive Sharpe ratios is suggestive.

2.8.2 Conclusions

- The two conventional interpretations of the RSI, the overbought/oversold interpretation and the divergence interpretation, fail to produce positive Sharpe ratios when applied with a priori reasonable formalizations.

- The selective strategy does produce a positive Sharpe ratio when applied to the RSI under the interpretation that high RSI values are bearish.

2.9 OBV: On Balance Volume

The on balance volume (OBV) attempts to measure under what conditions trading volume was generated – advances or declines. Mathematically, it's simple: if the volume was generated on an advance (if the close today is higher than yesterday), you add the volume to a running total; otherwise, you subtract. The intuition behind this is that trading volume generated on advances says different things about that stock than trading volume generated on declines. Formally:

- $X_t = sign(close_t - close_{t-1})$
- $OBV_t = OBV_{t-1} + X_t \cdot volume_t$

In order to test the on balance volume indicator, I apply the same strategies as used in the discussion of the accumulation distribution line above the standard four strategies – comprehensive, crossover, selective, and change, as well as the two divergence strategies.

OBV Results (High Values Bullish Direction)					
Strategy:	Sharpe	Returns	Std Dev	Decline	Long/Short
Comprehensive	-1.50	-11.05%	10.80%	-50.05%	36%/64%
Crossover	-2.60	-36.54%	16.03%	-92.31%	4%/4%
Selective	-1.13	-19.46%	21.83%	-74.06%	5%/5%
Change	-3.74	-50.34%	14.81%	-74.06%	7%/5%
Divergence I:	-.39	-.34%	13.91%	-32.57%	9%/5%
Divergence II:	-.64	-14.91%	31.43%	-77.15%	5%/2%

None of the strategies produce positive Sharpe ratios, although the second divergence strategy comes close. I present the results for the high values bullish interpretation below:

OBV Results (High Values Bearish Direction)					
Strategy:	Sharpe	Returns	Std Dev	Decline	Long/Short
Comprehensive	.17*	6.98%	10.76%	-30.03%	64%36%
Crossover	-.65	-5.29%	15.99%	-38.11%	4%/4%
Selective	.35*	12.79%	21.64%	-54.08%	5%/5%
Change	-4.11	-60.61%	16.00%	-38.49%	5%/7%
Divergence I:	-2.80	-33.98%	13.95%	-90.89%	5%/9%
Divergence II:	-1.76	-50.73%	31.68%	-98.24%	2%/5%

While none of these results beat our benchmark, two of them – the comprehensive strategy and the selective strategy – produce positive Sharpe ratios that are statistically significant by the bootstrap tests, and must be noted as intriguing possibilities for possible further exploration. Note that it is easily possible for strategies to fail to beat the benchmark and still register as statistically significant according to the bootstrap tests; remember that the bootstrap statistical significance only tests whether the results *are an accident of the data*, not if they beat the benchmark. The NAV series of the comprehensive and the selective strategies are plotted in figure 2.7. The NAV series for both OBV strategies follow the same basic pattern, essentially flat until the beginning of 2000 with a large drop in value, followed by a recovery to net positive levels. If the performance of these strategies were measured solely after the events of early 2000, their performance would be exceptional; but considered in the broader context, they are failures.

2.9.1 Conclusions

- The On Balance Volume indicator does not produce Sharpe ratios in excess of benchmark using any strategy.
- However, the comprehensive and selective strategies with the high values bullish interpretation do produce positive Sharpe ratios that are statistically significant under bootstrap testing.

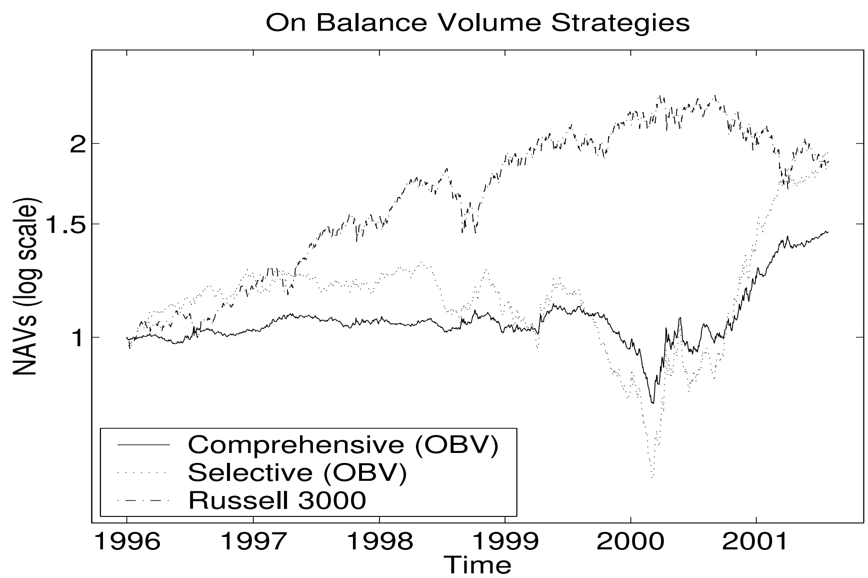

Figure 2.7: OBV Comprehensive and Selective Strategies NAV Plot

2.10 ADL: Accumulation/Distribution Line

The accumulation/distribution line (ADL) indicator attempts to measure if a stock is under accumulation. It's core idea is to identify whether or not volume is coming on advances or declines; if volume is coming on advances, then that hints there is buying pressure and accumulation even if the price isn't rising.

The core of the ADL is a measure called 'Close Location Value' (CLV), which attempts to quantify the notion of an advance or decline; this is then multiplied by the appropriate period volume. The CLV takes values over the range $(-1, 1)$. The value of 1 is achieved when the close is equal to the high, and is positive if the close is greater than the midpoint between the high and the low; correspondingly, -1 is achieved when the close is equal to the low. The ADL is simple the running sum of the daily CLV times the daily trading volume. Formally:

- $CLV_t = ((close_t - low_t) - (high_t - close_t))/(high_t - low_t)$
- $ADL_t = ADL_{t-1} + CLV_t \cdot volume_t$

According to conventional wisdom, signals produced by the accumulation/distribution line are bullish when there is positive divergence – the ADL is rising while the stock price is level or declining; bearish signals come out of a negative divergence, with a declining ADL with stable or rising stock price. Thus, I test the ADL indicator with the four standard strategies, and the two divergence strategies introduced above in section 2.8.1.

ADL Results (High Values Bullish Direction)					
Strategy:	Sharpe	Returns	Std Dev	Decline	Long/Short
Comprehensive	-1.91	-7.28%	6.51%	-42.84%	77%/23%
Crossover	-2.71	-50.66%	20.59%	-98.07%	3%/2%
Selective	-.57	-5.79%	19.14%	-46.34%	5%/5%
Change	-5.07	-75.20%	15.84%	-99.96%	7%/5%
Divergence I:	-1.30	-10.82%	12.27%	-57.10%	9%/5%
Divergence II:	-2.00	-31.40%	18.25%	-89.51%	6%/3%

Perhaps unsurprisingly, the results here are uniformly negative, although the selective strategy fails less catastrophically than the others. The results for flipping the interpretation of high ADL indicator values:

ADL Results (High Values Bearish Direction)					
Strategy:	Sharpe	Returns	Std Dev	Decline	Long/Short
Comprehensive	-.16	4.12%	6.47%	-42.84%	23%77%
Crossover	1.02 (*)	25.94%	20.40%	-23.58%	2%/3%
Selective	-.27	-1.00%	19.01%	-37.54%	5%/5%
Change	-1.61	-27.77%	20.42%	-23.64%	5%/7%
Divergence I:	-2.34	-23.57%	12.30%	-78.64%	5%/9%
Divergence II:	-1.88	-29.20%	18.21%	-86.04%	3%/6%

Here the divergence strategies also fail; the comprehensive and selective strategies come tantalizingly close to a positive Sharpe ratio and the crossover strategy produces a promisingly positive Sharpe ratio. However, it fails to achieve statistical significance in our test of Sharpe ratio differences:

Sharpe Ratio Differences over Benchmark		
Strategy:	Yearly	Quarterly
Crossover	.269 (.142)	.501 (.461)

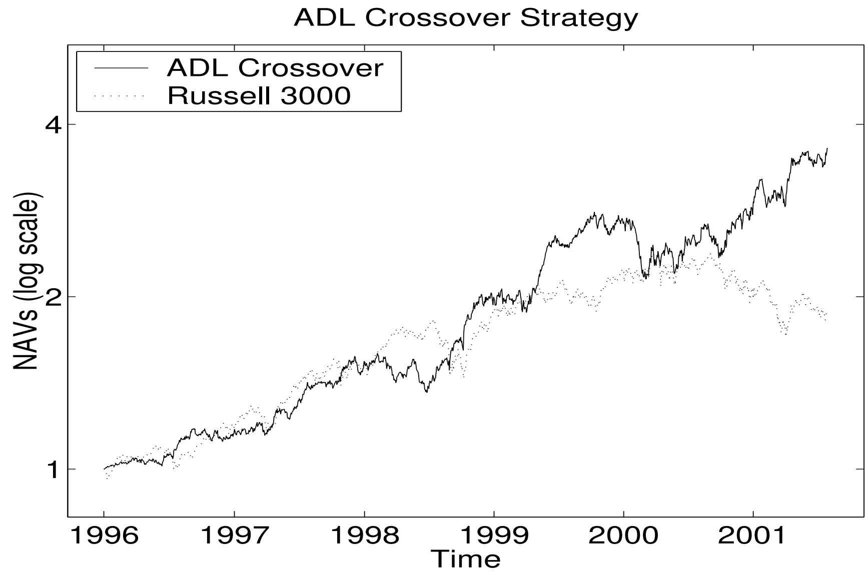

Figure 2.8: ADL Crossover Strategy NAV Plot

Figure 2.8 plots its NAV series, with the Russell 3000 for comparison. Intriguingly, its performance mirrors that of the Russell 3000 relatively closely (at least, compared to the other technical indicator/strategy combinations) until the Russell 3000 begins its decline in the spring of 2000, when the ADL indicator derived strategy maintains relatively steady growth. This helps explain the lack of statistical significance – performance to the Russell 3000 is nearly identical for most of the time period.

2.10.1 Conclusions

- The crossover strategy, applied to the ADL indicator under the interpretation that high ADL values are bearish, does produce a positive Sharpe ratio. However, the differences in yearly and and quarterly Sharpe ratios between this strategy the Russell 3000 are not statistically significant.
- Trading strategies employing traditional interpretations of the ADL indicator do not produce positive Sharpe ratios.

2.11 Stochastic Oscillators

The Stochastic Oscillator compares the current close to the high/low range over a window of past prices. The intuition behind this is that closing levels close to the top of the range indicate buying pressure; levels close to the bottom of the range suggest selling pressure. The Stochastic Oscillator takes on high values when recent

Formally:

- $n - daylow_{t,n} = min(low_{t-n}...low_t)$
- $n - dayhigh_{t,n} = min(high_{t-n}...high_t)$
- $\%K_{t,n} = 100 \cdot (close_t - n - daylow_{t,n})/(n - dayhigh_{t,n} - n - daylow_{t,n})$
- $\%D_n = SMA_3(\%K)$

In addition, there is an overbought/oversold interpretation with canonical values at 80 and 20. I test this strategy as well. The results are presented in the table below.

Stochastic %K (High Values Bullish Direction)					
Strategy:	Sharpe	Returns	Std Dev	Decline	Long/Short
Comprehensive	-4.78	-44.00%	10.28%	-96.10%	52%/47%
Crossover	-5.77	-48.85%	9.36%	-97.65%	36%/34%
Selective	-4.03	-84.05%	22.16%	-100%	5%/4%
Change	-4.72	-83.09%	18.68%	-100%	7%/4%
Overbought	-2.69	-19.49%	9.16%	-70.29%	40%/46%

Unsurprisingly, these results are uniformly negative. Now, with the high values Bearish interpretation:

Stochastic %K (High Values Bearish Direction)					
Strategy:	Sharpe	Returns	Std Dev	Decline	Long/Short
Comprehensive	1.28 (*)	18.03%	10.09%	-21.45%	47%/52%
Crossover	0.97 (*)	14.04%	9.63%	-15.78%	34%/36%
Selective	2.64 (*)	62.90%	21.89%	-28.31%	4%/5%
Change	-.55	-4.80%	18.24%	-65.03%	4%/7%
Overbought	-0.65	-.81%	9.16%	-29.86%	46%/40%

The results here are more promising, with the comprehensive, crossover, and selective strategies providing positive Sharpe ratios. The selective strategy in particular produces a strong 2.64 Sharpe ratio.

The differences in Sharpe ratios over benchmark are presented below:

Sharpe Ratio Differences over Benchmark		
Strategy:	Yearly	Quarterly
Comprehensive	1.642 (.145)	1.487 (.091)
Crossover	1.195 (.267)	.780 (.352)
Selective	4.967 (.175)	5.559 (.017)

Despite differences in Sharpe ratios, the both the crossover and comprehensive strategy fail to provide statistical significance for yearly and quarterly results. The selective strategy performs extremely strongly, producing strong differences across the board and statistical significance for the quarterly differences.

Figure 2.9 plots the results of the comprehensive and selective strategies below. The selective strategy performs extremely well, with stable, large returns except for a perhaps expected dip in the beginning of 2000. The comprehensive strategy doesn't produce returns as strong, but what returns it produces are accompanied by far less volatility.

2.11.1 Conclusions

- The Stochastic %K indicator produces positive Sharpe ratios in excess of benchmark with the comprehensive, crossover, and selective strategies, under the interpretation that high values of the indicator are bearish signals. However, only the selective strategy produces acceptable statistical significance.
- The traditional overbought/oversold interpretation applied to the Stochastic %K indicator fails to provide positive Sharpe ratios.

2.12 CCI: Commodity Channel Index

The Commodity Channel Index, developed by Daniel Lambert, is a somewhat baroque indicator requiring complicated computation. First, the true price (TP) is computed,

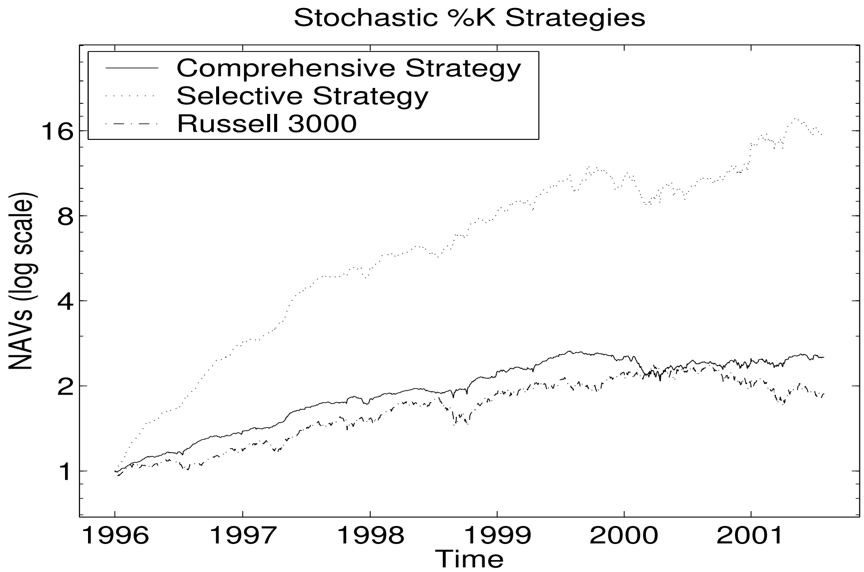

Figure 2.9: Stochastic %K Comprehensive and Selective Strategies NAV Plot

which is simply the average of the high, low, and closing price. Then, the simple moving average of the true price (SMATP) is computed over an n-day. Then, a factor called the mean deviation (MD) is computed by taking the mean of the difference between the current SMATP and the TP for each of the last n days. Finally, the CCI is computed as the difference between the TP and the SMATP, divided by the mean deviation times a scaling constant (canonically .015, designed to that the CCI is within the range $\{-100, 100\}$ 80% of the time. Formally:

- $TP_t = (high_t + low_t + close_t)/3$
- $SMATP_{t,n} = SMA_{t,n}(TP_t)$
- $MD_{t,n} = (1/n) \cdot \sum_{i=1}^{n} |SMATP_{t,n} - TP_{t-i}|$
- $CCI_{t,n} = (TP_t - SMATP_{t,n})/(.015 * MD_{t,n})$

The CCI was originally designed for cyclical commodities markets, and the typical choice of n is determined by the expected market cycle. Since here, I am applying the CCI to equities with no real cyclical component, I arbitrarily choose $n = 60$.

To test the CCI, I apply the comprehensive, crossover, selective and change strategies, as well as the overbought/oversold strategy (with the overbought level at 100 and the oversold level at -100), and the two divergence strategies. First, the results under the high indicator values are bullish interpretation:

CCI Results (High Values Bullish Direction)					
Strategy:	Sharpe	Returns	Std Dev	Decline	Long/Short
Comprehensive	-2.20	-20.05%	11.44%	-72.47%	56%/43%
Crossover	-3.95	-36.59%	10.57%	-92.35%	17%/16%
Selective	-3.17	-63.93%	21.78%	-99.68%	5%/5%
Change	-3.76	-72.14%	20.55%	-99.92%	7%/4%
Overbought	-2.51	-23.70%	11.47%	-78.57%	20%/26%
Divergence I:	-1.55	-31.53%	23.70%	-88.77%	9%/4%
Divergence II:	-1.23	-33.19%	31.29%	-90.10%	6%/2%

None of the strategies produce excess returns. Intriguing are the high figures in the long/short holdings columns – compare the crossover and overbought strategies; all of them have large long and short positions, indicating that the CCI is jumping all over the place, crossing the centerline and the overbought/oversold levels frequently.

The results for the high values are bullish interpretation are presented below:

CCI Results (High Values Bearish Direction)					
Strategy:	Sharpe	Returns	Std Dev	Decline	Long/Short
Comprehensive	-0.14	3.62%	11.28%	-34.52%	43%/56%
Crossover	-0.30	-2.00%	10.47%	-26.62%	16%/17%
Selective	2.63 (*)	61.52%	21.40%	-34.94%	5%/5%
Change	-1.35	-22.12%	20.18%	-76.37%	4%/7%
Overbought	-.02	4.87%	11.51%	-19.33%	26%/20%
Divergence I:	-1.42	-28.78%	23.98%	-89.05%	4%/9%
Divergence II:	-1.44	-40.26%	31.58%	-95.69%	2%/6%

Here, the selective strategy produces a positive Sharpe ratio, with statistical significant on the bootstrap tests; the rest are strongly negative, except for the overbought and comprehensive strategies. The table below presents the data for the yearly and quarterly differences in Sharpe ratio between the selective strategy and the Russell 3000.

Sharpe Ratio Differences over Benchmark		
Strategy:	Yearly	Quarterly
Selective	2.84 (.019)	3.48 (< .001)

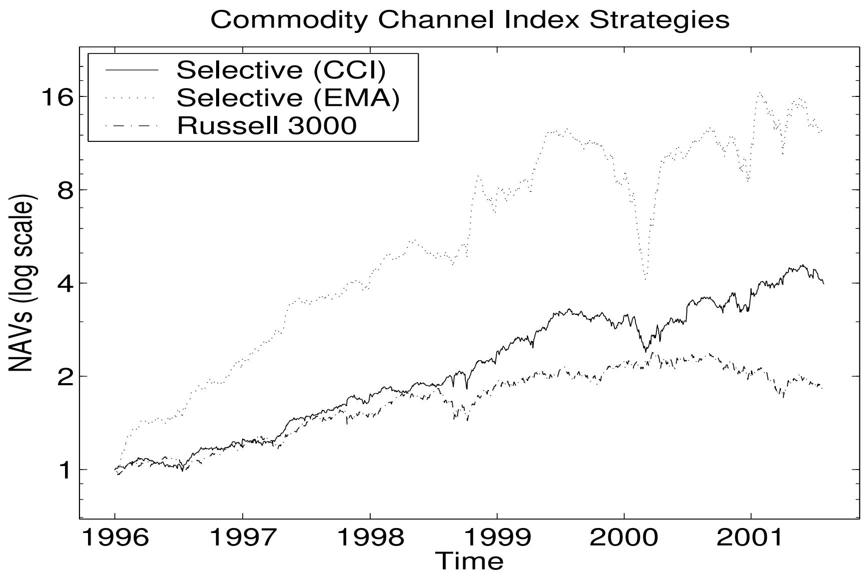

Figure 2.10: CCI with Selective Strategy NAV Plot

The differences in Sharpe ratio are strong, and statistically significant.

The NAV series for the selective strategy is presented below, in figure 2.10, with the exponential moving average selective strategy for comparison. Interestingly, the NAV profile of the CCI selective strategy looks similar to that of the EMA selective strategy, with the same significant drop in early 200, but with far less volatility.

2.12.1 Conclusions

- The CCI indicator produces a positive, statistically significant Sharpe ratio with the selective strategy, under the interpretation that high values of the indicator are bearish signals.
- The traditional interpretations of the CCI indicator – high values are bullish, with overbought/oversold and divergence interpretations – fail to produce positive Sharpe ratios.

2.12.2 PVO: Percentage Volume Oscillator

The Percentage Volume Oscillator is similar in concept to the MACD applied to volume instead of prices; it's simply the 12-day exponential moving average of the volume minus the 26-day moving average of the volume, divided by the 12-day exponential moving average. This number is than multiplied by 100 to get a score that varies from positive 100 to unlimited negative values. Formally:

- $PVO_t = 100 \cdot (EMA_{t,12}(V) - EMA_{t,26}(V))/EMA_{t,12}$

The results for the standard strategies applied to the PVO indicator are presented below:

PVO Results (High Values Bullish Direction)					
Strategy:	Sharpe	Returns	Std Dev	Decline	Long/Short
Comprehensive	1.68 (*)	12.36%	4.30%	-4.85%	42%/58%
Crossover	-.65	2.08%	4.70%	-13.78%	31%/34%
Selective	3.96 (*)	57.24%	13.17%	-11.88%	5%/5%
Change	-2.76	-27.48%	11.82%	-83.93%	7%/5%

These results are an enormously pleasant surprise. Immediately note the impressive 4.21 Sharpe ratio for the selective strategy – the best Sharpe ratio of any of the indicator/strategy combinations seen so far – and the 1.74 Sharpe ratio for the comprehensive strategy, the largest Sharpe ratio produced by the comprehensive strategy on any indicator. Also, this is the first technical analysis indicator to ever score well in the high scores bullish interpretation.

The yearly and quarterly Sharpe ratio differences are presented below, with the selective strategy significant across the board and the comprehensive strategy significant for quarterly differences.

Sharpe Ratio Differences over Benchmark		
Strategy:	Yearly	Quarterly
Comprehensive	1.231 (.044)	.647 (.354)
Selective	3.248 (.005)	3.239 ($<$ .001)

The results for the selective strategy are strongly statistically significant across the board. The comprehensive strategy, while just meeting the 5% cutoff on yearly

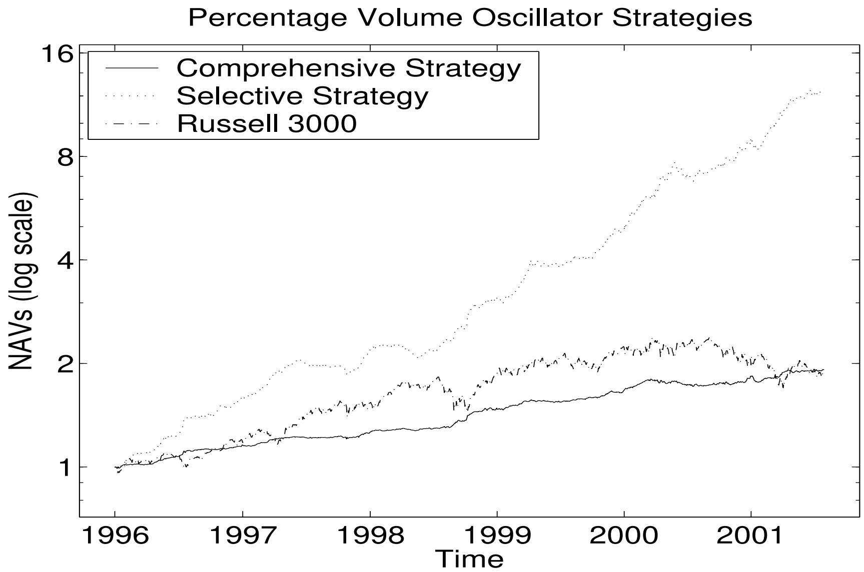

Figure 2.11: PVO Comprehensive and Selective Strategies NAV Plot

differences, fails on quarterly differences. The performance for the high values bearish interpretation is given below:

PVO Results (High Values Bearish Direction)					
Strategy:	Sharpe	Returns	Std Dev	Decline	Long/Short
Comprehensive	-8.92	-33.41%	4.33%	-89.70%	58%/42%
Crossover	-8.58	-35.47%	4.73%	-91.36%	34%/31%
Selective	-5.18	-63.67%	13.29%	-99.66%	5%/5%
Change	-6.67	-74.30%	11.92%	-99.95%	5%/7%

Given the strong performance in the high value bullish interpretation, it is no surprise that performance here is poor across the board.

To gain intuition, the results for the comprehensive and selective strategies applied to the PVO with the high values bullish interpretation are plotted below, in figure 2.11. The comprehensive strategy produces low, steady returns, while selective strategy behaves admirably, producing large, steady returns with low volatility. Note that neither of the strategies produce are large dip in early 2000 like many of the other successful technical analysis indicators.

2.12.3 Conclusions

- The PVO indicator produces a positive Sharpe ratio with the comprehensive and selective strategies, under the interpretation that high values of the indicator are bullish signals, although only the selective strategy meets our statistical significance tests in Sharpe ratio differences.

- The PVO is the first indicator to show positive results under the interpretation that high scores are bullish; it is also the only indicator that relies solely on volume and makes no use of past price information.

2.12.4 CMF: Chaikin Money Flow

The Chaikin Money flow indicator attempts to measure the notion of a stock being under accumulation. It consists of the n-day sum of the accumulation/distribution line divided by the n-day sum of the volume. The canonical n in this case is 21. Formally:

- $CMF_t = \sum_{i=0}^{n-1} ADL_{t-i} / \sum_{i=0}^{n-1} V_{t-i}$

The Chaikin Money Flow indicator is traditionally bullish when it is positive, bearish when negative, with the strength of signal depending both on the length of time the signal has been above or below zero, and the magnitude of the signal. However, in the practitioner literature the Chaikin Money Flow is rarely used in isolation, rather, as a confirmation of trends uncovered by other indicators.

CMF Results (High Values Bullish Direction)

Strategy:	Sharpe	Returns	Std Dev	Decline	Long/Short
Comprehensive	-1.27	-2.75%	6.24%	-29.60%	77%/23%
Crossover	-0.79	-19.32%	31.07%	-75.55%	1%/1%
Selective	-1.38	-10.48%	11.35%	-59.95%	5%/5%
Change	-7.11	-68.09%	10.30%	-99.84%	7%/5%

CMF Results (High Values Bearish Direction)					
Strategy:	Sharpe	Returns	Std Dev	Decline	Long/Short
Comprehensive	-0.59	1.50%	6.22%	-18.83%	23%/77%
Crossover	-0.44	-8.51%	31.02%	-66.57%	1%/1%
Selective	-0.29	1.89%	11.32%	-33.77%	5%/5%
Change	-3.62	-31.64%	10.18%	-88.54%	5%/7%

None of the strategies, under either of the interpretations, produces positive Sharpe ratios, although for the high values Bearish direction the results are mostly flat for the comprehensive, crossover, and selective strategies.

2.12.5 Conclusions

- The CMF fails to produce any positive Sharpe ratios, using any of the four standard strategies under either the high values are bullish or high values are bearish interpretation.

2.13 Summary of Results So Far

Now that I have run through all of the technical analysis indicators with appropriate strategies, I review which strategies produced interesting results. I break them into three categories:

- The Well-Behaved list: Indicator/strategy combinations that produce positive Sharpe ratios that are statistically significant both according to the bootstrap tests, and one of the benchmark Sharpe ratio differences tests.
  - EMA, crossover, high values bearish direction
  - EMA, selective, high values bearish direction
  - EMA, change, high values bearish direction
  - MACD, selective, high values bearish direction
  - MACD, change, high values bearish direction
  - RSI, selective, high values bearish direction

- Stochastic %K, selective, high values bearish direction
- CCI, selective, high values bearish direction
- PVO, selective, high values bullish direction

- The Almost list: Indicator/strategy combinations that produce positive Sharpe ratios in excess of our benchmark, but fail the difference in Sharpe ratio statistical significance tests.
  - ADL, crossover, high values bearish direction
  - Stochastic %K, comprehensive, high values bearish direction
  - Stochastic %K, crossover, high values bearish direction
  - PVO, comprehensive, high values bullish direction

- The Promising list: Strategies the produce positive Sharpe ratios not in excess of the benchmark. These are strategies that could conceivably be fine tuned into usefulness.
  - EMA, comprehensive, high values bearish direction
  - OBV, comprehensive, high values bearish direction
  - OBV, selective, high values bearish direction

2.14 Connections Between Indicators

Given that nearly all of the of the indicator/strategy combinations that produce results use the selective strategy, with the high values bearish direction, an obvious question to ask is how much information is duplicated between the various indicator/strategy combinations?

The table below presents the correlations of the return series between all of the indicator/strategy combinations that produced Sharpe ratios statistically significant in both the bootstrap and benchmark Sharpe ratio difference tests.

Correlation Matrix										
Strategy:	Russ 3000	EMA Cro	EMA Sel	EMA Chg	MACD Sel	MACD Chg	RSI Sel	Stoch Sel	CCI Sel	PVO Sel
Russell 3000	1	-.03	0	-.06	-.01	-.05	.08	.02	.01	-.08
EMA Cross	-.03	1	.35	.44	.48	.56	.56	.53	.66	-.06
EMA Selec	0	.35	1	.12	.67	.11	.53	.36	.63	-.10
EMA Change	-.06	.44	.12	1	.05	.71	.10	.44	.30	-.01
MACD Selec	-.01	.48	.67	.05	1	.11	.73	.39	.73	-.12
MACD Change	-.05	.56	.11	.71	.11	1	.32	.57	.46	-.02
RSI Selec	.08	.56	.53	.10	.73	.32	1	.54	.78	-.14
Stoch. %K Selec	.02	.53	.36	.44	.39	.57	.54	1	.55	-.11
CCI Selec	.01	.66	.63	.30	.73	.46	.77	.55	1	-.09
PVO Selec	-.08	-.06	-.10	-.01	-.12	-.02	-.14	-.11	-.09	1

Some interesting comments on these results:

- None of the indicator/strategy combinations are very strongly correlated with the Russell 3000 – the strongest correlation is the relative strength index (RSI) with selective strategy at .08.
- The EMA, MACD, RSI, and CCI selective strategies are in fact strongly correlated with each other, with correlation coefficients of greater than .5 in all cases. The Stochastic %K selective strategy is more weakly correlated with this group.
- The percentage volume oscillator selective strategy is essentially uncorrelated with any of the other strategies.
- The two change strategy entries – EMA and MACD, are highly correlated with a correlation coefficient of .71, and are both moderately correlated with the stochastic %K selective strategy.

In summary, while there is a cluster of strongly correlated indicator/strategy combination that feature the selective strategy, not all of the indicator/strategy combinations using the selective strategy are highly correlated with each other, and a few of the indicator/strategy combinations are not positively correlated with anything. Many of the indicators really appear to be uncovering genuinely unique information.

2.15 Practical Problems, or Why Isn't Everybody Doing This?

Given that so many of the indicators/strategy combinations produce Sharpe ratios clearly superior to an appropriate buy-and-hold benchmark, seemingly flying in the face of the efficient market hypothesis, a natural question arises: why isn't everybody using them?

I have no answer for this question. But, I suspect it revolves around two key assumptions: transaction costs and order fulfillment.

I assume that I can purchase or sell short these securities at the daily closing price. I also assume that one way transaction costs are .1% one way. Are these assumptions reasonable? I cannot directly test them, although they are in line with previous practice in the literature. However, I believe it is still important to think about them.

I can test the sensitivity of the above results to these assumptions, directly in the transaction cost case, indirectly in the order fulfillment case. The next two sections address those two issues.

2.15.1 Daily Turnover, and Maybe Order Fulfillment

I cannot test order fulfillment directly. However, almost by definition liquid stocks, stocks that trade frequently, are easier to trade than illiquid stocks. Therefore, understanding how the results presented above vary with the liquidity (which I proxy by mean daily turnover) gives at least some understanding of the problems these techniques could face in real application. In general, stocks with more daily turnover will be easier to buy and sell at the stated price. So, understanding how the techniques fare with stocks of varying turnover gives some idea of their potential application to the real world.

I split the stock universe into quintiles based on average turnover – daily trading volume multiplied by the daily closing process – and evaluates the successful indicator/strategy combinations on each quintile. The results are presented below:

Sharpe Ratios by Turnover Quintile					
Quintile:	0-20%	20-40%	40-60%	60-80%	80-100%
EMA Crossover	10.03	1.03	.01	-.16	-.65
EMA Selective	4.07	2.31	.70	.55	.17
EMA Change	19.34	2.59	.16	-.90	-1.48
MACD Selective	4.08	.94	.70	1.02	.47
MACD Change	13.73	2.47	.72	.11	.33
RSI Selective	1.40	1.13	.57	.30	.72
Stochastic %K Selective	1.63	1.65	.70	.05	.82
CCI Selective	5.51	2.00	.54	.67	.97
PVO Selective	2.00	3.27	2.20	1.63	-.56

Examining the data, it is clear that the success of the indicator/strategy combinations is strongly dependent on which group of stocks it's applied to. Note that for most indicator/strategy combination, by far the best Sharpe ratio is provided by the most illiquid quintile of stocks. For some – especially the EMA and MACD using the change strategies, and the EMA with crossover strategy, the differences border on the ludicrous, with absurdly high Sharpe ratios of over 15 in the lowest turnover quintile, and negative or very small Sharpe ratios in the higher quintiles.

Most of the indicators using the selective strategy – EMA, MACD, RSI, Stochastic %K, and the CCI, follow a similar, although less extreme pattern of high results in the low-turnover quintiles and declining success as turnover grows.

It's clear from this data that order fulfillment could potentially be a huge problem, especially as many of the strategies only work well on stocks with low turnover.

2.15.2 Transaction Costs

In this section I examine the performance of selected strategies as transaction cost increases.

Before I present the results, let me clear up a common confusion about transaction costs among the lay public. When brokerages advertise, they claim $10 per trade fees as the transaction costs. However, the true transaction costs are much higher, and depend on the difference between the bid and ask spread. Stocks don't trade at a single price: market makers offer stocks for sale at the bid price, and offer to buy stocks at the ask price. This spread is what allows market makers to make a profit and provide liquidity, and somewhere in the middle is the true price. The real transaction cost is the difference between what you pay, or receive, and this true price.

In any case, the table below presents Sharpe ratios for varying indicator/strategy combinations as transaction costs increase. The asterisks here represent statistical significance on quarterly differences in Sharpe ratio between the indicator/strategy combination and the benchmark, at the $p < .05$ level.

Sharpe Ratios as Transaction Costs Increase

Transaction Costs:	.1%	.15%	.2%	.25%	.3%
EMA Crossover	1.98*	.49	-.82	-1.97	-2.98
EMA Selective	1.43*	1.17*	.92	.69	.47
EMA Change	2.28*	-.57	-2.39	-3.54	-4.26
MACD Selective	1.13*	.96*	.80	.64	.49
MACD Change	3.37*	1.44	-.07	-1.25	-2.15
RSI Selective	1.14*	.31	-.40	-1.02	-1.55
Stochastic %K Selective	2.64*	.55	-.94	-2.01	-2.77
CCI Selective	2.63*	1.73*	.94	.24	-.38
PVO Selective	3.96*	2.42	1.07	-.09	-1.11

The performance of the change strategy, applied both to exponential moving average and the MACD, decreases rapidly as transaction costs increase. Few strategies maintain statistical significance at even the .15% one-way transaction cost level. However, many of the strategies do continue to produce strong Sharpe ratios with higher transaction costs.

The key to understanding sensitivity to transaction costs for each of the indicator/strategy combinations presented above is the holding period; the average length that each indicator/strategy combination holds a stock long or short. The table below presents the regression coefficients for the Sharpe ratio/transaction cost relation, as a measure of sensitivity to transaction costs, along with the mean holding period for each indicator/strategy combination.

Sharpe Ratios as Transaction Costs Increase		
Indicator/Strategy:	Regression Slope	Holding Period
EMA Crossover	-24.76	4.97
EMA Selective	-4.80	43.74
EMA Change	-32.10	2.20
MACD Selective	-3.20	23.09
MACD Change	-27.46	4.72
RSI Selective	-13.42	9.72
Stochastic %K Selective	-26.76	4.34
CCI Selective	-15.02	12.13
PVO Selective	-25.30	8.34

Unsurprisingly, there is a strong relationship between the holding period and the sensitivity to transaction costs; short holding periods produce high sensitivities to transaction costs.

For a computer scientist, the causal link is clear, and the temptation to tinker with increasing the holding periods as a way of coping with increased transaction costs is almost unbearable. But I will resist, for this chapter at least.

2.15.3 Problems, Problems

It's pretty clear that implementing these strategies in the real world faces potential obstacles from both higher transaction costs, and difficulties in trading low turnover stocks.

However, with both of these problems, there are plausible solutions – some indicators work better than others for high turnover stocks, and increasing holding periods of the trading strategies holds out the promise of ameliorating sensitivity to transaction costs.

Out of deference to charges of data mining, I do not attempt to address these issues here. But, it is a perfectly legitimate endeavor to address them through machine learning, and much of the next chapter does exactly that.

2.16 Conclusions

I boil the conclusions of this chapter down to the following points:

- Many of the technical analysis indicators can be used with strategies to produce statistically significant positive Sharpe ratios, both tested using bootstrap statistics against the null hypothesis that the results are produced by chance, and using traditional t-tests against the null hypothesis that the Sharpe ratios are less than or equal to the benchmark of the Russell 3000.

- However, very few of the strategies that produce positive Sharpe ratios correspond to the conventional interpretations of the technical analysis indicators.

- In particular, most technical analysis indicators are interpreted such that high or rising values of the indicators lead to bullish signals. I find that almost without exception, the strategies that produce positive Sharpe ratios interpret the technical analysis indicators in exactly the opposite way. The Percentage Volume Oscillator, the only indicator to depend solely on volume information, is the only exception.

- In nearly all of the indicators that produce statistically significant positive Sharpe ratios, there is a dip in returns in the early months of 2000, corresponding to the market downturn.

- The return series produced by the different indicator/strategy combinations are not correlated with the return series of the Russell 3000, and in general are not strongly correlated with each other, although there is a group of indicator/strategy combinations utilizing the selective strategy that are moderately inter-correlated.

- The indicator/strategy combinations show sensitivity to transaction costs, which is strongly related to the mean number of days each indicator/strategy combination holds positions in each stock.

- When the stock universe is split into quintiles by average daily turnover, there is great variation in performance for the indicator/strategy combinations over each quintile. In general, the indicator/strategy combinations perform better on the lower-turnover quintiles, although that trend is not universal.

Chapter 3

Learning Trading Rules

3.1 Introduction

The previous chapter cataloged the performance of a wide variety of technical analysis indicators with reasonable trading strategy formulations. In many cases, the data was suggestive that some tinkering with the algorithm could produce improved results. I carefully refrained from such engineering, out of concern for charges of data mining in the economic sense.

But applying machine learning to improve the algorithms is a legitimate enterprise – so long as the learned algorithms maintain their success on data drawn outside of the training sample. In fact, using machine learning helps to guard against charges of data mining in the economic sense. Investigating individual trading rules leaves open the charge of selection from thousands of possible trading rules; however, by attempting to learn trading rules that generalize out of sample, I are selecting from the universe of methods of learning training rules, which is far smaller, and was developed with a careful eye on issues of overfitting and out of sample generalization.

But the key question remains: can trading rules be learned that produce excess Sharpe ratios out of sample?

3.1.1 Goals of this Chapter

In this chapter, that question is put in the service of several goals.

- Reinforce the conclusions of the previous chapter. Observing that many technical analysis derived trading rules produce high Sharpe ratios is certainly a strong result. But showing that these relationships can be learned – that they stay valid for out of sample data – is stronger evidence.

- Understand what adjustments need to be made to traditional machine learning methodologies to cope with the specific challenges of learning trading rules in a financial context

3.1.2 A Note on Representation

Although the possible techniques and representations that can be applied to this problem are innumerable, this chapter focuses on the specific problem of learning complex trading rules based on simple technical trading indicators. I offer the following reasons for this decision:

First, as chapter 2 and a significant amount of previous research (discussed in section 2.2) suggests, technical analysis indicators do have predictive power under some conditions. Second, given their widespread use by practitioners, technical analysis indicators are subjects of study in their own right, and a deeper understanding of their use as building blocks for more complex strategies aids in our understanding of the indicators themselves. Third, technical analysis indicators are a relatively straightforward representation, and trading rules built out of a combination of technical analysis indicators are transparent to human understanding – certainly when compared to techniques such as neural networks. And finally, the next two chapters deal around new data derived from message boards and news stories, and this representation offers a natural way to integrate that data and explore its effectiveness.

3.1.3 Chapter Organization

The format of this chapter is straightforward: first, a discussion of previous work, focusing both on work originating from economics and on work from artificial intelligence. Second, a brief note about the data set and methodology used in the subsequent experiments. Then, the core empirical work of the chapter, split into two parts. The first part focuses on how *not* to learn trading rules – in particular, how the noise in financial data forces a rethinking of some machine learning assumptions.

The second part presents a positive methodology for learning complex trading rules from atomic components of technical analysis indicators. I start with the challenges in learning atomic rules, and build up to issues of combining into complex rules and populations of rules.

3.2 Previous work

3.2.1 The View from AI and Machine Learning

Forecasting financial markets is one of the canonically sexy tasks in Artificial Intelligence. Probably every AI and machine learning algorithm ever developed has been applied to forecasting financial markets. The amount of work is staggering. However, since this chapter is designed to focus on a specific tradition of economics research, I will focus my discussion there, and refer the reader to my discussion in the introduction, in section 1.3.2, for a summary of interesting applications of AI to markets.

3.2.2 The View from Economics and Finance

The history of AI in research on the finance side of the fence is relatively limited.

There is a strong tradition of applying AI to model human reasoning – as economics has attempted to move away from pure analytically tractable equilibrium models, AI has often been used to model human learning. The rational expectations literature is full of applications of techniques originally developed in artificial intelligence.

But, given the power of the efficient market hypothesis, economists have traditionally not been able to justify the application of artificial intelligence towards forecasting markets. Recently, however, the use of machine learning methods as diagnostics for financial market predictability has made its way into some finance papers.

The key path of economics research that this chapter expands upon evolved out of the trading rule research of Brock, Lakonishok, and LeBaron [22] and subsequent papers, which explored the use of pre-defined trading rules.

The shift to applying machine learning was straightforward: if simple trading rules produced excess returns, why not learn the trading rules using genetic programming techniques? Allen and Karjalainen's paper [15] learning trading rules on the S&P 500 was probably the first effort in this direction; Chris Neely and his collaborators further extended this school of thought with papers on foreign exchange markets [71, 70]. This is the extent of this line of economics research.

I feel this basic approach – using machine learning-derived trading rules as an operational test of the efficient market hypothesis – is extremely promising. It has one key advantage: using pre-chosen technical analysis indicators leaves open the critique of data mining over the space of technical analysis indicators, as in White's critique [93]; by learning the technical analysis indicators, this charge is ameliorated.

Of course, accusations of data mining specific technical analysis techniques can morph into accusations of data mining, and this is a serious matter, but there I use methodologies to address this, discussed below.

However, what work has been done in this line of research has suffered from several problems. First, the economists have paid little attention to tailoring machine learning techniques to the financial domain – they have treated them like a black box. And worse, they have used a technique – specifically, genetic search – that is customarily implemented without the safeguards against overfitting that are standard procedure on other techniques, such as pruning decision trees ([77]). In addition, the algorithm design process presented was not transparent, leaving open charges of data mining the algorithm. And finally, these approaches were tested on small datasets – single indexes or small sets of securities.

There is some work which has applied genetic learners with appropriate concern for noise. In a somewhat different context, LeBaron [58], uses genetic algorithms to evolve neural network topologies for foreign exchange prediction; here, the genetic learner is used for model selection, and the evaluation of fitness is augmented by bootstrap and crossvalidation techniques.

It is the goal of this chapter to take the basic technique pioneered in the economics literature – genetic search over technical analysis-like rule constructions, and put it in a sound foundation by handling some of these methodological issues. First, in algorithm design, the top priority is fighting noise; making sure that the algorithm is appropriate to the domain. Second, I take steps to fight the possibility of overfitting the algorithm, discussed in section 3.3.1. And finally, I dramatically increase the number of securities under scrutiny (although, admittedly, the time window I explore is shorter, the total amount of data is dramatically increased).

3.3 Methodology & Data

For all of the following experiments, I use the following dataset, identical to that used in the technical analysis explorations of chapter 2: stocks present in the Russell 3000 as of June, 2001, which have a continual return series from March 1, 1998 to February 28, 2002. This represents 1768 stocks and 1005 trading days. Stock prices – closes, opens, highs, and lows, as well as trading volume – were gathered from the Yahoo! Finance quote server at `http://finance.yahoo.com` [9]. These prices are both split- and dividend-adjusted.

This data set differs from that discussed in the previous chapter for two reasons. First, I wanted to make some of it non-overlapping with the dataset from the previous chapter, to avoid charges that I am simply learning technical analysis rules that I know work already work on a specific time period. The final holdout set does not overlap temporally with any of the data from the technical analysis results in chapter 2.

Second, eventually I plan to integrate this learner with news data, and so the holdout set corresponds to the same time period for which I have news data.

3.3.1 A Note About Testing

Given the slippery nature of financial data – I am competing against the hypothesis that it's essentially all noise – it is essential to guard against overfitting. Financial data is qualitatively different from the data usually handled by machine learning, and requires special care.

Good Data Mining vs. Bad Data Mining

There is a confusion of terminology between AI and economics that needs to be addressed here. Both use the phrase "data mining" – but to mean dangerously different concepts. In economics, "data mining" is the process of making spurious generalizations are artifacts of the noise in the data and that to not generalize to out of sample data.

In AI, "data mining" is a term for automating the process of finding patterns in data that *do* generalize out of sample. One is left with the terminologically troubling dictum that in order to perform data mining in the AI sense, one must avoid data mining in the economic sense.

Given the confusion these two meanings engender, the term "data snooping" emerging to cover the economic sense of "data mining". I certainly approve of this shift in terminology; however, data mining still is the most

The word the machine learning community uses to mean the bad kind of data mining, in an algorithm context, is overfitting (although term multiple hypothesis testing problem is also sometimes used). To (hopefully) avoid confusion, from now on I will use the term overfitting to describe the undesirable outcome of reaching conclusions on the training set that do not generalize to novel data.

Guarding Against "Bad" Data Mining and Overfitting

In any case, financial data is so noisy that overfitting is an enormous concern in finance. The canonical worry is that by trying a large enough number of trading rules, some will be found to be effective by mere chance. There is a fair amount of work examining the relationship between the size of the universe of rules being considered (see, for example, [93, 96]) and the testing of statistical significance.

In a sense, the current approach is a guard against this kind of overfitting – instead of selecting a few rules from a large universe of possible rules, the AI approach selects a *method of generating rules* from a smaller universe of methods of generating rules. However, this opens the approach to criticisms that I am overfitting over the universe of possible methods of generating rules (ie, the universe of algorithms). There is an unavoidable conflict here. One of the core elements of AI research is process of iterative algorithm tweaking and performance measurement, but this carries the corresponding danger that the tweaking will narrow the scope of the algorithm so much that it will fail to generalize to novel data. This danger looms particularly large in financial data, given the levels of noise present. To guard against this, I establish the following policy:

- Hold out a final testing set of data. This data will not be touched until the algorithm design process is complete.
- Split the remaining data into training and testing sets.
- Perform algorithm design on only this data – develop the algorithm by examining performance on the test set, guided by sound a priori principles from finance

and artificial intelligence.

- Then, only when the algorithm has been settled, verify the conclusions based on the 'holdout' set.

I perform this split in time; I set aside the last quarter of the data as the holdout set. All algorithm design conclusions made in sections 3.5 and 3.5 are reached using the third quarter of the data as a test set. Then, they are checked to see if they hold over the holdout dataset, the last quarter of the data, in section 3.7. In concrete temporal terms, the test set data for the algorithm design phase runs from February 28th, 2000 to February 25th, 2001. The holdout set data runs from February 26th, 2001 to February 28th, 2002.

It is interesting to compare this approach specifically to White's reality check method [96]. In the Reality Check method, the statistical significance tests are adjusted by a factor that depends on the size of the universe of trading rules, to adjust for potential data snooping biases.

In a machine learning context, I am no longer searching over a universe of trading rules; rather, I am designing an algorithm which selects trading rules. The real danger is data snooping among the universe of *learning algorithms*. The the situation is less clearcut than in a universe of static trading rules; there is a potentially enormous universe of trading rules to select from. However, selecting a machine learning algorithm does not consist of a data snooping-style search over possible algorithms; rather, tradition and experience have shown what techniques work well in broad application. Rather than applying a Reality Check style test that controls for the size of the possible algorithm universe, I believe that the approach taken here – committing to a final algorithm before testing results on out of sample data is the proper approach.

3.4 Representation

At the core of what I am trying to learn is the idea of an atomic trading rule, a method to map past price and volume information into buy and sell signals.

This section describes the structure of these atomic trading rules, built with components studied in chapter 2. Given that the learning discussed here involves stochastic search over the space of possible rules, I also discuss how to generate random rules, and how to perturb rules – the algorithm I use to generate a new rule from an old rule that is still similar in some sense.

Each atomic trading rule consists of a technical analysis indicator and a trading strategy, just like the combinations evaluated in chapter 2, along with appropriate parameters. Each technical analysis indicator transforms past price and volume data into a corresponding time series; then a trading strategy is used to transform this time series into a actual buy and sell signals.

3.4.1 Generating Atomic Rules

I generate random atomic trading rules are generated as follows:

- Pick an indicator, each with uniform probability. (The indicators are listed below, and discussed in detail in sections 2.6 through 2.12.2).
- If the indicator has a relevant parameter, randomly select a value for it, uniformly selected across the range given for each parameter.
- Pick a strategy, each with uniform probability. (The strategies are listed below, and discussed in detail in section 2.5.2)
- If using the selective or change strategy, set the initial selective percentage (the percent of stocks that generate a signal based on the technical analysis indicator) at 5%.
- If using the crossover, change, or selective strategy, pick a holding period (the number of days to hold a stock long or short after a signal is first issued) uniformly chosen over the range $[1, 20]$.
- Pick a direction (high values are bullish vs. high values are bearish), with equal probability.

I allow all of the technical indicators examined in the previous chapter. To review:

- Exponential moving averages (section 2.6). Relevant parameter: length of moving average window, range $[1, 200]$

- Simple moving averages (section 2.6). Relevant parameter: length of moving average window, range $[1, 200]$
- Moving Average Convergence Divergence (section 2.7). No parameter.
- Relative Strength Indicator (section 2.8). Relevant parameter: length of window, range $[1, 50]$
- Stochastic %K (section 2.11). Relevant parameter: length of n-day low/n-day high window, range $[1, 50]$
- Commodity Channel Index (section 2.12). Relevant parameter: range $[1, 100]$.
- Percentage Volume Oscillator (section 2.12.2). No parameter.
- Accumulation/Distribution Line (section 2.10). No parameter.
- On Balance Volume (section 2.9). No parameter.

I set the parameter ranges to reasonable values, according to the following guidelines. For parameters whose traditional values are less than 30, I set the range to be $[0, 50]$. For parameters whose traditional values are greater than 30 but less than 100, I set the range at $[0, 100]$. And for parameters whose traditional values are greater than 100, I set the range at $[0, 200]$. Note that these initial values do not constrain the parameter values during search.

To review the possible strategies:

- Comprehensive: Go {long,short} every stock whose indicator is {above,below} its centerline.
- Crossover: For each {upcrossing,downcrossing} of the technical analysis indicator centerline, go {long,short} for n days.
- Selective: For each stock a technical analysis indicator in the {top,bottom} p% of all stocks on a given day, go {long,short} for n days.
- Change: For each stock whose overnight change in the the technical analysis indicator is in the {top,bottom} p% of all stocks on a given day, go {long,short} for n days.

Note that these strategies are slightly modified from those used in chapter 2. In section 2.5.2, the selective and change strategies functioned on a daily horizon – stocks would be held long or short for only one day in reaction to a signal. In addition, the change and selective strategies have a parameter – the percent of stocks that generate signals – that was fixed at 5%; here, it is allowed to vary.

Note that the results of chapter 2 clearly demonstrate that most of the indicator/strategy combinations are not useful for building trading rules that produce positive Sharpe ratios. However, I explicitly do not want to take advantage of this knowledge. The point of this chapter is to see how those relationships can best be learned.

3.4.2 Perturbing Atomic Rules

Key to the process of learning effective trading rules is searching over the space of possible trading rules; essential to this process of search is the method used to perturb a candidate atomic trading rule. This section only handles the atomic trading rules; I discuss the composite trading rules below, in section 3.4.4. Given that nearly all the complexity in the perturbation procedure is in the atomic trading rules, I discuss the atomic trading rules first before describing the process to build more complex rules.

In traditional genetic algorithms, this perturbation is straightforward: flipping a single bit, or each bit with some random probability. In more structured representations, perturbation becomes more complicated. Given a lack strong theoretical direction, the algorithm below was chosen for *a priori* reasonableness with no post hoc optimization.

- 50% chance of perturbing the indicator:
  - 50% chance of picking a new indicator at random, and, if a relevant parameter exists, generating it from scratch by picking it randomly over the appropriate range.
  - 50% chance of altering the indicator's parameter, multiplying the current value by a random number uniformly distributed across the range $[.75, 1.25]$, then rounding up to the nearest integer.
- 50% chance of perturbing the strategy:

  - 25% chance of picking a new strategy at random, and generating appropriate parameters from scratch.
  - 25% chance of altering the holding period, adding a number uniformly distributed across the range $[-5, 5]$, with a minimum holding period of 2 days.
  - 25% chance of altering the selectivity, multiplying it by a factor uniformly distributed across the range $[.75, 1.25]$.
  - 25% chance of flipping the direction (changing from a high values bullish interpretation to high values bearish interpretation, or vice versa).

3.4.3 Composite Trading Rules

Learning atomic trading rules is important, but hopefully the next step is to take those atomic rules and combine them into composite rules. In order to build composite trading rules, it is essential to have operators that take the signals produced by two trading rules and produce one set of signals.

The obvious inspiration are the logical operations 'and' and 'or', but since I have three signals: long, short, and neutral, they have to be modified slightly. In addition, I introduce an operator that simply combines signals, as if the trading rules were individual securities to be combined in a portfolio. The exact details:

- And:
  - long and long $\Rightarrow$ long
  - short and short $\Rightarrow$ short
  - long and neutral $\Rightarrow$ neutral
  - short and neutral $\Rightarrow$ neutral
  - long and short $\Rightarrow$ neutral
- Or:
  - long and long $\Rightarrow$ long
  - short and short $\Rightarrow$ short
  - long and neutral $\Rightarrow$ long

- short and neutral ⇒ short
- long and short ⇒ neutral

- Portfolio: average the two signals together.

Of course, defining such connecting operators is only part of building composite trading rules. As the specific representations I use vary between experiments, I will leave those details to the relevant parts of sections 3.5.

3.4.4 Perturbing Composite Trading Rules

Section 3.4.2 discussed how to perturb the leaf nodes, or atomic trading rules. Now, we discuss how to perturb the composite trees. The procedure for perturbing a composite rule are as follows:

- Pick a node, uniformly chosen over all nodes, leaf and connector
- 50% chance of changing the structure of the tree
  - If the chosen node is the root node, expand the tree by:
    - Create a new connector node
    - Shift the current subtree down to be the left or right child subtree of the new connector node
    - Create a terminal node to fill the other child subtree
  - 50% chance of expanding the tree, as above
  - 50% chance of contracting the tree by:
    - Deleting the current node
    - Randomly picking either the left or right child subtree and promoting it to the current node
    - Deleting the other child subtree
- 50% chance of perturbing the node itself
  - If that node is a leaf node, perturb according to the directions in section 3.4.2.

– If that node is a connector node, pick a new connector, choosing a connector equal probability from the list of AND, OR, and the portfolio operator described above in 3.4.3.

Of course, if the depth of the tree is fixed, as in experiments below, expanding or contracting the tree does not apply

3.5 Learning Trading Rules

As discussed previously, there is an interesting strain of work in economics that uses genetic programming techniques to learn trading rules. Out of sample excess returns are considered evidence for predictability of the time series and thus an argument against the traditional conception of the efficient market hypothesis.

The machine learning methodology of these papers is straightforward: take simple technical analysis operations (moving averages, etc) and use them as the leaf nodes in a genetic programming [57] learner to induce arbitrarily complex technical analysis rules.

Although I will not argue against the details of the methods used in these papers, this section does argue that this general method – straightforward application of a genetic programming learner to induce complex trading rules – is not the appropriate approach to learn financial trading rules.

The core of this argument is as follows: In any machine learning endeavor, there is a trade-off between representational complexity and noise. Increased representational complexity brings with it the possibility of more accurately matching up with the true state of the world, but always at an increased risk of overfitting. And financial data is so noisy – much more noisy than the kinds of data machine learning techniques were developed on. As a result the customary trade-off is skewed; the balance needs to be focused on fighting noise.

Let me be clear: learning trading rules over financial data isn't a search problem, it's a noise problem.

Genetic algorithms are especially sensitive to this problem. Genetic algorithms were developed as optimization methods – with noiseless objective functions and no need to generalize to out of sample data – and so are especially vulnerable to problems of overfitting. Many machine learning techniques have assumptions built into their representational frameworks that help to enforce a priori reasonable representations; genetic algorithms have no such implicit safeguards (of course, the representation chosen always places implicit constraints; this point becomes crucial to the conclusions of this chapter). The customary genetic algorithm approach to this problem is to take part of the data set and use it as a validation set.

And yet, I don't want to abandon the idea of combining atomic trading rules based in technical analysis indicators into complex rules altogether. As the previous chapter demonstrated, technical analysis indicators do potentially have predictive power, and promise to be well-suited building blocks for building complex rules, somehow. The next section, 3.5, will deal with the right way to build complex trading rules.

The rest of this section provides empirical evidence to support the idea that genetic algorithms, or any other method that performs stochastic search over an a priori reasonable complex rule space made up of combinations of atomic trading rules, is counterproductive.

The next section outlines the genetic programming inspired structure for complex rules used to explore the properties of complex rule spaces. The following section, 3.5.1 provides results from an explicit genetic programming simulation.

3.5.1 The Basic Genetic Programming Approach

The basic genetic programming methodology I use is as follows:

- Generate an initial population of rules, and evaluate their fitness over the training set. The population size is set to 10.
- For n generations, repeat:
  - Generate a new population of rules by generating a copy of each rule in the population and mutating it.
  - Evaluate the fitness of these new rules over the training set, and the validation set.
  - Use pairwise selection to eliminate half of the population.
  - Repeat until the average performance on the validation set has failed to improve to 10 generations.

- Select the best performer of the final population based on the performance in the validation set.

Those familiar with the genetic algorithm/genetic programming literature will have many questions about the design choices made here. There is no crossover, and many of the other choices – pairwise selection ([67],[20]), population size ([82, 14]), for example – have been extensively discussed in the genetic search research literature.

This approach was designed to be as simple as possible; the focus of most genetic algorithm research is on improving search. However, for my purposes, the search capabilities of the basic approach are powerful enough, too powerful, probably. The focus of the next sections is on showing that for this data, the search process needs to be constrained to near ludicrous levels to help prevent overfitting – thus, I feel justified in not engaging in detailed tweaking to make search more efficient.

3.5.2 Representational Complexity

One way to fight overfitting is to limit the representational complexity; to constrain what kinds of functions the learner can represent. Although this technique is not a strong part of the research tradition of genetic techniques, it is certainly present in other areas of machine learning. Decision trees ([77, 78]) have an explicit 'pruning' step where the representation of the decision tree is trimmed by criteria derived from a validation set. The technique of optimal brain damage [60] changes the topology of neural networks as learning proceeds, removing does that do not directly contribute to the function being learned.

In the genetic programming learner, the easiest way to do this is to think about the *depth* of the tree; the deeper the tree, the more atomic trading rules, and the more complex their logical combinations.

The obvious place to think about limiting representational complexity is to limit the depth of the genetic programming tree; this section measures performance of the genetic learner, limiting the tree to varying depths

Figure 3.1 presents the results of running the learner as a function of depth of the tree. In addition to the standard variable depth tree (limited here to a maximum depth of three), I run the algorithm with several tree representations of fixed complexity. Trees of depth one consist of a single atomic trading rule; trees of depth two consist of up to two trading rules linked by a single logical connector; and trees of depth three consist of up to four trading rules linked by three logical connectors.

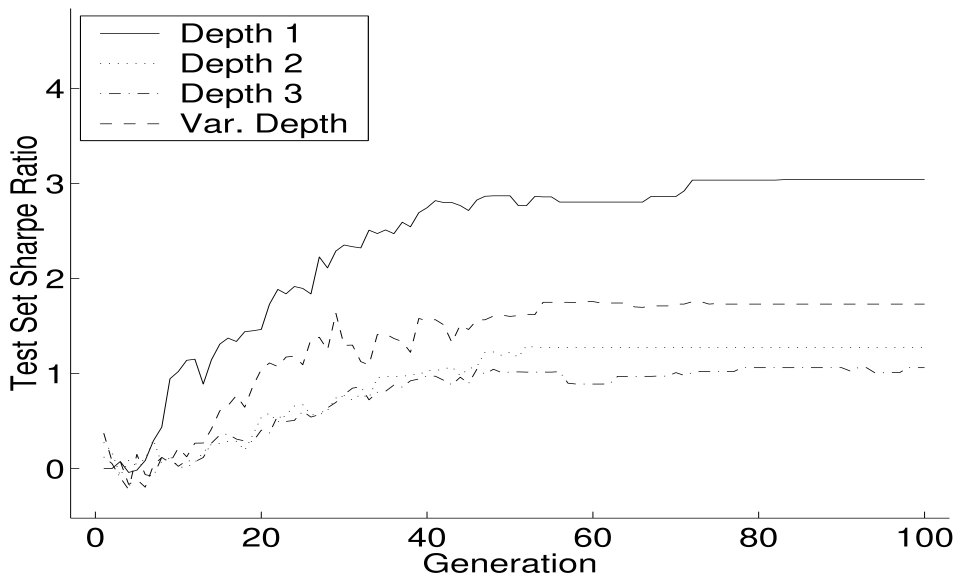

Figure 3.1: Mean Test Set Sharpe Ratio by Generation

Examining the graph, the trend is clear. The learner limited to the simplest possible representation is clearly superior, with a Sharpe ratio of approximately 3; the more complex trees perform worse, with Sharpe ratios of less than 1.5. The flexible tree in the middle of performance, In addition, for the most part, maximum performance is produced by generation 70.

Figure 3.2 confirms these results, presenting the mean final Sharpe ratios for each tree depth, with standard error bars.

The implications of these results are stunning: for best performance, limiting the representations to the simplest possible representations – individual atomic trading rules – is essential. Furthermore, given this simple representation, it is likely that the exact search method is inessential.

Both of these facts are unusual for machine learning: the limited complexity, and limited search required – less than 1000 function evaluations – are at the far end of machine learning practice. One could easily apply exhaustive search over the parameter space.

It is interesting to compare these results with those of the most successful techniques presented in chapter 2. Exact comparison is impossible – the time periods

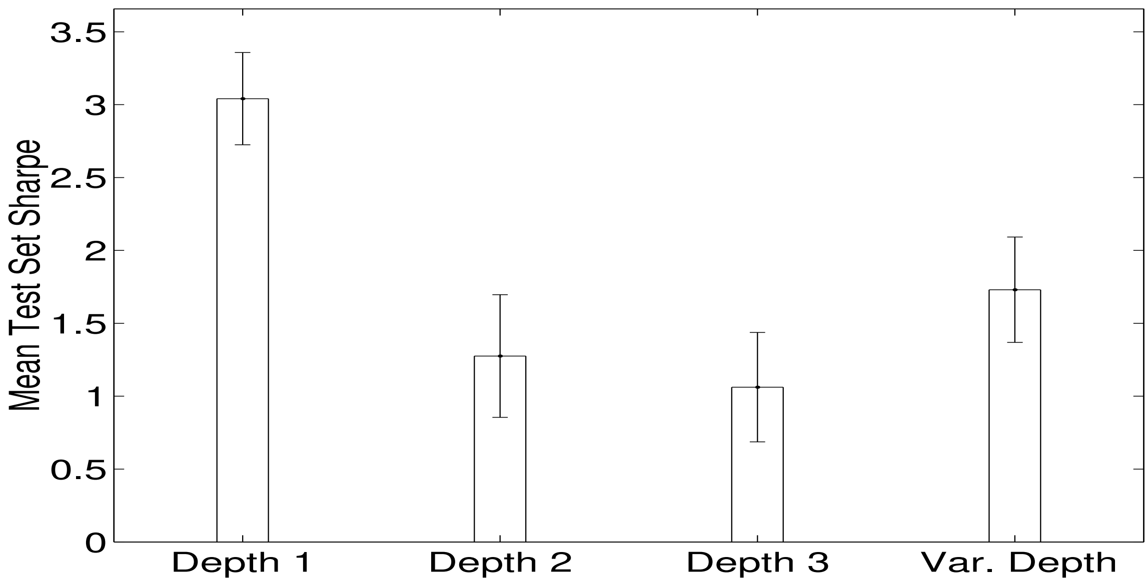

Figure 3.2: Average Test Set Sharpe Ratio by Tree Depth

differ (the previous technical analysis work covers a longer period). But comparing these results to the technical analysis results over the top two quintiles presented in section 2.15.1 shows that performance from the learner compares favorably – the best Sharpe ratios on the top two quintiles are all less than two, compared to the learner's stronger performance of a Sharpe ratio of approximately 4.

3.5.3 The Final Simple Learner

The above sections have provided empirical evidence for various design choices. Here, I present the resulting final algorithm. Due to the complexity of generating and perturbing atomic rules used by the learner, I refer back to the appropriate sections instead of duplicating the information here.

- Start with training set Tr, validation set V, and test set Te.
- Generate initial population of 10 atomic rules, according to the procedure detailed in section 3.4.1.
- For i=1 to 100
  - Copy and perturb each atomic rule according to the procedure detailed in

section 3.4.2, giving a new population of 20 rules – the 10 original rules, and 10 perturbed rules.

  - Evaluate the Sharpe ratio for each atomic rule over Tr and V.
  - Randomly pair up each of the twenty rules. For each pair, throw out the rule with worst Sharpe ratio on Tr.
  - If the mean Sharpe of the remaining population over V has not improved over its maximum during the last 10 generations, terminate the procedure.

- Evaluate the performance of each learner in the population on Te, and report the mean.

Key numbers here – 100 generations, a population of 10 atomic rules, where chosen to be a priori reasonable

3.6 Adding Ensemble Methods

I am left with the following approach: genetic algorithm style search over a universe of atomic rules, with a quick cutoff determined by validation set performance. So far, this approach works surprisingly well. Yet this leaves one unsatisfied. Does machine learning really have nothing to offer to improve this approach?

As it turns out something can help: ensemble methods. Ensemble methods are meta-techniques; they are applied on top of traditional machine learning algorithms to make them work better. The basic idea behind ensemble methods is simple: instead of using a single induced function for forecasting, ensemble methods use a committee of many induced functions. Instead of one induced function predicting, there are many induced functions whose predictions are combined together.

Although ensemble methods vary in details, they share the same basic approach. They take a machine learning method and apply the learning algorithm multiple times on the same problem under slightly tweaked conditions. This produces a set of similar, yet non-identical induced functions. The outputs of each member of this set of functions are then combined together to produce a single output. Instead of a single predictor, we have a committee of predictors. Of course, this description leaves out the exact method by which this set of predictors is generated; specific algorithms are discussed below, in section 3.6.1. But the intuition why this works spans variants of the basic ensemble method approach: having many different functions learned over slightly different conditions will smooth out the effects of noise in the data or learning process. Given the severely noisy nature of financial data, this is an especially promising prospect.

Ensemble methods have demonstrated convincing effectiveness when applied a broad spread of machine learning techniques, and have become a significant thread of machine learning research. Dietterich [28] provides an excellent introduction to ensemble methods. The best known ensemble methods are boosting, developed by Shapire (nicely summarized in [91]) and bagging, developed by Breiman ([21]).

In this section, I apply three ensemble methods on top of the basic approach of learning atomic rules, with hopes of improving performance.

3.6.1 How do Ensemble Methods Work?

To give a deeper understanding of how ensemble methods work, this section steps through one of the methods I use – bagging – in detail.

In the standard supervised learning problem, there is a set of training examples of the form $x_1, y_1, ..., x_n, y_n$. In classification, the y_n are class labels, and in regression, they the y_n are real valued.

The goal is for a learner to induce a function that maps an arbitrary x_i onto corresponding y_i. Ensemble methods are completely agnostic (ok, really, *almost* completely agnostic, but in ways that don't matter here) about how the learner functions.

The basic template for ensemble method is:

- Run learner multiple times, under slightly tweaked conditions
- Combine the outputs of each iteration of the learner to get the final output.

Each of these points raises questions about implementation. Combining the outputs is primarily a function of the problem type – regression outputs can be averaged; class labels can be voted upon, or averaged to produce a rough confidence estimate.

It's how the conditions are tweaked to generate slight variations among the learned functions, that really differentiate varying ensemble methods. The method used by bagging – the most common method among ensemble methods – is manipulating the training set.

Specifically, bagging generates perturbed training sets using bootstrap replication. Given a training set of n entries, bootstrap replication makes n draws with replacement over the original training set. This produces a new set of data, the same size as the original training set, with many duplicate data points.

- Start with a training set Tr containing n data points.
- For i iterations:
  - Generate a novel training set Tr'_i using bootstrap replication, by performing n draws with replacement over S.
  - Train the learner on Tr'_i,
- Average the predictions of the i learners to get final results

The number of iterations is customarily in the range of $[10 - 50]$. Customarily, it is chosen by a mix of trial and error, and consideration costs of computation time; in practice, the law of diminishing returns usually sets in after around 20 iterations.

Other Ensemble Methods?

There are many other potential ensemble methods. In addition to bagging, I apply two other ensemble methods:

- N-Fold Crossvalidated Committees: This differs from bagging only in how the perturbed training sets are generated. Given an original training set with m data points, it is is split into n disjoint sets of equal size, m/n data points each. Then, n training sets are generated by leaving out one of the disjoint sets, producing n new training sets, each with $m - (m/n)$ non-duplicated data points. I use $n = 20$.
- Simple Committees: Here, the learner is simply trained on the same training set multiple times, and then the results from the multiple learners are combined.

  Although in this case the training set is identical over each iteration, because the genetic learner is not deterministic, the learner will produce a different learned function each time it is run on the training set.

N-fold crossvalidated committees ([19]) is a technique developed by Parmanto and colleagues that has seen some serious interest and application to machine learning techniques.

The simple committee method does not exactly correspond to an existing method in the literature. In a sense, I include this method as a baseline for comparison against the bagging and n-fold crossvalidated committee methods. *If* the bagging and n-fold cross-validated committee methods prove superior to the simple voting method, then it is clear that the manipulation of the training sets is key to their success.

However, there is a strain of research in ensemble methods centering on perturbing the learner, not by manipulating the training set, but rather by introducing noise into the learner ([29, 79]). Given the randomness inherent in the genetic search method, the simple committee approach can be viewed as an example of one of these techniques.

Of course, the randomness in the genetic search is present while using bagging and n-fold crossvalidated committees as well, so it is certainly not the case that the simple committee approach is a strict test of the idea of randomness in the learner compared to other ensemble methods. In fact, the combination of randomness in search combined with the manipulation of training sets put these algorithms in the same category as those which combine non-deterministic neural network learning with the manipulation of training examples, as in the work of Raviv and Intrator [79].

There is one other ensemble method that I will discuss here, even though I do not apply it: boosting. I mention it because it is the best known of the ensemble methods – but, it is not easily applicable to the problem of learning trading rules, for reasons described in the next section.

Boosting, developed originally by Shapire [90] and extended by Freund and Shapire into AdaBoost ([38, 39]), is more sophisticated in its manipulation of the training set than bagging or n-fold crossvalidated committees. Instead of simply changing which points are included in the training set, boosting tweaks the importance of each datapoint in the training set in response to the algorithm's performance on that data point.

During each iteration, the learner minimizes the aggregate error over the training set, and the error the learner is computed on each data point. Then, each data point is assigned a weight by this error – data points with high errors have high weights; low errors mean low weights. Then, during the next iteration, the aggregate error function for the learner is computed over these weights – so the learner is is biased towards minimizing the error on exactly those data points for which it performed poorly during the last iteration.

3.6.2 Adapting Ensemble Methods to Learning Trading Rules

The trading rule problem faced in this thesis differs materially from the standard supervised learning problem that ensemble methods were designed to work over, making a seamless application of ensemble methods impossible.

In particular, there are two characteristics of the standard supervised learning that my trading rule learning problem does not share.

1. Boosting, bagging, and n-fold crossvalidated committees require that the training set is composed of discrete data points, which can be used to generate new training sets with omitted or duplicated data points.

2. Boosting requires that the success of the learner can be evaluated on each training example independent of the other training examples.

In contrast, my trading rule learning problem presents the following complications:

1. There is no clear-cut concept of what a single data point is. Is it the time series of a single stock? Is it a cross sectional single day's stock prices across all stocks? Is it a a single day's closing price for a single stock?

   Furthermore, the representations I use don't depend on the current day's data, they depend on moving averages and other summary statistics operating over potentially large time windows.

2. The success of the learner can only be evaluated when aggregated over the entire data set – over time, because of temporal dependencies caused by transaction costs, and over stocks, because of dependencies in portfolio construction.

Point two is important because it prevents the application of methods such as boosting, which crucially depends on a notion of how effective a learner is on a given data point – data points for which the learner fails to work well are emphasized in future learning, and data points for which the learner works well are de-emphasized.

Of course, such a notion could be approximated by looking at the daily return, perhaps a beta-adjusted return over the market return, but such an attempt is beyond the scope of this work.

Point one complicates the potential application of any ensemble method that depends on manipulating the training set. It raises two key issues: The first is how to define data points to produce perturbed training sets for bagging and n-fold crossvalidated committees. The two breakdowns that suggest themselves are by time and by stock.

Breaking down the training set by stock is straightforward. Simply leaving out, or duplicating, some of the stocks in the stock universe presents no complications.

Breaking the training set down by time is not as straightforward, and raises the second key issue: how to deal with the fact that the summary statistics used by the atomic trading rules in my representation are calculated over large time windows. Considering an individual data point to be a single day of data becomes problematic. Bagging, which requires the duplication of many data points, would destroy any notion of temporal order. The N-fold crossvalidated committee approach, which simply leaves out points, is conceivable, but this would still distort the relationship between the actual inputs to the learner and the desired outputs.

The answer is to think of an individual data point as a large moving temporal window of data. That preserves the temporal relationship between the inputs to the learner, while allowing for some of the data to be left out or duplicated in a fashion that makes temporal ordering irrelevant.

Practically, this is easy to implement. The technical analysis statistics needed by the atomic learner are calculated over the entire time series, properly ordered. But when evaluating the performance of the algorithm, the perturbed training set of days is used. Since the computation of the appropriate performance measure – Sharpe ratio – does not depend on a proper temporal order of the data, it can easily be computing according to the structure of the various ensemble algorithms.

One final implementation issue is how to combine the output of the multiple learners. In the typical machine learning problems, classification and regression, the process is relatively straightforward. For classification the group of learned functions can vote on the class label; for regression, the results can be averaged together. When learning trading rules, the outputs of the learners are portfolio weights; for each stock, on each day, a learner spits out what weight that stock should have in our portfolio, as a long or short position.

The sensible way to combine outputs is simply to average the portfolio weights together, and re-normalize the resulting portfolio weights.

3.6.3 Integrating Ensemble Methods with the Simple Learner

In section 3.5, empirically I showed that the most effective learner was simple search over a single atomic rule. In this section, I formally describe the application of ensemble methods layered on top of this simple learner.

- For $i = 1$ to 10
  - Generate the perturbed training and validation sets Tr_i and V_i according to the specific ensemble method chosen, described below.
  - Run the genetic learner detailed in section 3.5.3 on Tr_i and V_i to generate a population of atomic rules.
  - Out of the final population, select the atomic rule that produces the highest fitness over V_i.
  - Apply this atomic rule to the Te to generate portfolio P_i.
- Average portfolios $(P_1, P_2, ..., P_n)$ together, rebalance to ensure that aggregate long positions equal aggregate short positions, giving a final portfolio $P_{ensemble}$.
- Evaluate the performance of this $P_{ensemble}$ on the test set.

I have outlined several different ensemble methods. The key difference between them is how the perturbed training and validation sets are constructed. Here I formalize those differences.

Remember that each datapoint is a set of data for a given stock at a given day. Specifically, it is the time series histories of closing prices, daily highs, daily lows, and daily trading volumes for that stock – the entire history leading up to that day.

This gives a set of datapoints $D_{s,t}$, where $s \in S$ indexes by stock, and $t \in T$ indexes by time. Each ensemble algorithm that perturbs the training/validation sets alters either I or T – not both.

- Simple committees: Do not perturb the training/validation sets; take them as they are.
- Bagging by time: Construct T' by performing $|T|$ random draws with replacement over T.
- Bagging by stock: Construct S' by performing $|S|$ random draws with replacement over S.
- Crossvalidated committees by time: Split T into twenty random subsets $Tsub_1 ... Tsub_{20}$. Then, each T'_i is constructed by eliminating $Tsub_i$ from T, ie $T'i = T \cap \neg Tsub_i$
- Crossvalidated committees by stock: Split S into twenty random subsets $Ssub_1 ... Ssub_{20}$. Then, each S'_i is constructed by eliminating $Ssub_i$ from S, ie $S'i = S \cap \neg Ssub_i$

Given that there is no obvious a priori reason to prefer either the by stock or by time breakdown, their comparative effectiveness is investigated empirically below, in section 3.6.4 below.

3.6.4 Results

I test the effectiveness of the ensemble methods by applying them to my standard data sets. As in the experiments in section 3.5, I apply the new learner to the top half (by daily turnover) of my stock universe.

Specifically, I need to answer the following questions about the application of ensemble methods:

1. What works better, breaking up the data set by time or by stock?
2. Which ensemble method – bagging, n-fold crossvalidated committees, simple committees – works best?
3. Most importantly, do ensemble methods produce superior results to the plain learner?

The experimental results presented below answer these questions. The plot in figure 3.3 presents the final results for all the ensemble techniques. To answer the questions raised, in order:

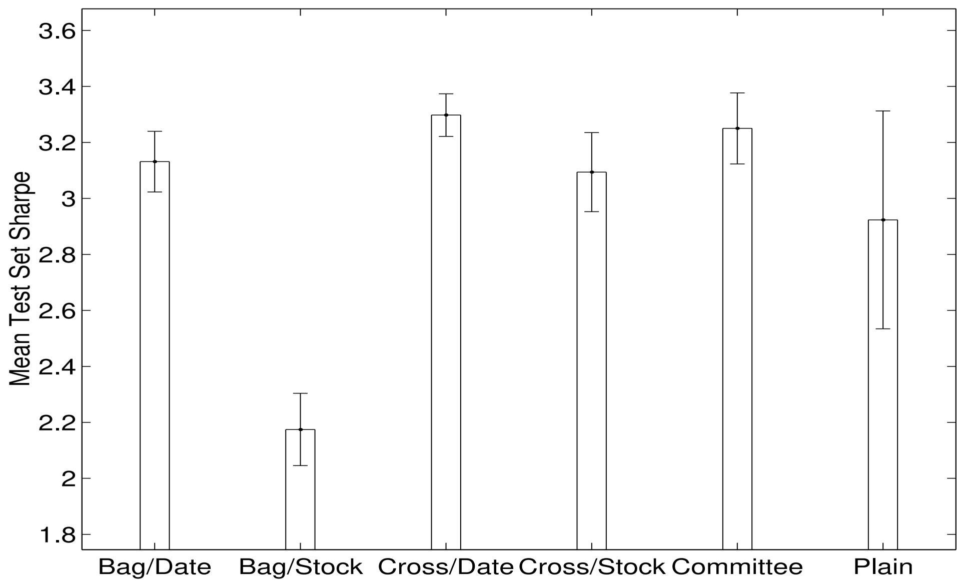

Figure 3.3: Final Performance of Ensemble Methods

1. There is a clear difference between breaking up the data set by stock and by date with the bagging technique – handling things by date is clearly superior. For the 10-fold cross-validated committees, handling by date is still superior to handling by stock, although the difference is not huge.

2. 10-fold crossvalidated committees produces better results than bagging, especially when handling by stock. Crossvalidated committees are slightly better than the simple committee method, although the difference is not significant.

3. With the exception of the bagging by stock approach, all of the ensemble methods look consistently better than the plain learner, both in terms of absolute Sharpe ratio and variability across iterations. However, the differences are not enormous, and the large standard error in the plain learner prevents statistical significance.

The only iron-clad conclusion is that handling by stock is inferior to handling by date. Still, it seems that ensemble methods – at least when handling by stock – are superior to the plain learner, and I will proceed accordingly.

To gain an intuition about the performance as a function of each perturbed dataset, figure 3.4 presents the results of three ensemble methods – bagging by stock, 10-fold crossvalidated committees by stock, and the simple committee system – plotted against the iteration of the algorithm. To be clear what I mean by 'iteration of the algorithm': the ensemble methods generate 10 new training sets. As I train a new learner on each succeeding training set, I incrementally add its prediction and evaluate.

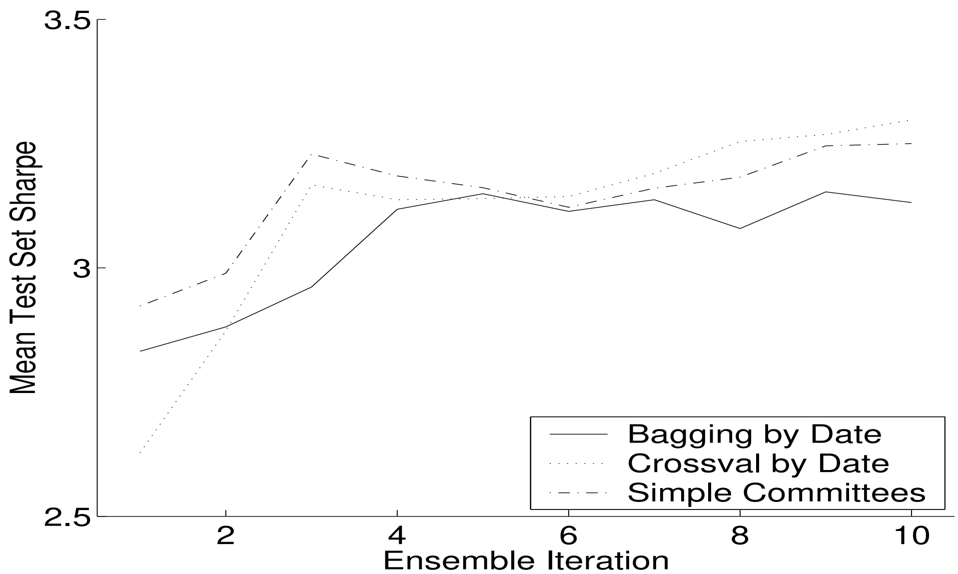

Figure 3.4: Performance of Ensemble Methods, by Iteration

The additional information here is not too surprising; all methods improve as iterations are added. Note that the simple committees approach at iteration 1 is identical to the results for the simple learner.

3.7 Performance on Holdout Data

In section 3.3.1, I explain why the noisy nature of financial data requires extra care. In this section, I confirm the conclusions reached in the above sections by measuring performance on a set of holdout data, comprising the last quarter of the full data set by time. In the algorithm design phase, I used the first half of the data (March 1st, 1998 to February 25th, 2000) for training and validation, and the next quarter (February 26th, 2000 to February 25th, 2001) for testing. Here, I use the first three-quarters (March 1st, 1998 to February 25th, 2001) for training and validation, and the final quarter (February 26th, 2001 to February 28th, 2002) as the test set.

Here, I commit to a final algorithm – the simple learner described above in section, together with 10-fold crossvalidated committees, and evaluate its results on the holdout set.

3.7.1 Does the Final Algorithm Really Work?

The most important conclusion is whether or not the core approach developed above – a basic genetic rule learner that learns atomic rules over a representation based on technical analysis, augmented by 10-fold crossvalidated committees – works on the holdout data set.

Table 3.1 presents results over the holdout test data set for the final algorithm. For comparison, I also the performance of the Russell 3000, as well as a equally weighted portfolios of the stock universe. Examining the table, the performance of the learner is strong: a Sharpe ratio of 1.43, driven by high returns and high volatility.

By comparison, neither the equally-weighted portfolios nor the Russell 3000 manages reasonable returns, with negative Sharpe ratios.

Table 3.1: Results of Learner on Holdout Data

Strategy:	Sharpe	Returns	Std Dev	Decline
Learner	1.43	88.43%	58.20%	-23.61%
Universe Portfolio	-.24	-0.21%	22.05%	-30.83%
Russell 3000	-.70	-12.32%	23.39%	-26.73%

I performed statistical testing using bootstrap techniques. I ran 100 trials, letting the learner operate normally over the training set; I then scrambled the test set data according to the procedure described in section 2.4.2, and evaluated the performance of the learned algorithms on the scrambled test set. In both cases, the testing revealed p-values of less than .01.

Figure 3.5 presents the mean NAV plots for the sample runs of the learner as well as NAV plots of the Russell 3000 and the equally weighted portfolios. Performance shows steady, if uneven gains for the first half of the data before increasing in the second half. The Russell 3000 and the equally weighted portfolio, by comparison, mostly hover around even.

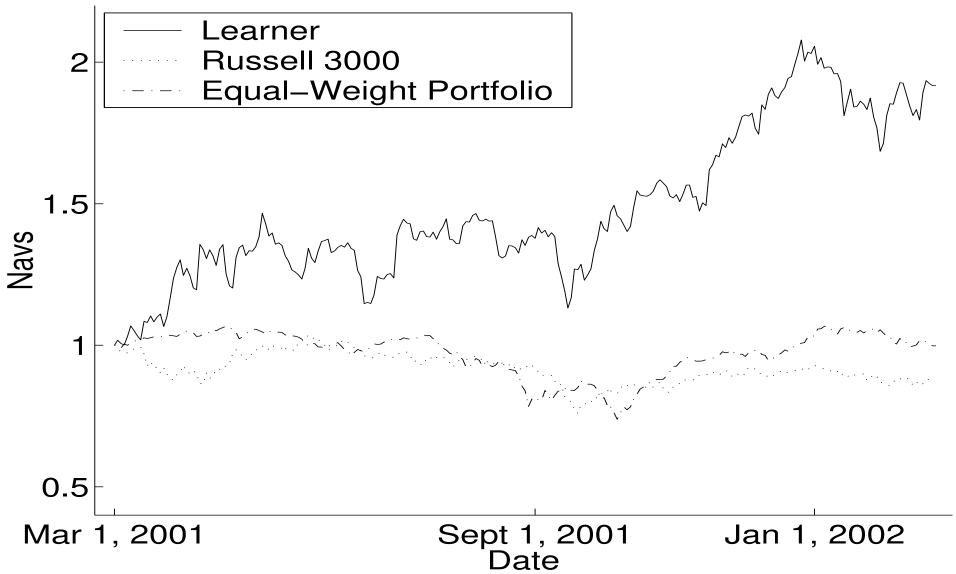

Figure 3.5: NAV Plots of Final Algorithm Results on Holdout Data

3.7.2 What is the Learner Learning?

One advantage of using the representation presented here is that it can be examined to understand what exactly the learner is learning.

Examining the results of the learner on the holdout set, we get the following breakdown. For strategies, 80.5% use moving averages, 15.5% use exponential moving averages, 3% use the on balance volume indicator, .5% (a single example) use moving average convergence/divergence, and .5% (a single eample) use the relative strength index. Overwhelmingly – 96% of the time, the strategies employed use moving averages of one form or another.

As far as the length of the moving average employed, the average period length is 7.16, very short. The strategy applied was always the selective strategy, in the reverse direction – effectively, the strategy looks for stocks that have the highest divergence above and below their short term moving averages, and then takes an opposite position – long positions for underperforming stocks, short positions for overperforming stocks.

Once a position is taken, the average holding period parameter was 2.29 days. *However*, the average time each stock was held was actually much greater: an average of 12.41 days for the final ensemble method learners.

This is because of two factors: first, the ensemble methods combine the results of multiple trading strategies into a single portfolio. Also, if a given stock continues to meet the criteria for a long or short position, it will continue to be held – so stocks that stay well below their moving averages for a few days will be held longer.

Similarly, the average percentage parameter (the percentage of stocks to take a position in) was low, at 2.39%; but the functional average percentage of the stocks held long and short at any given day was much higher, at 10.32% and 12.29% respectively.

This result is strikingly similar to that reached by Hellström [46] in his paper on predicting rank order of stocks. Hellstroöm, instead of trying to predict absolute returns of stocks, simply tried to predict stocks would have high and low returns relative to other stocks. He found that stocks that produced high relative returns one day tended to produce low relative returns the next day, and stocks that produced high relative returns one day tended to produce low relative returns the next. Hellström's work was performed on a completely different market. This suggests that the results here are unlikely to be an artifact of a single specific market.

3.8 Conclusions & Future Work

The conclusions of this chapter can be summarized as follows:

- The core conclusion of this chapter is that the simple algorithms developed here can learn trading rules that perform better than a buy and hold strategy over the Russell 3000 over a holdout set, tested with bootstrap hypothesis testing.
- Limiting representation to a single parameterized trading rule, and cutting search off quickly produce the best results.
- Given the limitations on representation and search, the exact search method is likely irrelevant; performance depends far more on representation and fighting overfitting.
- Ensemble methods show minor, although consistent improvement; however, it is difficult to find a strong preference for one ensemble method over another.
- The learner is inducing rules that depend on short term moving averages; stocks strongly above their short term moving averages are shorted, those strongly below their short term moving averages are held long. At any given time a little

over 10% of the stocks are being held long or short, and the average holding period for each stock is over 12 days.

3.8.1 Future Work

This chapter opens many doors of potential research. Given the basic framework of the methodology, simple trading rule learners using representations from technical analysis, augmented by ensemble methods, each area could benefit from additional research.

Boosting does not cleanly fit into the trading rule problem, since it is impossible to get an exact evaluation of the success of the learner on a single data point. However, this success could be approximated, and if the tradeoffs made in this approximation – ignoring transaction costs and dependence on other stocks for portfolio – end up not crucially impairing performance, then boosting could be a valuable method.

The validation methodology used above depended on splitting the available training data into two temporally contiguous sets, the first used for training, the second validation. However, this doesn't have to be the case. So long as I preserve the temporal relationships of the data for calculating the summary statistics used by the algorithm, there is nothing compelling me to use temporally contiguous daily return values for calculating performance.

It is interesting to reflect on exactly what is happening with the learner described here – basically, the learner is inducing multiple slight variations of the same strategy using short term moving averages.

Given this, I have my doubts about using machine learning as a stand-alone trading rule learner – rather, I suspect the more practical approach is to assume such a strategy (as Hellström does), and use the machine learning algorithms to essentially fine tune that strategy, using stochastic search over a reasonable parameter space and ensemble methods to induce a set of slightly varied functions which, hopefully, prove more robust to noise than any single parameterization.

Chapter 4

Message Boards

4.1 Introduction

This chapter investigates the first source of non-traditional financially relevant data, internet message boards. Specifically, I address data drawn from message board traffic off of the stock specific message boards of yahoo.com [11], focusing on market reactions to changes in message volume. The work in this chapter relies both on methodology drawn from economics, and also techniques and methodology investigated in chapters 2 and 3.

The organization of this chapter is as follows. First, a brief discussion of previous work, followed by a subjective introduction to the content of message boards. Then, I explore some descriptive statistics about message board posts – for example, temporal frequency and their relation to trading volume. The final section of the chapter examines using message board volume in place of trading volume in some of the technical analysis indicators explored in the previous chapter. Finally, section 4.8 summarizes and proposes future work.

4.2 Previous Work

Internet Message boards are a relatively new phenomenon, and as such their economic implications have not been deeply explored. Perhaps the first serious exploration in the literature is Wysocki's paper [99] investigating the effect of stock bulletin board posting volume on stock price behavior. As of yet, there hasn't been much other work.

4.3 Data Set

I collected the times stamps and subject line of every message posted on the yahoo.com message board server at `http://messages.yahoo.com/yahoo/Business___Finance/Investments/Sectors/index.html` [11], for every stock in the Russell 3000 as of July of 2001. Messages were collected from the inception of the system (early 1998) until August of 2001.

For market data, I downloaded split adjusted closing, high, and low prices, as well as and trading volume off of the yahoo.com quote server at `http://finance.yahoo.com` [9] for each stock. These prices are both split- and dividend-adjusted.

There are a number of other stock related message boards, including those on ragingbull.com [10] (now owned by Lycos) and The Motley Fool [6], as well as broader topic forums (often organized by sector rather than individual stocks) on sites like Silicon Investor [5].

But, like markets, activity breeds activity and success breeds success, and activity seems to be concentrating at Yahoo. For example, as of January 1, 2001, the Microsoft board at ragingbull.com has approximately 95,000 posts, the one at The Motley Fool had 84,000 posts, and the Yahoo! board had over 450,000. On January 1, 2003, corresponding numbers were 118,000 for ragingbull.com, for 110,000 for The Motley Fool, and 590,000 for Yahoo! (clearly, the pace of discussion has slackened with the sliding market). Thus I concentrate my efforts on the Yahoo! boards.

4.3.1 Special Data Handling

There are several features of message board traffic that require special handling.

First, while market data comes, by definition, attached to specific times during a trading day, message board postings occur throughout the day. Since I have consistently used closing prices so far, I will define the message board content of a trading day to be all the messages posted from market close on the previous day up to close on the current day. This allows us to talk about using, for example, Tuesday's message board volume to issue buy and sell signals for Tuesday's close.

In addition, message board traffic happens every day, but equities trading happens only on weekdays (and not every weekday at that). In order to handle this, we aggregate message traffic over the weekend for making conclusions about Monday's market activity (holidays are handled similarly; the non-trading day's message traffic is added to the previous period's message traffic). I acknowledge that this is not a perfect solution, but I feel it is the best alternative available.

Also, some days produce more messages than others. As demonstrated by the discussion below in section 4.5.1 and figure 4.2, weekdays see more message traffic than weekends. Since the interesting factor in message board traffic is *deviation from expected message counts*, I adjust each days' message traffic by a factor computed from the expected message traffic, so that the expected adjusted message traffic is the same for each day of the week. For weekends and holidays, we adjust the total messages for the period by the total expected messages for the period.

The pioneering work in this area has been done by Wysocki ([99]), and I use his approach as a starting point for the explorations here.

4.4 A Subjective Introduction to Message Boards

Message boards are an interesting phenomenon. In the way that the advent of Usenet and the broader internet is a democratizing force – *everyone's* voice can be heard – so too are the stock-related message boards. Discussion of investing has been taken out of the hands of a jealously guarded caste of professionals and handed over to the people.

Of course, anyone who's ever looked at free exchanges of ideas on the internet will understand why this isn't necessarily a good thing – open discussion is notorious for degenerating into irrelevant tangents and personal attacks. Stock message boards are no different, and they come with an additional danger – manipulation, which I will discuss briefly in section 4.4.2.

Nearly all posters on internet stock message boards claim not to be professional investors. The truth of those claims is essentially unverifiable. However, all of them are posting to the board because they have a strong interest in the stock – an actual long or short position, or sometimes just a purely psychological investment in the success or failure of a stock.

This leads to spirited discussion, as posters dig hard for any information relevant to a stock, analyze it, and argue over its meaning. The level of analysis is not always high, and often evolves into personal feuds that take on a life of their own. Some posters will push distorted interpretations or post outright lies in an attempt to manipulate the price. But, for almost any significant event, there is intelligent discussion, and every once in a while someone will post something that makes it clear that they have some sort of inside information.

This mix of flames, outright deception, and genuine information makes message boards intriguing. Almost certainly, they contain valuable information. And it's the job of this chapter to get it out.

Because the message board data is so messy from point of view of extracting textual meaning, I take the approach of using it as a source of numerical data. I do not attempt to classify or analyze the textual content; rather, I simply use the amount of discussion to generate a numerical signal counting how many messages were posted.

Clearly, going forward, diving into the textual content of the data is an interesting problem, but not one I handle here.

4.4.1 Why do People Post on Message Boards?

People post to message boards for a variety of reasons. After reading through thousands of messages, I believe the biggest categories are:

- Frank discussion of a stock's prospects: This is what an outside observer would expect to be the primary function of stock message boards. This debate can often get vicious. Proponents of long and short positions, as well as partisans of competing companies, get into vigorous arguments that often descend into name calling and flamewars.

- Analysis of current market events: When a stock price does something unusual – large price movement, large spike in trading volume – posters immediately start discussing why. If there is news afoot, they are surprisingly good at ferreting it out.

- Emotional support: Investing can be an emotional experience – think of the rise and fall of the internet bubble. Investors sometimes turn to others on the message boards to provide support. Other investors are often happy to provide a cheerleading section.

- Sense of community: Over time, people exhibit distinct personalities as posters, and regulars often get to know each other as people. Sometimes, discussion turns into out-and-out personal conversations, and some stock-specific message boards have spawned boards dedicated to community interaction instead of the stock itself.

4.4.2 Manipulation

The history of stock manipulation is long and storied. Message boards, with their broad audience, lack of control or supervision, and ability to reach thousands of investors in a matter of minutes, are perfect for manipulating illiquid stocks.

The case of Jonathan Lebed ([65],[12]) exemplifies this. Lebed, a fifteen year old high school student, parlayed $8,000 into over $800,000 through stock manipulation in months, until the SEC finally caught up with him and made him give back $285,000. His modus operandi was simple: he identified intriguing stocks, took large positions, and then talked them up on message boards, posting rumors of acquisition or simply of an explosive rise in the price of the stock. The stocks would rise, Lebed would sell, and move on to a new target. Lebed claimed he never posted any outright false information; often, simple repetition – posting predictions of a stock's rise under many different usernames – was enough.

It is likely that post-crash, this activity has greatly decreased. People are less likely to get caught up in this kind of speculative frenzy when the market as a whole isn't immersed in a speculative bubble.

Unfortunately, these kinds of manipulation are very difficult to detect automatically – difficult enough that this research does not attempt to do so.

- "Pump and Dump" a stock: Posting enthusiastic projections about the stocks future prospects, waiting for the stock to rise, and then selling. Sometimes, simple repetition is enough; other times, misleading rumors are posted. This is essentially what Lebed did.

To understand what is difficult about detecting this kind of manipulation, think of the following question: what is the difference between an enthusiastic and vocal supporter of a stock and outright manipulation? The distinction is not clear. In the SEC's eyes, the distinction is more one of intent than anything else, and is usually only determined ex post, by the trading actions of the manipulator. The occurrence of multiple posters touting the same stock could be genuine, and it could be a concerted manipulation attempt – there's no way to know, simply from examining the text itself.

As far as rumors go, they wouldn't be rumors unless they were nearly impossible for humans to disprove. The task is far harder for computers.

- Outright deception: Actually lying about some relevant event to force a stock up or down.

  It is very difficult to detect untruths without a comprehensive understanding of the world. To detect even an obvious lie about a relevant financial ratio would require a constantly updated model of every relevant financial statistic of every stock, together with an understanding of plausible events that could cause a sudden change in condition. A mistake in updating this knowledge base – which would likely occur far more often than any outright deception on message boards – could lead to a false positive.

In summary, what is really difficult about detecting manipulation is context. A piece of text on a message board is only considered manipulation in conjunction with trading activity.

Methods of computer text classification, as discussed in section in section 5.7, of the chapter on news stories, are good at looking at pieces of text in isolation.

They are not good at looking at chunks of text, understanding the relevant information, and putting it in proper context with regards to a whole world of background information – that is a much harder problem.

There are research programs that try to do this – Lenat's Cyc ([61],[62]), for example, certainly wants to walk this road – but the effort required to build them is measured in man-decades, and, while an interesting research project, are certainly beyond the scope of the research program of this thesis.

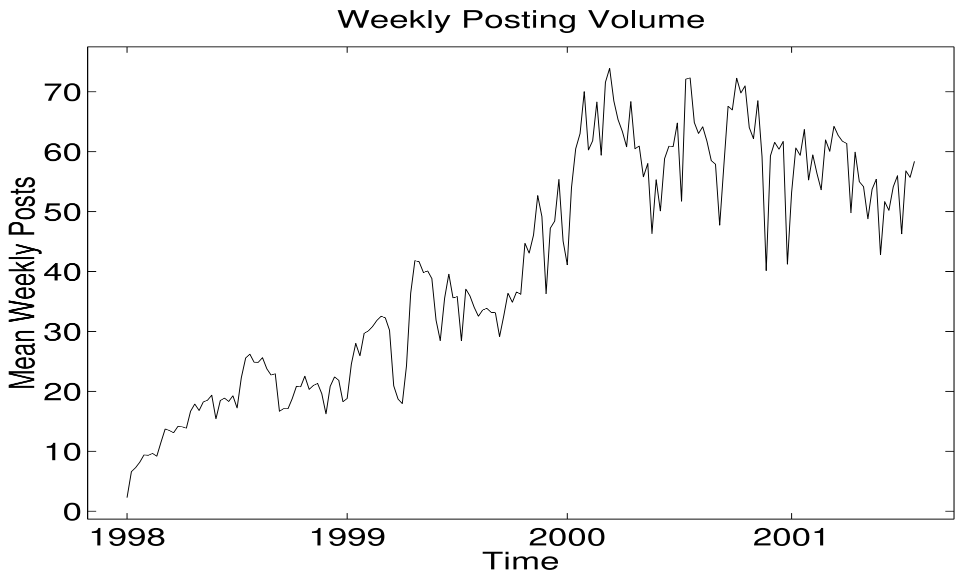

Figure 4.1: Average daily message counts for Yahoo message boards

4.5 Basic Message Traffic Measurements

This section presents some basic data about the message board traffic. First, we measure weekly message board traffic. Figure 4.1 plots the mean weekly message count for all stocks in the universe. Examining the graph, posting volume starts small in 1998 and grows relatively steadily until the beginning of 2000. Then, in early 2000 there is a noticeable drop, and weekly posting volume trends very gently downward with increasing volatility as time goes on.

4.5.1 Temporal Variation

Message posting varies considerably from day to day. The pattern is presented below in figure fig 4.2. The left graph plots shows the share of posts for each day, from market close to market close; the right graph shows the share of posts for open-to-closing hours on trading days. Unsurprisingly, weekdays generate far more posts than the weekends, with traffic increasing over the work week until reaching a maximum on Thursday and dropping slightly in Friday. The figures for Monday are interesting; given that the market close to market close measure for Monday includes Sunday evening, it is unsurprising that it would be substantially lower than the other trading

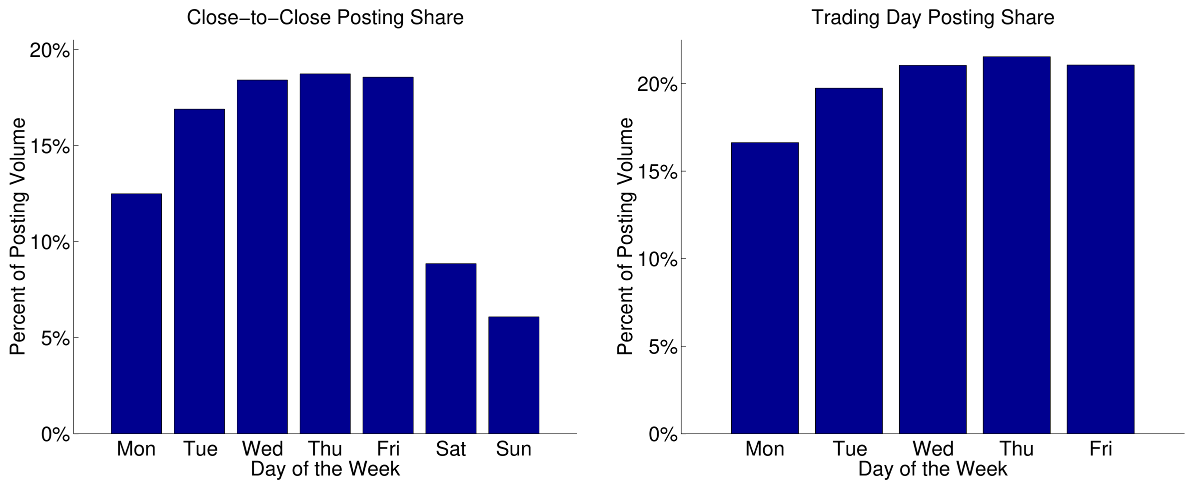

Figure 4.2: Message count share by day of week

days. However, message traffic during market hours on Mondays is still substantially lower than the other trading days, with no clear explanation.

Message traffic also varies by hour, unsurprisingly. Figure 4.3 presents a bar graph of the share of total messages broken down by hour. Note that the x-axis represents the hour of the day, in military time, with hour 0 representing midnight to 12:59 AM, hour 6 representing 6 AM to 6:59 AM, and hour 23 representing 11 PM to 11:59 PM. Market open (9:30 AM) happens during hour 9, and close (4 PM) is marked at hour 16. Unsurprisingly, message traffic is highest between market open and close, before trailing off gently after hours until about midnight. It drops to almost nothing in the dead of the night (3-4 AM), before starting to ramp up again around 8 AM.

Figure 4.4 breaks this data up into weekdays and weekends (weekends are defined as 12:00 AM Saturday morning to 11:59 PM Sunday night), which is an inexact proxy for trading and non-trading days. The bar graph on the left represents weekdays, the graph on the right weekends. The graph on the left is similar in general trend to the every day graph presented above, although the distinction between trading hours and non-trading hours is stronger. In the right graph, representing hourly message share on weekends, the message share stays essentially stable from noon to midnight, as users are presumably not tied to a traditional work schedule on weekends.

These results mimic the familiar U-shaped pattern observed in intraday trading volume ([45, 43], yet another sign of a strong relationship between trading volume and message volume.

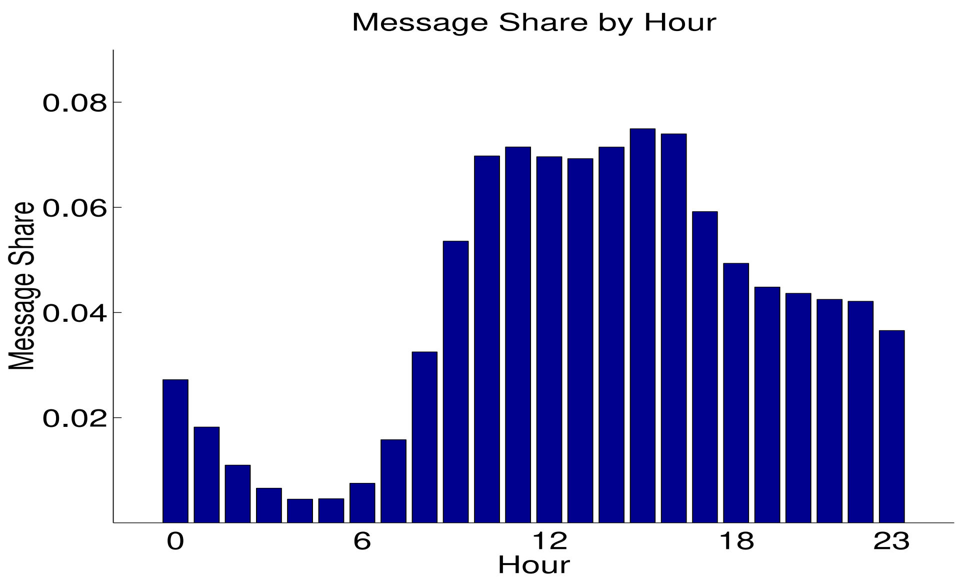

Figure 4.3: Yahoo! Message Board Share per Hour

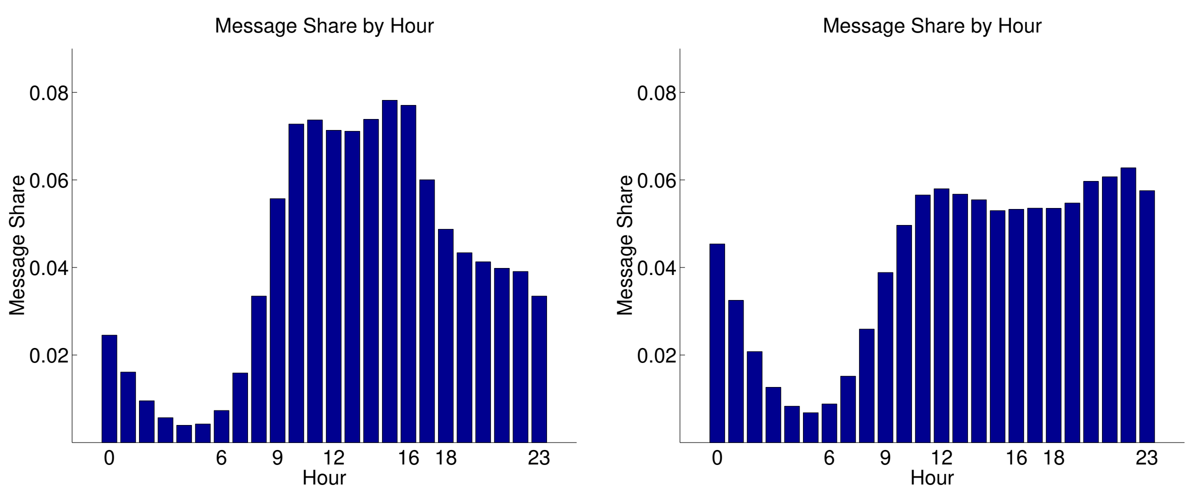

Figure 4.4: Yahoo! Message Board Share per Hour, for Weekdays and Weekends

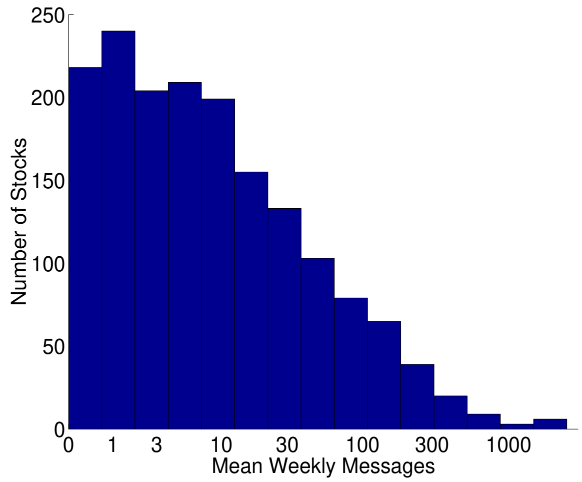

Figure 4.5: Histogram of Weekly Message Counts by Stock

4.5.2 Distribution Across Stocks

Message board traffic is heavily concentrated on a few stocks. Figure 4.5 gives a histogram (The x-axis is in log scale) plotted by the mean weekly message count for each stock in the universe. Intriguingly, the distribution appears flat out to roughly 10 messages per week, and then drops off dramatically as expected.

A breakdown by deciles appears in the table below. The first row, counts, presents the mean weekly count for the decile. The second row, weight, presents the percent of total posts about stocks in that decile. The third row, cumulative weight, presents the cumulative percentage of total posts by all stocks in that decile or lower. Examining the counts row, it's clear than 20% of stocks produce less than one message per week, 50% produce less than one per day, and 70% produce less than two per day. Clearly, getting meaningful signal out of message board posts for the vast majority of stocks would be difficult.

Decile:	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
Mean Counts:	.14	.83	1.66	2.96	5.05	8.34	13.93	25.49	56.63	295.43
Weight:	.04%	.2%	.4%	.7%	1.2%	2.0%	3.4%	6.2%	13.8%	72.0%
Cum. Weight:	.04%	.24%	.64%	1.4%	2.6%	4.6%	8.0%	14.2%	28.0%	100.0%

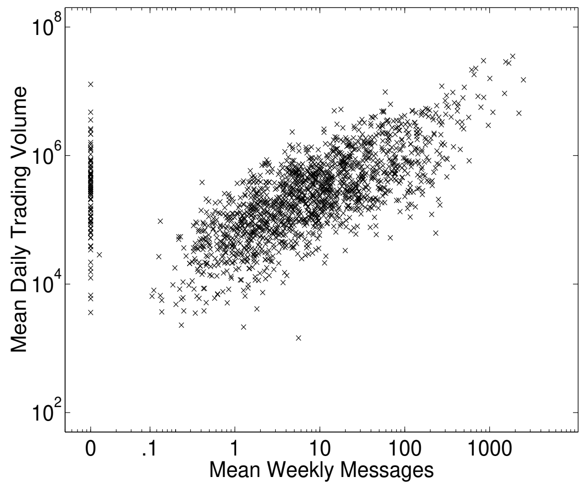

Figure 4.6: Scatter plot of correlations between mean trading volume and mean message counts

4.5.3 Correlation with Trading Volume

Trading volume and message volume are quite closely related; this section presents statistics examining that relationship.

The plot in figure 4.6 presents a scatter plot of mean trading volume against mean message counts (both axes are in logarithmic scale). This is a very well behaved scatter plot; except for the long line of stocks with zero message traffic on the far left, the cloud of points shows a strong positive correlation; the actual correlation coefficient is .7288.

In addition to being strongly correlated across stocks, daily message volume and trading volume have a strong relationship. The following table presents the correlations between trading volume, lagged trading volume, close-to-close message traffic, lagged close-to-close message traffic, overnight message traffic, and trading day only message traffic, averaged across all stocks.

Mean Correlations Across All Stocks

	Volume	Lag Volume	C-to-C	Lag C-to-C	Overnight	Trading Day
Trading Volume	1	.437	.287	.181	.207	.295
Lag Trading Volume	.437	1	.239	.287	.236	.175
Close-to-Close	.287	.239	1	.442	.809	.727
Lag Close-to-Close	.181	.287	.442	1	.422	.330
Overnight	.207	.236	.809	.422	1	.368
Trading Day	.295	.175	.727	.330	.368	1

Note also that the correlations between lag trading volume and trading volume, and between close-to-close message counts and lag close-to-close message counts, are similar, at .437 and .442 respectively. The high correlations between that both overnight message volume and trading day message volume and close-to-close message volume are not surprising, since the close-to-close message volume is by definition composed to the overnight message volume and the trading day message volume.

The table below presents the mean correlations for the top decile of stocks by message volume. The correlations are qualitatively similar, although higher across the board.

Mean Correlations Across Top Decile of Stocks by Message Counts

	Volume	Lag Volume	C-to-C	Lag C-to-C	Overnight	Trading Day
Trading Volume	1	.594	.551	.390	.464	.567
Lag Trading Volume	.594	1	.466	.551	.466	.399
Close-to-Close	.551	.466	1	.731	.925	.891
Lag Close-to-Close	.390	.551	.731	1	.710	.645
Overnight	.464	.466	.925	.710	1	.689
Trading Day	.567	.399	.891	.645	.689	1

The plot presented in figure 4.7 take each decile of stocks, ranked by message traffic, and plots the mean correlations for the decile between trading volume and lagged trading volume, close-to-close message counts, and overnight message counts. Inspection shows that the correlations with trading volume increase with the amount of message traffic – both for lag trading volume, but more quickly for close-to-close message counts and overnight message counts.

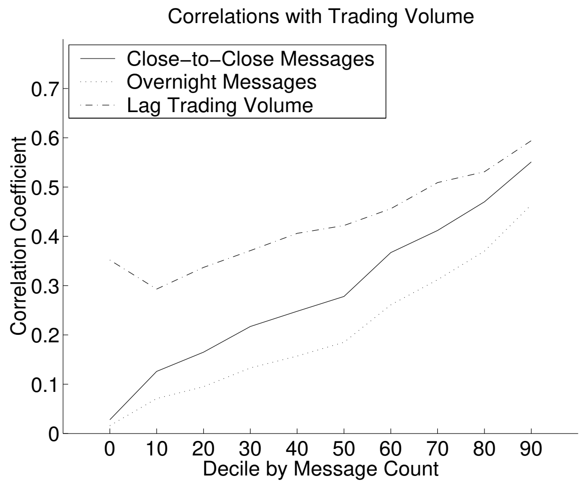

Figure 4.7: Correlations between trading volume and lag trading volume, close-to-close message counts, and overnight message counts

4.6 Economic Analysis

This section provides an economic analysis of message board counts; in substance, it is strongly similar to that provided by Wysocki [99], save for the much larger larger stock universe and time period.

4.6.1 Message Board Traffic and Trading Volume

This section explores the relationship between message volume and trading volume. Like Wysocki, I take the daily trading volume for each stock and regress it against lagged trading volume, the absolute value of the previous days abnormal returns, and message volume from both ragingbull.com and yahoo.com bulletin boards, split into posts made during the previous trading day, and posts made from the end of the previous trading day to the opening of trading on day t. The trading volume and message board post volume figures are all divided by the appropriate thirty day moving averages.

The results presented are for the top quintile of stocks, ranked by aggregate message traffic over the whole sample period. Note that unlike Wysocki, I do not include earnings reports as independent variables.

The following chart presents the regression results, with the dependent variable trading volume at day t (divided by the 30 day moving average).

	Mean Coefficient	Std error of Mean	t-statistic
Constant	.5471	.0046	$< .001$
|Abnormal Return(t-1)|	.7210	.0856	$< .001$
Trading Volume(t-1)	.3369	.0044	$< .001$
Yahoo.com trading hours(t-1)	-.0071	.0020	$< .001$
Yahoo.com non-trading hours(t-1)	.1225	.0052	$< .001$
Adjusted R^2	19.64%		

These results largely mirror those presented by Wysocki. For Yahoo posts, overnight postings have a significant effect on next day trading volume, while postings from the previous trading day actually have a negative effect.

Easily the most interesting coefficient is the non-trading hours message volume. In figure 4.8 below, I examine how the coefficient fares over a set of moving one year windows. The first period is the first year of data covering from January 1, 1998 to January 1, 1999; each succeeding data point shifts the data window by one month.

The graph plots the mean regression coefficient for the top 20% of stocks, together with standard error bars. Examining the graph it is clear that the predictive power of non-trading hour message volume is relatively stable, slowly growing from a regression coefficient of .1 in the first window to close to .2 in the year 2000. This demonstrates that the regression coefficients seen above are not simply due to an abnormal, short-lived spike in relevance.

I have explored the link between message volume and trading volume for those stocks that generate the most message board discussion; but, an open question is does the same relationship apply for stocks that generate less message volume.

Figure 4.9 presents a view of the regression coefficients for all stocks in our stock universe. The left hand plot is a scatter plot of the mean coefficient for non-trading hour message volume against the log of mean daily trading volume. The right graph plots the mean regression coefficient of the non-trading hours message volume by decile, together with standard error bars. Unsurprisingly, both graphs demonstrate a clear relationship between higher message volume for a stock, and the predictive power that message volume holds over that stock's trading volume.

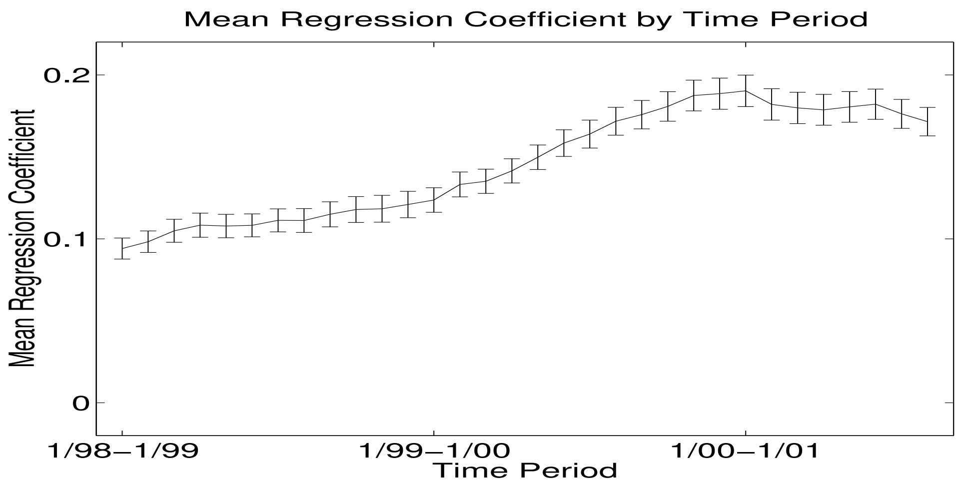

Figure 4.8: Message Volume Regression Coefficient over Moving Time Periods

The plot by stock decile is a near perfect exponential curve, indicating that the influence of message volume grows with the mean message volume. The scatter plot, by the very nature of scatter plots, is not as blindingly clear; however, it is clear that increasing message volume leads to higher regression coefficients.

4.6.2 Message Board Traffic and Trading Volume

This section examines the relation between message volume and the absolute value of abnormal returns. I repeat the same techniques as used in the previous section; the only change is that the dependent variable in the regressions is now the absolute value of abnormal returns at day t. I present the regression results below:

	Mean Coefficient	Std error of Mean	t-statistic
Constant	.0237	.0005	$< .001$
|Abnormal Return(t-1)|	.1370	.0016	$< .001$
Trading Volume(t-1)	.0016	.0002	$< .001$
Yahoo.com trading hours(t-1)	.0002	.0001	.0047
Yahoo.com non-trading hours(t-1)	.0002	.0001	.0085
Adjusted R^2	2.13%		

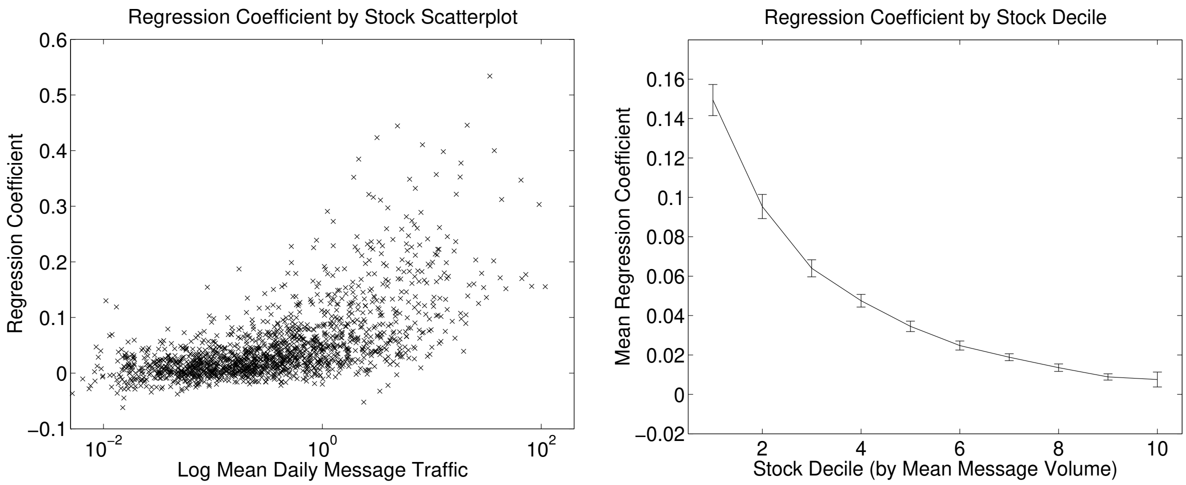

Figure 4.9: Message volume and trading volume, entire stock universe

Two differences from Wysocki's results stand out. First, the adjusted R^2 value is much lower; this is probably due to my omission of earnings report data as dummy variables. Secondly, although a rise in message volume, either during trading hours or non-trading hours, does indicate a rise in the absolute value of abnormal returns, the effect seems to be very small.

4.7 Technical Analysis

Given that message board traffic is strongly correlated with volume, and that some technical analysis indicators heavily dependent on volume produced strong results, It is an obvious question to ask if message traffic volume can be substituted for trading volume in the appropriate indicators. There are three appropriate indicators: on balance volume indicator (discussed in section 2.9), the accumulation/distribution line (section 2.10), and the percentage volume oscillator (section 2.12.2). The table below presents the results for the relevant indicator/strategy combinations.

The first conclusion is that none of the indicator/strategy combinations produce strong positive Sharpe ratios. However, for the ADL and OBV indicators, at least the results aren't completely atrocious – typically Sharpe ratios in the range $[-1, 0]$. The PVO indicator, which also happens to be the only indicator depending completely on volume, performs far worse, with Sharpe ratios worse than -3.

ADL Results (High Values Bearish interpretation)					
Strategy:	Sharpe	Returns	Std Dev	Decline	Long/Short
Comprehensive	-0.59	1.74%	5.53%	-10.30%	51%/42%
Crossover	0.53	13.07%	15.24%	-16.01%	4%/4%
Selective	-0.68	-9.88%	22.04%	-45.14%	5%/5%
OBV Results (High Values Bearish interpretation)					
Strategy:	Sharpe	Returns	Std Dev	Decline	Long/Short
Comprehensive	-0.15	4.07%	6.13%	-10.53%	67%/26%
Crossover	-0.77	-8.35%	17.38%	-40.26%	4%/4%
Selective	-0.17	1.28%	21.31%	-25.54%	5%/5%
PVO Results (High Values Bullish interpretation)					
Strategy:	Sharpe	Returns	Std Dev	Decline	Long/Short
Comprehensive	-2.91	-8.18%	4.52%	-26.86%	38%/55%
Crossover	-4.12	-14.04%	4.62%	-42.09%	28%/31%
Selective	-3.18	-34.43%	12.39%	-77.91%	5%/4%

Since, as noted in section 4.5.2, most stocks produce very little message traffic, examining only those stocks which do generate significant message traffic – say, the top 10% – is potentially valuable. The results for the above indicator/strategy combinations applied to the top 10% of stocks is presented below:

ADL Results (High Values Bearish interpretation)					
Strategy:	Sharpe	Returns	Std Dev	Decline	Long/Short
Comprehensive	-0.51	-3.43%	16.46%	-25.78%	63%/37%
Crossover	1.17	97.09%	78.81%	-35.78%	5%/4%
Selective	-0.85	-39.68%	52.55%	-84.48%	5%/5%
OBV Results (High Values Bearish interpretation)					
Strategy:	Sharpe	Returns	Std Dev	Decline	Long/Short
Comprehensive	0.24	10.38%	22.10%	-23.24%	77%/22%
Crossover	0.03	7.60%	76.68%	-58.06%	6%/5%
Selective	0.29	21.04%	55.72%	-55.99%	5%/5%
PVO Results (High Values Bullish interpretation)					
Strategy:	Sharpe	Returns	Std Dev	Decline	Long/Short
Comprehensive	-0.51	-2.72%	15.18%	-36.12%	48%/52%
Crossover	-0.31	-0.88%	18.68%	-42.44%	28%/29%
Selective	-0.59	-27.19%	54.67%	-82.06%	5%/5%

For comparison, the results for the above indicator/strategy combinations applied to the same top 10% of stocks, using trading volume in its appropriate place, are presented below:

ADL Results (High Values Bearish interpretation)					
Strategy:	Sharpe	Returns	Std Dev	Decline	Long/Short
Comprehensive	-0.65	-3.77%	13.48%	-25.34%	43%/57%
Crossover	0.88	47.91%	48.50%	-40.44%	4%/4%
Selective	-0.64	-14.96%	30.80%	-58.26%	5%/5%
OBV Results (High Values Bearish interpretation)					
Strategy:	Sharpe	Returns	Std Dev	Decline	Long/Short
Comprehensive	-0.04	4.11%	22.27%	-51.12%	52%/47%
Crossover	-0.06	2.35%	44.51%	-40.08%	5%/5%
Selective	0.16	11.20%	37.95%	-64.53%	5%/5%
PVO Results (High Values Bullish interpretation)					
Strategy:	Sharpe	Returns	Std Dev	Decline	Long/Short
Comprehensive	-0.07	4.17%	11.94%	-17.51%	44%/56%
Crossover	-0.94	-7.55%	13.37%	-30.00%	30%/32%
Selective	0.53	30.92%	49.36%	-45.81%	5%/5%

One conclusion is that the OBV and ADL indicators using message volume instead of trading volume produces worse results when applied to all stocks, but, when applied to the top decile of stocks by message volume, using message volume arguably performs better than trading volume. For the PVO, these conclusions do not hold; performance differences the PVO between trading volume and message volume are inconclusive, although using trading volume does provide a Sharpe ratio of .53 on the selective strategy, while using message volume produces no positive Sharpe ratios.

Another interesting conclusion is to note that the volume based technical analysis indicators perform poorly on the top decile of stocks by message volume, compared to the results presented in sections 2.10,2.9, and 2.12.2. Given the different time period of the data sets used in these two cases (The chapter 2 dataset runs from January 1, 1995 to August 31, 2001, whereas the results presented here use data starting with January 1, 1998), the results are not directly comparable.

The reason for this is unclear. Is this because there is something about that subset of stocks that creates problems, or is it just that the indicators/strategy combinations need to work on a huge set of stocks (1789 instead of 179) to be successful? The answer requires further exploration.

4.7.1 A Promissory Note

Of course, using message volume as a plug-in replacement for trading volume is only one potential method to exploit message board data. I explore another way – integrating the message board counts with the trading rule learner developed in chapter 3 – in the next chapter, in section 5.8, where I discuss it at the same time I discuss the use of news stories.

4.8 Conclusions & Future Work

This chapter provided a mostly descriptive overview of message board data drawn from a wide universe of stocks. The conclusions are straightforward. Much of the work here has been descriptive, trying to portray a feel for what message boards are about, including breakdowns by time and by stock of message board traffic. Also included are regression results that show that yahoo.com messages appear to have a strong connection to trading volume, and perhaps a relationship with the absolute value of returns that is not accounted for by trading volume. Finally, slotting message volume in place of trading volume in several volume-dependent technical analysis indicators show promise, although they do not produce strong results.

Extending this line of thought further, one might be able to think more deeply about the relationship between volume and message board traffic. Message board traffic is a kind of noisy measure of volume – given the message board posts are more likely reflecting current exchange activity.

However, I think what is most interesting is the portion of message board traffic that does not reflect trading volume – one obvious experiment would be to build a factor model of message traffic and examine the residuals once trading volume was accounted for for connections to market activity.

4.8.1 Future Work

The biggest potential are of future work is in expanding the kind of processing performed on message board posts. This chapter treats message board posts as a simple number; clearly they contain more information than this – discussions of real substance. For reasons discussed above, getting at this data is not easy, far harder than automating the classification of the content of news stories. But, undoubtedly, something can be done.

Another issue is how best to utilize the data – slotting the message volume in as a replacement for trading volume is not a strong effort to extract useful forecasting signal from message volume. I do address this challenge somewhat in the next chapter, in section 5.8, but beyond that there are many possibilities; I only scratch the surface in this thesis.

Given the relationship between trading volume and message board volume, one intriguing idea is to think about using message board traffic to help forecast trading volume. This could be used to help gain an idea of the future liquidity of a stock.

Chapter 5

News

5.1 Introduction

This chapter is the heart of this thesis. It introduces idea of using news data for financial forecasts, develops techniques for classifying and exploiting this data, and finally applies them, together with the methodologies developed in chapters 2, 3, and 4 to real financial data.

In section 5.2 I discuss relevant previous work, both in AI and economics. Then, in section 5.3, I summarize the basic approach I take to the challenge of integrating news with the numerical trading rule learner developed in chapter 3. Section 5.4 introduces the data set I will be using throughout the chapter, and section 5.5 provides some basic statistics about the distribution of news stories. Section 5.6 gets at the meat of this chapter – the ontology, the category breakdown by which I classify news. Section 5.7 discusses classification methods, the methodology I used, and its resulting accuracy. Finally, sections 5.8 and 5.9 get to the whole point of this exercise – studying the market impact of these news stories. The core of section 5.8 is an examination applying the categorized news data to augment the performance of the trading rule learner developed in chapter 3, and section 5.9 brings traditional event study methodology to bear. Section 5.10 wraps up the chapter and summarizes conclusions.

5.2 Previous Work

The previous work relevant to this chapter exists in an interesting state: while there is a huge amount of work on text processing in the AI/NLP literature, there is virtually no work on automatically processing news stories and integrating them with trading strategies.

Some work has been done. Wütrich and his students [98, 94] focused on forecasting major market indexes, using a keyword based system. Lavrenko et al [95] built a system that used naive Bayesian classifiers to link news stories to trends in intra-day trading for prediction. Previously, Fawcett and Provost adapted a fraud-detection developed in [36] to forecast price spikes based on news stories [35]. Other than that, little work that explicitly builds text processing into trading strategies has been done.

There has been an explosion in the field of statistical text classification techniques over the past twenty years. While the work in this chapter does not use these techniques – it uses a far simpler methodology inspired by – but it certainly borrows their methodology. Such methods are discussed below, in section 5.7, which discusses text classification methods in depth.

From the finance perspective, the idea of understanding the relationship between events in the world and market behavior has of course been an important concern. The key issue has been methodology – the notion of events in the world is not trivial to quantify.

The primary approach has been the event study, where an economist manually compiles a list of a certain class of events – mergers, for example – and then measures market reaction in the time surrounding the event. The price dynamics for each event are then aggregated over all events to give a view of the typical reaction. The event study literature is huge and varied; classic examples of event studies, applied to corporate takeovers, include the work of Jarrell and Poulsen [52] and Asquith and Mullins [18]. However, these studies are all limited by events that are trivially formalizable, and they require considerable human intervention.

The approach I take in this chapter has the potential to greatly expand the event study methodology – given the power to classify text, nearly arbitrary events could be defined. Of course, with automatic classification, classification accuracy becomes a problem, but in the worst case human intervention.

5.3 The Basic Approach

The basic issue this chapter addresses is straightforward. Nearly all of the work in using trading strategies as an operational test of the efficient market hypothesis – including the trading rule learner developed in chapter 3 – have depended on numerical data generated by the markets themselves.

However, it is clear that what really moves markets are events in the outside world, and news stories are clearly the best proxies for these events available. The challenge this chapter faces is simple: extract information from news stories that is integrable into the quantitative frameworks of both machine learning classification, and traditional economic analysis.

I take the following approach:

- Develop an ontology of financial news – a breakdown into categories designed by an human observer.
- Build classifiers that put novel news stories into these categories.
- Treat the counts of each category of classified news stories as numerical time series.
- Examine the market impact of these specific categories.
- Attempt to integrate these counts with the the trading rule learner developed in earlier chapters

In addition, I perform some event study like analysis on the individual categories in section 5.9, but the heart of the chapter is the integration of news data with the existing trading rule learner.

5.4 Data Set

For market data, I downloaded split adjusted closing, high, and low prices, as well as and trading volume off of the yahoo.com quote server at `http://finance.yahoo.com` [9] for each stock. These prices are both split- and dividend-adjusted.

For our dataset of news stories, I use business news headlines from the Yahoo! finance website [9]. Each stock has several pages of information, including price quotes, message boards, and news stories. For news stories, there is a page which lists recent headlines, time stamps, sources, and links to either on-site or off-site pages containing the full news story text. These news stories come from a variety of sources, including Business Wire [1], CBS Marketwatch [2], Reuters [4], Forbes.com [3], the Wall Street Journal Online ([8], and many others. The distribution of sources is discussed in more detail in section 5.5.3 below. Given the previous chapter's work on message boards, one could ask why I didn't also use the message board data. Unfortunately, the message board content is very messy – informal discussion that often degenerates into personal flamewars – and extremely difficult to extract meaning from. News headlines, on the other hand, are always concerned with relevant news, and are written in relatively standardized form.

I collected every headline attached to each stock in the Russell 3000 (as of July 2001) between the dates of March 1, 2001 and April 30, 2002. However, since the work in the following sections was performed at different points in the data collection process, different sections use different temporal slices of the news data. In addition, for work that involved stock prices, I need to limit myself to stocks that have a continuous price series over the appropriate time period, and furthermore, since I am investigating the relevance of news data, I limit myself to those stocks that produced significant amounts of news. The exact datasets used by each sections are listed below:

- News traffic statistics (section 5.5): March 1, 2001 to February 28, 2002, full Russell 3000
- News classification (section 5.7): March 1, 2001 to February 28, 2002, full Russell 3000
- Integration with trading rule learner (section 5.8): March 1, 2001 to April 30, 2002, top 50% of stocks, by turnover, of the stocks in the Russell 3000 that have a complete price series for the time period.
- Event studies (section 5.9): March 1, 2001 to April 30, 2002, top 20% of stocks, by news volume, of the stocks in the Russell 3000 that have a complete price series for the time period.

There are a few potential objections to the use of this data set. First, one could object that using only news from Yahoo is an inadequate sampling of business news. I realize that the universe of financially relevant news is hardly limited to those stories which are linked to by Yahoo finance. However, given the diversity of original sources of news present at Yahoo, I feel it is an adequate sampling.

Secondly, using only headlines ignores much of the information found in full news stories. However, examining the data, it is clear that headlines do contain most of the key information. I certainly agree that full text news stories contain far more information than headlines; whether they contain significantly more relevant information is not clear.

Journalists are taught the "inverted pyramid" structure – the commonsense dictum of putting the most important information first: the most important sentence in a paragraph should be the first sentence; the most important paragraph in a story should be the top paragraph. Similarly, if one reads many news stories, it's easy to see that the most important piece of information in a news story gets put in the title. If the core news event in a news story is a merger, then that will be clear from the title. For this reason, I believe headlines are an adequate starting point.

The time period of data used (fourteen months) is far shorter than the periods used for work in previous chapters. Clearly, this weakens the conclusions of the chapter; however, I feel that given the richness of the dataset, a little over one year is enough to begin investigations.

Note also that by using news already attached to a specific stock by Yahoo, I avoid what is admittedly a large, thorny problem: how to automate the classification of news with specific stocks. This is a difficult problem, and one that I do not address here. However, since real world sources of news that are piped in to financial professionals are often labeled with relevant stock symbols.

5.5 Basic News Volume Measurements

This section provides basic data about the raw volume news stories; the analysis mirrors that of message board traffic, presented in section 4.5. For data, I use a stock universe of every stock in the Russell 3000 as of July 1, 2001, and I collected every headline of every news story posted for each stock on the corresponding yahoo.com news page from March 1, 2001 to April 30, 2002, giving a fourteen months' worth of news.

Section 5.5.1 presents information about the variability of news story occurrence over time, section 5.5.2 discusses the concentration of news story volume across stocks, and section 5.5.3 discusses the different sources of news.

5.5.1 Temporal Variation

News volume does not vary considerably from day to day, except for a large difference between weekdays and weekends. The pattern is presented below in figure 5.1. The left graph plots shows the share of news stories for each day, from market close to market close; the right graph shows the share of news stories for open to close solely on trading days. Observing the left graph, clearly and unsurprisingly weekdays generate far more news stories than the weekends. The low news share on Monday is misleading, since it includes Sunday night in its count; similarly, the fact that Saturday looks like it generates more news stories than Sunday is partially misleading because the figure for Saturday includes Friday afternoon. A fairer view of variation by week day is given in the right graph, which presents only news stories posted during trading hours. Here, it is clear that the volume of news posted is essentially identical across weekdays.

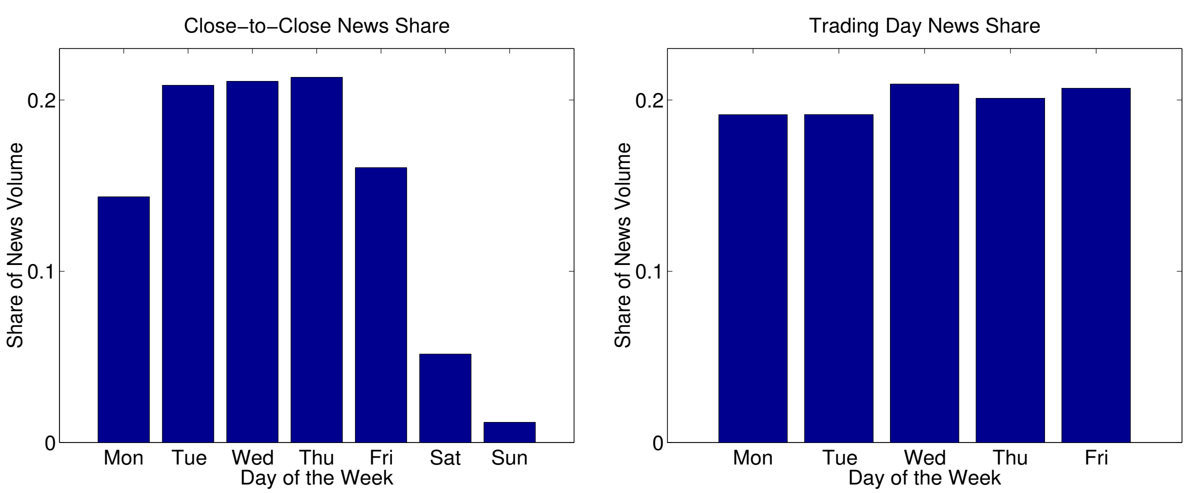

Figure 5.1: News volume share by day of week

News volume varies considerably by hour. Figure 5.2 presents a bar graph of the share of the number of news stories broken down by hour. Note that the x-axis represents the hour of the day, in military time, with hour 0 representing midnight to 12:59 AM, hour 6 representing 6 AM to 6:59 AM, and hour 23 representing 11 PM to 11:59 PM. Market open (9:30 AM) happens during hour 9, and close (4 PM) is marked at hour 16. There are two large spikes in news volume: the first in the 8 AM to 9 AM hour, just before market opening, and the second from 4 PM to 5 PM, as the market is closing.

The occurrence of news drops considerably in the evening, and almost completely to zero after midnight. This contrasts somewhat with the case of message board traffic, presented in figure 4.3 in section 4.5.1, where traffic stays strong into the evening. Of course, given that message board posts depend on individual posters who may be glued to their computers in the evening, the difference is not surprising.

Figure 5.3 breaks this data up into weekdays and weekends (weekends are defined as 12:00 AM Saturday morning to 11:59 PM Sunday night), which is an inexact proxy for trading and non-trading days. The bar graph on the left represents weekdays, the graph on the right weekends. The graph on the left is essentially identical to the everyday graph presented above. In the right graph representing hourly message share on weekends note that the spike at the 8 AM hour is still present, despite the fact that there is no concept of a market opening time on the weekends.

The results in Figures 5.3 and 5.2 mimic the familiar U-shaped pattern observed in intraday trading volume ([45, 43]). Unlike message boards – which are plausibly driven by trading volume – news probably does not have short term reactions of trading volume. However, it is clear that whoever controls the timing of news release, there is bias towards the opening and closing of the markets. Why this should be is unclear, and an interesting prospect for further investigation.

5.5.2 Distribution Across Stocks

Volume of news stories is heavily concentrated on a few stocks. Figure 5.4 gives a histogram (The x-axis is in log scale) plotted by the mean weekly news count for each stock in the universe. A few stocks have literally zero news stories, but the majority of stocks average less than two per week, with the distribution dropping off dramatically as expected.

A breakdown by deciles appears in the table below. The first row, mean counts,

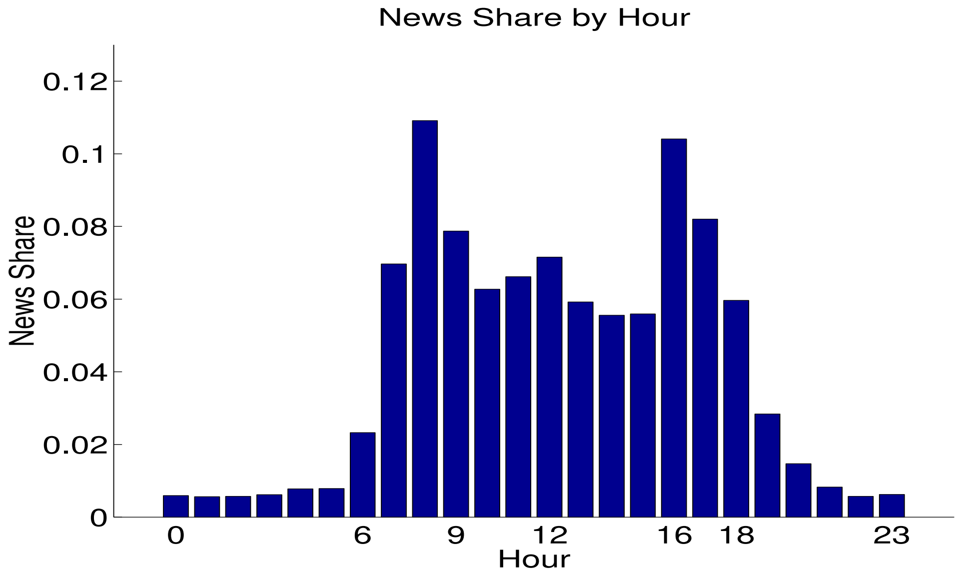

Figure 5.2: News Share per Hour

presents the mean weekly count for the decile. The second row, weight, presents the percent of total posts about stocks in that decile. The third row, cumulative weight, presents the cumulative percentage of total posts by all stocks in that decile or lower. Examining the counts row, it's clear than 30% of stocks produce less than one news story per week, and that 90% produce less than one news story per day.

Decile:	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
Mean Counts:	.50	.76	.93	1.08	1.28	1.52	1.84	2.40	3.85	15.2
Weight:	1.7%	2.6%	3.2%	3.7%	4.4%	5.2%	6.3%	8.2%	13.2%	51.5%
Cum. Weight:	1.7%	4.3%	7.5%	11.2%	15.6%	20.8%	27.1%	35.4%	48.6%	100.0%

5.5.3 News Sources

Yahoo! finance gathers news from a variety of different sources. Table 5.1 presents both a percentage breakdown and an absolute count of source frequency, for each source with a 1% share of the toal news. The data is presented organized both by all stocks in the universe, as well as by the top quintile by news volume.

A couple of interesting trends emerge from this table. PR Newswire and Business Wire, which are largely outlets for company press releases, together make up 30% of the news stories for all stocks, and 20% of the news stories for the top quintile.

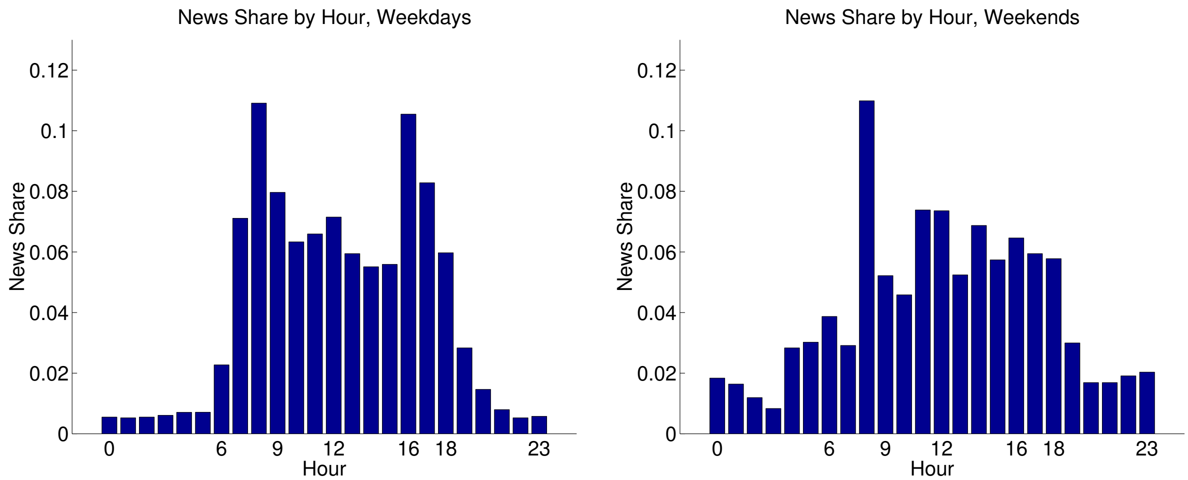

Figure 5.3: News Share per Hour, for Weekdays and Weekends

Although the differences in the breakdowns between the top quintile of stocks and the all stocks case is not dramatic, the stocks in the top quintile by news volume have a much lower percentage of news from sources such as CCBN, EDGAR, and First Call, which focus on reporting financial data. Each stock produces a similar amount of financial data – discussions of earnings reports, etc, while there is great variation between stocks in other kinds of news.

The news sources vary in what kinds of news they report. An exhaustive analysis is beyond the scope of the work here. However, to give a flavor for the variation in news sources, I present a brief description of some of the major sources below.

- Reuters: General business news
- PR Newswire: A distribution channel for company press releases
- Business Wire: A distribution channel for company press releases
- CBS Marketwatch: Market centered news (i.e., mostly news about the current performance of a stock)
- Thestreet.com: Market centered news
- CCBN: Earnings releases
- EDGAR: Online source for SEC document filings

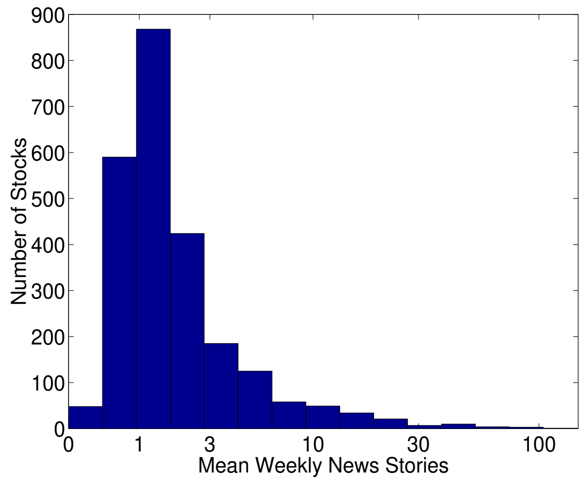

Figure 5.4: Histogram of Weekly News Story Counts by Stock

- Briefing.com: Market centered news
- Thomsonfn.com: Market centered news
- ON24: Market centered news, invariably an audio excerpt of an interview with an analyst.
- First Call: Cleaning house for earnings estimates and reports
- Internet Wire: A distribution channel for company press releases

5.6 Ontology

This section proposes a comprehensive ontology for the kinds of business news that exist. Business news – news relevant to specific stocks – comes in many flavors. Some kinds of news are going to be more important than others. Conventional wisdom would certainly suggest that news of a merger, or legal issues, or labor problems is likely to have a more material effect on stock prices than news about a charitable donation.

I have no rigorous derivation for the ontology; I have been guided by common sense, previous relevant work in economics (discussed above in section 5.2, and experience from thorough study of the kinds of news that shows up on the internet. The ontology has two levels: a layer of broad categories, with nested sub-categories.

Table 5.1: News Source Percentage Breakdown

Source	All Stocks	Top Quintile by News Volume
Reuters	21.66% (79947)	28.59% (54198)
PR Newswire	17.35% (64043)	11.02% (20890)
Business Wire	13.22% (48807)	7.82% (14825)
CBS Marketwatch	8.20% (30280)	12.00% (22751)
Thestreet.com	4.19% (15494)	6.48% (12297)
CCBN	3.79% (13992)	.86% (1633)
EDGAR	3.77% (13910)	.98% (1860)
Briefing.com	3.43% (12663)	1.79% (3397)
Thomsonfn.com	2.86% (10585)	.71% (1361)
ON24	2.85% (10505)	4.25% (8066)
Wall Street Journal	2.07% (7659)	3.20% (6080)
Forbes.com	2.02% (7473)	3.33% (6327)
Yahoo!	1.81% (6688)	2.71% (5151)
AP	1.73% (6396)	2.59% (4921)
First Call	1.47% (5456)	.34% (649)
Internet Wire	1.40% (5183)	1.43% (2715)

The complete ontology is below. For each sub-category, I give a brief description, and an example headline. Note that I will not attempt to build classifiers for all of these categories.

- Analyst Activity: This category contains news items about the actions and evaluations of stock analysts and credit rating agencies. It contains the following subcategories:
  - Coverage Initiated: Analyst initiates coverage of a publicly traded company. ("Coverage initiated on Sprint PCS by Dresdner Klnwrt Wasserstein", PCS, 3/5/2001)
  - Analyst Upgrade: Report of an analyst upgrading their rating of a publicly traded company. ("Marsh McLennan upgraded by Bear Stearns", MMC, 3/9/2001)
  - Analyst Downgrade: Report of an analyst downgrading their rating of

a publicly traded company. ("Motorola downgraded by Merrill Lynch", MOT, 3/9/2001)

  - Credit Upgrade: One of the major ratings agencies raises their credit rating of the company. ("Moody's Upgrades Senior Secured Bank Debt Ratings of Nextel Partners to B1", NXTP, 4/5/2001)
  - Credit Downgrade: One of the major ratings agencies lowers their credit rating of the company. ("Moody's lowers GMs credit outlook", GM, 4/7/2001")
  - Stock Analysis: General discussion of a stock's prospects by an analyst. This is a broad, catch-all category, intended to cover any discussion of a stock by an analyst not easily classifiable. ("Analyst: Lucent May Have Difficulty Monetizing New Superconductor", LU, 3/8/2001)
  - Sector Analysis: General discussion of broad sector's prospects by an analyst. ("Lehman Analyst Sees Some Entertainment Companies as Recession-Proof", PKS, 3/7/2001)
  - Conference Call: Mention of a company's scheduled quarterly analyst conference call discussing quarterly earnings. ("AppliedTheory Earnings Call scheduled for 10:00 am ET today", ATHY, 3/9/2001)

- Corporate Control: This category contains news about corporate control issues. It contains the following subcategories:

  - Merger: News about a merger involving a publicly traded company. ("James River Bankshares to Merge With First Virginia Banks", FVB, 3/5/2001)
  - Acquisition: News about an acquisition involving a publicly traded company, either another company or specific assets of another company. ('Convergys to Acquire U.K. Billing Software Maker", CVG, 3/6/2001)
  - IPO/Spinoff: News about a publicly traded company spinning off one of its divisions. ("Kraft Foods Files for $5 Billion IPO", MO, 3/16/2001)
  - Tracking Stock: News about a tracking stock, a stock designed to mirror the performance of a unit within the company, rather than the performance of the company as a whole. ("Cabelvision Announces Tracking Stock Distribution Date", CVC, 3/6/2001)

  - Minority Investment: News about a company taking, or selling, a minority stake in another company. ("Motorola Inc. Anchors Second Round Financing in WatchPoint Media, Inc.", MOT, 3/8/2001)
  - Shareholder Meeting: News about a shareholder meeting, or other shareholder issues. ("Maxim Pharmaceuticals Announces Results of 2001 Annual Stockholders Meeting", MAXM, 3/9/2001)
  - Divestment: News about a public company divesting assets. ("BorgWarner Divests Fuel Systems Unit", BWA, 4/20/2001)
  - Bankruptcy: News of a company filing for bankruptcy. ("Finova Files Chapter 11", FNV, 3/7/2001)

- Legal/Regulatory Issues: This category contains news about a company's interaction with legal and regulatory systems. It contains the following subcategories:

  - Lawsuit: News about a lawsuit (potential or actual) or arbitration involving the company or one of its key employees. Most of these news stories are announcements of class action suits. ("Metromedia Trial Set for March; Shareholder Activist Seeks Right to Inspect Company's Books and Records", MMG, 3/5/2001)
  - Regulatory action: News about regulatory action – at the state, national, or potentially surpra-national level – involving the company. This includes investigations by regulatory agencies. ("SEC Probing Bezos Stock Sales, New York Times Reports", AMZN, 3/9/2001)
  - SEC Filing 10-K: Report of a SEC form 10-K filing, a yearly earnings report. These mentions are almost invariably from Edgar, and follow a simple, universal format. ("EARTHLINK INC - Annual Report (SEC form 10-K)", ELNK, 3/9/2001)
  - SEC Filing 10-Q: Report of a SEC form 10-Q filing, a quarterly earnings report. As with 10-K reports, these are almost invariably from Edgar, and follow a simple, universal format straightforward. ("LITTON INDUSTRIES INC - Quarterly Report (SEC form 10-Q)", LIT, 3/8/2001)
  - SEC Filing 8-K: Report of a SEC form 8-K filing, the all purpose "special event" filing. As with 10-Q and 10-K reports, these mentions are almost

invariably from Edgar, and follow a simple, universal format. ("NEUBERGER BERMAN INC FILES (8-K) Disclosing Other Events", NEU, 3/9/2001)

- Earnings: This category contains news about company earnings. It contains the following subcategories:
  - Earnings Warning: A company issues an earnings warning, lowers earnings estimates, or warns about coming revenue shortfalls. ("ON The Move: New Focus Clearly Warns", NUFO, 3/6/2001)
  - Earnings Report: News about an earnings report ("Q4 2000 Guess Earnings Release - After Market Close", GES, 3/8/2001)
  - Earnings Outlook: News about estimated future earnings, revenues, and profits. ("Inforte Corp. Remains Comfortable with March 2001 Quarter Guidance", INFT, 3/8/2001)
  - Earnings Restatement: News about a restatement of past earnings by the company. ("Kroger to Restate Earnings by 'Minor Amounts"', KR, 3/5/2001)
  - Earnings Charge: News about an earnings charge – a non-recurring expense deducted from the current quarter's earnings. ("NCR Anticipates Charge Related to ATM Equipment Distributor.", NCR, 3/6/2001)
- Capital Markets Activity: This category contains news about the capital markets activity of a company. It contains the following subcategories:
  - New Financing: Discussion of issues of new debt, shares, other securities, or loans. ("Charter Communications files $4 billion shelf", CHTR, 3/9/2001)
  - Stock Buyback: News of a stock buyback. ("Waddell & Reed Financial Announces Share Repurchase", WDR, 3/5/2001)
  - Debt Repayment: News of debt repayment. ("NHI Further Reduces Indebtedness, Lowers Interest Rate", NHI, 3/7/2001)
  - Stock Split: News about a stock split. ("Frontier Airlines' 3-for-2 Stock Split to Take Effect on March 6, 2001", FRNT, 3/5/2001)

  - Options: News of a company's treatment of employee options. ("TIBCO Software Announces Stock Option Exchange Program for Employees", TIBX, 3/8/2001)

- Product/Sales Activity: This category covers stories related to a company's products. It contains the following subcategories.

  - Product Use: Description of product in use. ("State of Wisconsin Uses Keynote Systems to Develop and Launch New High Performance Government Web Site", KEYN, 3/14/2001)
  - Product Announcement: A catch-all category, covering the announcement of a planned or new product, or a modification to an existing product. ("Unifi Launches Fyberserv", UFI, 3/6/2001)
  - Product Sales: Announcement of a major product sale or contract, or discussion of company market share. ("Uzbekistan Chooses Two Boeing 767s for Fleet Expansion", BA, 3/16/2001)
  - Recall/defect: News of a product recall or defect. ("CPSC, McDonald's Announce Recall of 'Scooter Bug' Happy Meal Toys", MCD, 3/7/2001)
  - Product Research: Discussion of product research, primarily the granting of patents and clinical drug trials. ("AVANT Announces Changes to Its Phase IIb Trials of TP10 In Infants Undergoing Cardiac Surgery", AVAN, 3/6/2001)
  - Licensing: News about one company licensing another's technology or intellectual property. ("ON The Move: Immunogen Soars on Licensing Deal", IMGN, 3/6/2001)
  - Price Cuts: News about cutting prices. ("Intel, AMD Unveil More Price Cuts", AMD, 3/5/2001)
  - Price Increases: News about raising prices. ("Eastman Announces Price Increase for Oxo Alcohols", EMR, 3/5/2001)

- Stock Activity: This category contains stories regarding the short-term performance of the company's stock. The relevant time period under discussion is usually a single market day. It contains the following subcategories:

  - Stock Activity: News about the recent performance of the stock. ("Intel , AMD, TXN all down in early trading", INTC, 3/9/2001)

- Sector Activity: News about the recent performance of a sector relevant to the stock. ("Affymetrix, biotech stocks slide", AFFX, 3/7/2001)
- Market Activity: News about the recent performance of the entire market, or a significant index. ("Nasdaq up 3 days in a row", BRCM, 3/7/2001)
- Commodity Activity: News about the recent performance of commodities. ("Oil shares, crude prices inch up", KMG, 3/7/2001)

• Public Relations: This category contains news stories related to a company's explicit efforts to affect public opinion. It contains the following subcategories:

- Marketing: News about a public company's marketing or branding campaigns, including sponsorships of major events and conferences, and contests and sweepstakes. ("SunTrust Bank Launches New Branding Effort; Customer Service at Helm of 'Unexpected' Campaign", STI, 3/6/2001)
- CEO Comments: A story about the CEO publicly commenting on issues concerning the company. Occasionally comments are made by the CFO or COO, or other high ranking official instead. This does not include CEO presentations at conferences, which are counted under "Public Presentation". ("Ameritrade chair doesn't see M&A in top e-brokers", AMTD, 3/5/2001)
- Charity: News about charitable activities or sponsorships. ("BroadVision CEO And His Wife to Donate $15 Million to Fund Particle Astrophysics and Cosmology Institute", BVSN, 3/6/2001)
- Public presentation/speech: News about a public presentation by the company, or speech by a corporate officer. Sometimes these events are targeted at other industry participants, sometimes they are targeted at investors ("Internet Capital Group to Present At Merrill Lynch Internet Conference", ICGE, 3/8/2001)
- Award: News about awards won, records set, or special recognition of the company the company, or of employees of the company ("eXcelon e-Business Products Win Best of EnterpriseVision Award", EXLN, 3/5/2001)

• Corporate Relations: This is category involves news stories that center around a company's relations with other companies.

  - Business relationship: News concerning Joint ventures, partnerships, and other relevant agreements concerning the company. ("FutureLink Europe Announces Partnership With Ericsson", FTRL, 3/6/2001)
  - Supplier relationship: Story about one company purchasing another company's products. ("Indus International to Provide Transocean Sedco Forex With Robust Enterprise Asset Management Solution", IINT, 3/7/2001)
  - Industry Group: News about a company's involvement in industry organizations or trade groups. ("Medallion Homes Joins Growing Builder Consortium Developing a New Home Marketing Channel", MDC, 3/5/2001)
  - Standards: News about a company's involvement in developing industry-wide standards. ("StorageTek and CompTIA Join Forces to Promote Vendor Neutral Storage Standardization", STK, 3/7/2001)

- Internal Management Issues: Stories about the internal workings of a company and its business. It includes the following subcategories:

  - Belt Tightening: News about layoffs and restructuring, production cuts ("Copper Mountain cuts jobs, VIPs resign", CMTN, 3/7/2001)
  - Management Change: News concerning turnover in key management position, such as CEO or other corporate officer. ("Nolan's Tech Tidbits: Wal-Mart.com CEO May Be Leaving", WMT, 3/8/2001)
  - Reorganization: News of corporate reorganization. Sometimes this is the straightforward reorganization of business units; often it is a cover for job cuts ("AOL Time Warner Combines TV Networks Under Turner Broadcasting", AOL, 3/6/2001)
  - Executive Compensation: News about the compensation of key executives. ("Whirlpool CEO's bonus cut in half", WHR, 3/9/2001)
  - Labor Issues: News about strikes and other labor issues. ("Machinists union opposes United, US Airways merger", 3/7/2001)
  - Production: News stories about both production facilities, or natural resource reserves. ("New Well Holds Promise for Parker Drilling", 3/8/2001)

- Miscellaneous: This is a grab-bag category of story types that do not cleanly fit the other categories.

– Insider activity: News about a company insider buying or selling stock, including allegations of insider trading. (”BindView Re-Opens Insider Trading Window for Purchases Only”, BVEW, 3/8/2001)

– Force Majeure: Natural disasters, political upheavals, or other unanticipated events that affect company's business prospects. Examples include earthquakes, or allegations of past involvement in Nazi atrocities. (“Starbucks workers say quake no grounds for worry”, SBUX, 3/5/2001)

– Dividend: News relating to a stock's dividend. (“Wal-Mart Increases Dividend By 18 Percent”, WMT, 3/9/2001)

– Economy Analysis: Analysis of the broad economy, including Fed actions, bond market behavior, and macroeconomic data. (“Treasurys [sic] rally after France Telecom, Beige Book”, YHOO, 3/7/2001)

– Fund Activity: Discussion of actions made by mutual funds (“Gabelli Blue Chip Value Tastes Tech”, CPQ, 3/9/2001)

– Fund Analysis: Discussions of mutual funds of actions made by mutual funds (“The Cheapest No-Load Fund Families”, YHOO, 3/8/2001)

– Index adjustment: News about a stock being added to or eliminated from an index. (“Armor Holdings, Inc. Added to S&P 600 Smallcap 600 Index”, 3/6/2001)

– Exchange Activity: News about the actions of (“Krispy Kreme to Move to Big Board from Nasdaq”, KREM, 3/17/2001)

– Asset Sale: The sale of a significant asset, such as real estate. (“Deltic Timber Announces Land Sale”,DEL, 3/5/2001)

– Certification/Accreditation: The certification or accreditation of a product or service by a government or private group. (“McAfee Receives ICSA Anti-Virus Certification for WebShield SMTP Software”, NETA, 3/9/2001)

– Hackers/Virus: Problems revolving around hackers or viruses. This is often a virus warning from an anti-virus software manufacturer (“McAfee.com Issues High-Risk Virus Alert For ”Naked@MM”; Internet Worm Self-Replicates, Deletes Files”, MCAF, 3/7/2001)

5.6.1 This Ontology is Sadly Incomplete

The news ontology given above has well over 50 categories. Even so, casual perusal of business news demonstrates that this is not enough.

For example, take the category of "merger". There are many different kinds of news stories about mergers. Take the following examples, drawn from the time surrounding the Hewlett-Packard and Compaq merger announcement:

1. HP, Compaq Agree to Merge in $25-Billion Deal (HWP, 09/04/2001, 12:29 AM)
2. HP-Compaq merger boosts Asian techs (HWP, 09/04/2001, 3:17 AM)
3. The Merger: Great for Compaq Holders, a Mess for H-P (HWP, 09/04/2001 10:25 AM)
4. HP-Compaq merger May Be Tough to Pull Off (HWP, 09/04/2001, 5:09 PM)
5. Investor Revolt May Put H-P, Compaq Merger on Ice (HWP, 09/05/2001, 1:59 PM)

Headline 1 is the announcement of the merger itself; headline 2 is a report of effects the merger has on Asian markets, headlines 3 and 4 are analysis of the potential effects of the merger, and item 5 is a report of investor discomfort at the merger. No reasonable observer would expect these disparate events to have the same kind or degree of impact on Hewlett-Packard or Compaq's stock price. In fact, intuition suggests that headline 1 (the merger announcement) and headline 5 (discussion of trouble for the merger) might well have contradictory effects on the market. Capturing these nuances would require splitting the merger category into sub-categories such as merger announcement, merger analysis, and merger problems.

Many other categories evince this added complexity. In management change, for example, firing a CEO is different than hiring a CEO. People can be forced out or retire; employees can be brought in, or stolen away by other companies. Fleshing out the ontology could easily double or triple the number of categories, requiring subtle distinctions between related classifications. It is likely that such added sophistication would make the use of news more powerful, but for now, it must wait for further work.

However, some categories are straightforward and undifferentiable. Save for exceptional cases, news about an SEC filing means only one thing: someone has filed an SEC form.

5.7 Classification

Now that I have identified categories of news, the challenge becomes automating their classification – building algorithms that map chunks of text onto pre-defined categories.

This is a necessary step in my study of news and markets, for two reasons. First, it is only possible to hand-classify a relatively small number of news stories. The first two weeks provided nearly 10,000 headlines; clearly, hand classification cannot reasonably be extended to an entire year. In order to study the importance of various kinds of news over a sizable time period, automatic classification is a must.

In addition, if my work here is to serve as a model for decision support technologies for real traders, the classification process must be automated, and be able to run in real time.

This section introduces the concept of text classification. First, I discuss the classification problem itself. This is followed by a discussion of current research on methods of classification, with notes on some special treatment that the unique nature of this data requires. Then, I describe the metrics used to evaluate classification algorithms, and finally apply my hand-built classifiers to the real data and evaluate their success.

5.7.1 The Text Classification Problem

The canonical form of the text classification problem is simple. There is a set of documents – each of which could be anything from a single sentence to a full book – each with a corresponding category label.

The goal of text classification is to build a black box that takes arbitrary documents and maps them onto proper class labels. There are many approaches to this problem, described in the following section.

The implication of this technology for finance is potentially enormous. At its heart, text classification is a particular instance of the general classification problem, which maps data onto predefined class labels. Much of the attempts to apply AI to financial forecasting – including the work in chapter 3 – can be viewed classification applied to numerical data, trying to map past price patterns onto labels of "long" or "short".

But if there is information in news that is not materially represented in past prices, than the ability to extend classification's reach to text data, specifically financially relevant news stories, offers intriguing possibilities to augment traditional methods.

5.7.2 Methods of Text Classification

Text classification is an important problem in natural language processing, and much research has gone into it. There are several differing approaches to this problem, each with its own research tradition.

One useful way to think about the varying methods is to compare them to how humans understand language. It has long been the big dream of AI to build computer systems that genuinely duplicate human capabilities, and much AI research attempted to model itself on how human beings function. Unfortunately, attempts to reverse engineer the human brain – at any level – are truly still in their infancy, and so even AI's best attempts to duplicate human functioning are at best poor facsimiles. In the search for real world solution, AI researchers have realized that often the best functional algorithm bares little resemblance to the way the human mind functions.

In this subsection I present three major strains of relevant research, in order of decreasing resemblance to human behavior.

Full Text Understanding

Text understanding is a deep approach which attempts at least part of the "big dream" of AI – full understanding of the news, in the same the way actual living humans understand news. Perhaps the best known text understanding research project is the Cyc project, led by Douglass Lenat ([61, 62]). They involve sophisticated language parsing, and more importantly, a comprehensive knowledge model of the world.

These methods hold out the highest promise of text classification – if they succeed in a reasonable computer-represented understanding of the text, then simply assigning it a label is easy.

Unfortunately, the amount of effort required is immense – the Cyc project started in the mid 1980s and arguably still hasn't accomplished its goals. Developing one from scratch is a monumental effort, far far beyond the resources of a single researcher.

Information Extraction Methods

Information extraction methods are a significant step away from human functioning. Instead of building a general purpose mechanism, information extraction systems are built to extract specific information out of pre-defined domains. They parse sentences, and apply a set of predefined sentence templates to each sentence; if there is a match, slots in the template store the desired information.

Information extraction is not primarily a classification methodology; rather, its primary goal is to extract key pieces of information. However, it can easily be extended to classification purposes, as in work by Riloff and Lehnert [83].

This methodology promises high accuracy without the world-spanning effort required by a full text understanding approach. However, significant knowledge engineering is still required, as templates have to be built for each kind of sentence important to the category.

Bag of Words/Vector Space Models

While the two methods described above are promising techniques, in practice most text classification is handled by methods that have no concept of meaning, sentences, syntax, or even word order. These methods apply statistical classification techniques to representations of raw word counts, usually referred to as the "bag of words" or "vector space" representation. The statistical techniques used vary widely; both Yang [101] and Mladenic [69] present good surveys.

Statistical methods are the shallowest method described here; they make no attempt to exploit sentence structure or a model of human language; rather, they depend on statistical regularities across the distribution of individual words.

They don't even use full words – they use stemmed words, words stripped of derivational and inflectional suffixes, reduced to bare core of semantic meaning. Inflectional suffixes are things like plurals and verb tense markers; derivational suffixes differentiate root words into different parts of speech. For example, the verb “to incorporate”, the adjective “incorporated”, and the noun “incorporation” share a common stem, “incorp-”, and are considered the same word by these algorithms. So even the concept of part of speech is lost.

All of these algorithms follow the same basic framework:

- Stemming: Each word words are reduced to morphological and inflectional stems, stripping off inflectional suffixes

  The traditional stemming method is the heuristic-based Porter stemmer [75], although recently more sophisticated stemmers have grown in importance.

- Tokenization: Split the document into individual words. Sometimes bigrams (two word combinations) or trigrams (three word combinations) are used instead. Each document is then represented by a vector of word (or bigram/trigram) counts.

- Feature selection: Statistical algorithms identify rank the importance of stems by their usefulness in statistical classification, typically using a measure like information gain or mutual information (see [102] for a survey of methods). The less important stems can be discarded from the feature space.

- Statistical classification: A machine learning algorithm learns to categorize over the feature space. Common types of statistical classification include:

  - Decision trees: Decision trees learn map the input data onto classification labels via a tree of nodes, where each node is a split on a single attribute. Decision trees are little known outside of AI – perhaps because they do not share the buzzword-friendly qualities of neural networks – but they are highly effective. Decision trees were developed by Quinlan [77, 78].
  - Naive Bayes: Naive Bayes methods [64] simply calculates the probability Bayesian probabilities of each document class conditional on the individual features, assuming independence between feature probabilities.
  - K Nearest Neighbor: K-Nearest Neighbor methods [100] are conceptually simple: a novel document is classified by examining its K nearest neighbors in some distance space (defined by the vector of word counts), and assigning it the label shared by a plurality of those neighbors.

Within this basic framework, there are of course a large number other techniques for handling the classification, such as TFIDF [86] and support vector machines [54], as well as refinements such as Latent Semantic Indexing [89], which compresses the vector space in line with latent semantic content.

These methods have gained popularity because they are fast to build – they need only a set of labeled training examples – and they are completely domain independent. In addition, statistical methods generally work far better than one expects at first glance. Despite a complete avoidance of any concept of meaning above the level of the simple stem, the statistical properties of text seem to be strong enough for this approach to work.

My Method

Instead of using one of the above methods, I opted for simple hand built classifiers. For each category, I studied a series of examples, and hand crafted classifiers using logical combinations of regular expressions.

This approach was feasible largely because of the structure of the data set. First, since I was only dealing with single-sentence headlines, it was easy to design combinations of regular expressions that approximated noun/verb combinations.

Secondly, for most categories, business news is fairly formalized – writers are consistent in using the same vocabulary to describe similar events. In addition, journalists are taught the commonsense style – referred to as the "reverse pyramid structure" – of putting the most important information in a news story in the headline. If a news story is about a merger, the word "merger" appears in the headline. If the story is about an acquisition, then the words "acquisition" or "acquire" appear in the headline.

Of course, it is not quite this easy. Sometimes companies merge divisions instead of merging with other companies; sometimes they acquire technology instead of other companies. In addition to looking for key combinations of words that identify positive examples of a category, I have to look for key words that identify contexts that did not belong to the category. For example, when building the "Earnings Charge" category, when words "take" and "charge" appear, they almost always indicate a story about an earnings charge – *except* when they occur as part of the phrase "take charge of".

The hand built classifiers take the form of logical combinations of regular expressions; the exact regular expressions and composite rules used are presented in Appendix A

To build the classifiers, I had the following data sets. I fully hand classified the week of March 5, 2001 to March 12, 2001 to use for examples for classifier construction. I used the second week of hand classified examples to measure the accuracy of the classifiers, detailed in section 5.7.3 below.

In addition, I used the month of April during the classifier building process. When I had a rough version of a classifier, I would run it on the April data and examine the results for false positives.

Although the procedure I followed is not strictly formalizable, since it depended in large measure on my judgment, I followed the following procedure for each classifier:

- Examine examples of the category from the hand classified examples, for the week of March 5-12, 2001.
- Identify key combinations of words that indicate the presence of the category, usually:
  - Unique noun phrases (such as "class action lawsuit")
  - noun/verb pairs (such as "resign" and "CEO")
- For each key combination of words, construct a logical combination of regular expressions that detect those phrases. This serves as a rough draft of the classifier.
- Run the classifier on the headlines from April of 2001.
- Examine the false positives generated, and for each regular expressions, and identify key words that consistently occurred in incorrectly classified examples; these were then used to augment the regular expressions.
- Repeat this process until a satisfactory classifier is reached.

The end result of this process is a classifier consisting of a set of "phrase detectors", each consisting of a set of regular expressions linked by logical AND and NAND operators. By way of example, here is a part of the "new financing" classifier:

1. ”offer(s)?—issue(s)?—sale(s)?—pricing—launch(es)?—placement—raise(s)?—tap(s)?” AND ”shares”

2. ”public offering(s)?” NAND ”initial—IPO(s)?”

The expressions in quotes are regular expressions; the “AND” and “NAND” in all caps correspond to logical AND and NAND. So “regex1” NAND “regex2” would mean a positive match if regex1 matches and regex2 does not match. To those unfamiliar with regular expressions, they are a general purpose method for matching strings. The syntax may seem confusing at first, but is relatively simple. The “—” corresponds to the “OR” operator. The “?” means that the string in parentheses immediately preceding it is optional. So, the regular expression “placement—raise(s)?” matches if any of the strings “placement”, “raise”, or “raises” are present.

Each classifier is structured as a set of these sets of regular expressions, implicitly linked with “OR” operators; for example, if either of these two phrase detectors listed above registers positive, the headline gets classified as belonging to the “new financing” class.

The first phrase detector picks out sentences where the word “shares” co-occurs with a number of relevant nouns and verbs, that usually indicate discussion of new financing. In the second phrase detector, if the phrase “public offering” occurs and the words “initial” or “IPO” do not occur, likely the headline refers to new financing. When “public offering” occurs in conjunction with “initial” or “IPO” occurs, the headline is usually about an IPO, and I wish to exclude such headlines from the “new financing” category.

Of course, the full “new financing” classifier is more complex; the full set of classifiers is presented in the appendix.

I had two primary reasons for not using the statistical methods. First, many of the categories I wanted to classify had few positive examples in the set of hand-classified headlines – less than 10, in some cases. Also the amount of text per example is limited, since I am using only headlines. Statistical methods work best when they have a fair amount of text to work with.

Clearly, moving forward, one of the prime tasks ahead is to bring statistical methods into the mix. Now that I have good rough classifiers, it becomes much easier to generate training sets for each category. But for this initial work, I feel that these classifiers are adequate.

5.7.3 Measures of Success

The issue of evaluating the success of news classification is non-trivial. Simple measures of accuracy – how many right, how many wrong? – provide a shallow understanding of the properties of a classification algorithm, insufficient for tailoring the algorithm for specific needs.

For example, let's say our dataset consists of a set of stories, of which 50% are about management change. Take two algorithms: one that classifies every story as belonging to the management change category, the other that classifiers every story as *not* belonging to the management change category. Both algorithms will have identical accuracy – 50% – but their subjective behavior will be very different. In the first algorithm, we catch all of the management change stories, at the cost of seeing many false positives. In the second algorithm, I see no false positives, but miss all the real management change stories. These are extreme examples, but this exact trade-off is key to understanding performance in news classification.

The concepts that formalize these intuitions are *precision* and *recall*. Precision answers the question, "If I make a guess, how often am I right?" Recall answers the question "Of those stories I want to identify, how often do I successfully identify them?"

In addition, there is the F-measure, a single measure that balances precision and recall, with a parameter β that controls the relative importance importance of precision vs. recall. Setting $\beta = 1$ puts equally weight on both precision and recall, and is the value I use for all subsequent calculations.

Given the definitions in table 5.2, the precise definitions of precision, recall, and F-measure are given below:

Table 5.2: Definitions for Categorization

	Item actually is:	
Classified As:	Positive	Negative
Positive	True positive	False positive
Negative	False negative	True negative

- Precision: true positives / (true positives + false positives)

- Recall: true positives / (true positives + false negatives)
- $F - measure = ((\beta^2 + 1) \cdot precision \cdot recall)/(\beta^2 \cdot precision + recall)$

5.7.4 Measured Classifier Accuracy

I measure the accuracy of the hand-built classifiers over two different datasets. Remember that I labeled by hand every headline from every stock in the Russell 3000 for the first two full weeks in March of 2001, spanning from March 5th to March 18. The first week, March 5–March 11, 2001, was used in the design of the classifiers, leaving the second week, March 12-18, to fully test the accuracy of the classifiers with precision and recall.

In addition, I measure the precision of the classifiers over the entire month of May 2001. This is feasible (although time consuming) because calculating precision requires counting only true positives and false positives, which by definition only requires the examination of those headlines picked out by the classifier.

Calculating recall over a data set is not possible without labeling the entire data set, since false negatives are by definition positive class examples that the classifier doesn't pick up – I would have to look at every headline not picked out by the classifier, requiring the examination of the entire data set. The results are presented in 5.3. Precision is in general quite high, over 90% in all but 6 categories for the second week of March and all but 8 categories over all of May.

Recall is not as high on average, and far more inconsistent, especially for categories with small numbers of examples. For several of the categories with less than 25 examples, recall is under 30%, sinking to a low of 6.3% for charity.

5.7.5 Class Frequencies

Table 5.4 presents the frequencies of the classes in the ontology. As discussed in section 5.7.4, I hand-classified every headline from every stock for the two week period of March 5, 2001 to March 18, 2001. The first column presents the actual occurrence frequencies of each class.

The remaining columns are an attempt to approximate the frequencies over the entire sample period from March 1, 2001 to February 28, 2002. To do this, I first identify how many occurrences the hand-built classifiers classify over the entire year period. I then use the precision and recall numbers for each class to estimate the total occurrences over the year period, according to the following formula:

- estimated occurrences = (classified occurrences * precision) / recall

Where precision and recall are taken from the measurements presented in table 5.3. Of course, this estimation is only possible for categories for which I have written classifiers. Furthermore, it depends crucially on the measures of precision and recall, which for many classes depend on less than fifty class category examples, and so the estimated numbers should not be thought of as definitive. So in addition to presenting estimates used with measured recall values, I present more conservative estimates made with an assumed recall of 1, to generate a conservative lower bound.

5.8 Integrating News with the Trading Rule Learner

The classification of news headlines is an interesting problem in itself, but it is only a step towards the real goal: using that data to develop trading strategies.

Instead of trying to develop trading strategies based purely on news data, I feel it a more effective approach is to integrate the news data with the trading rule learner developed in chapter 3. Since I have already explored in detail the design of the trading rule learner, I will take its structure as given, and focus on how to combine it with the news data. In fact, I use the exact learner and dataset from chapter 3.

I should be clear about this: I am literally using the results I presented in section 3.7. I take the trading strategy it induced as a given, and I explore the performance of integrating the news data with that strategy over the same set of stocks and almost exactly the same time period used as the holdout set for the machine learning explorations.

5.8.1 Data Set

For most of the integration experiments below, I use the same data – in terms of stock universe and time period – as in chapter 3.

Time-wise, there is a slight difference: in chapter 3, the time period stretched from March 1, 1998 to February 28, 2002; since I have news data stretching from March 1, 2001 to April 30, 2002, I extend the time period for two months until April 30th. For the trading rule learner, I use the results – the learned trading rules – produced in section 3.7, where it was trained on the time period from March 1, 1998 to February 25th, 2001.

For the stock universe, I start with the Russell 3000 (as of July 31, 2001) and eliminate all stocks that do not have a full price series over the news data time period, leaving 2427 stocks. I further limit the stock universe to the top 50% of those stocks measured by the turnover – daily closing prices time trading volume – giving a final universe of 1214 stocks. This set of stocks is identical to that used in chapter 3

The Small Data Set and Concerns About Overfitting

I'd like to call attention to the amount of news data I have. Although I do have every news story for thousands of stocks for fourteen months, that is still only 294 trading days. This is not much, especially given the slippery nature of financial data, and conclusions in this section should be thought of as provisional.

As in chapter 3, I split the date into a training, test, and holdout set. I design basic approaches using the training and test sets, and then evaluate on the holdout set. The training set corresponds to March 1, 2001 to July 19, 2001; the test set July 20, 2001 to December 6, 2001; the holdout set December 7, 2001 to April 30, 2002.

5.8.2 A Note About Hypothesis Testing

Where appropriate, I test the conclusions in the following sections using bootstrap hypothesis testing. I discuss the use of bootstrap testing in section 2.4.2; the key fact in understanding how to apply bootstrap testing is defining the null hypothesis.

My null hypothesis is that using news stories as a gating signal to the trading strategy does not improve performance. To implement this in the bootstrap tests, I generate bootstrap datasets in the following manner: whatever the relevant signal is – total new counts, occurrence of individual news stories – I simply randomly scramble each stock's time series of that single over the appropriate time period. I then apply the appropriate gating signals to the scrambled signals, and evaluate the performance of the resulting portfolios.

5.8.3 First Idea: Earnings?

Let me start with earnings. It is well known that price movements can be severe around earnings announcements. Given that the rules induced by the trading rule learner make decisions based on price movements, it is reasonable to investigate whether large price movements around earnings announcements are materially different than large price movements that occur away from earnings announcements.

The application of this intuition to the trading rule learner is simple: when the text classifier detects an earnings-related news story for that stock, have the trading rule learner avoid taking a position in that stock. Specifically, I propose the following method of integrating the news classifier and the trading rule learner:

- If there is an earnings-related news story of a given class on that day, take no position in that stock for 15 days (three weeks of trading days).

I applied this rule to the news categories of earnings warning, earnings report, earnings outlook, earnings restatement, and earnings charge over the first third of the news data – the idea being, identify if categories work on the first third of the data, and then test results on the middle third, and verify on the final third.

The results are presented in table 5.5. The left column presents the Sharpe ratio of the trading rule learner augmented by the news classifier rule given above (note that the first row presents the bare trading rule learner). The right column presents the difference in Sharpe ratio between the bare trading rule learner and the trading rule learner with the news classifier, with a base Sharpe ratio of 1.74. These results should not be taken too seriously – the data in consideration comes from a time period of less than five months.

Examining the figure, the earnings report and the earnings outlook categories create the most improvement, at .38 and .46. Sharpe ratio differences respectively, other categories actually hurt the strategy the earnings warning category gives a negative difference of -.32, earnings restatement -.13, and earnings charge, -.23. Unfortunately, none of these differences are statistically significant by bootstrap hypothesis testing, but the result is still promising.

5.8.4 Expanding to All Categories

Given that it at least some categories can improve the results of the trading rule learner when used as a gating signal, the obvious extension is to ask if other categories can produce differences as well. The logical approach is to examine all the categories for which I have classifiers, examine Sharpe ratio differences they produce when used as a gating signal, and then use those categories as a gating signal on out of sample data.

Of course, the differences above were produced on the training set – the first third of the news data. The real test is to see if performance on the training set generalizes to the test set. To that end, I propose the following approach:

- Identify those news stories that produce a positive difference in Sharpe ratio when used as a gating signal on the training set
- Aggregate those news stories in the test set, count the total number of occurrences of news stories belonging to those categories.
- Use the presence of those news stories as a gating signal for the trading rule learner

Since I am aggregating multiple types of news stories, and there are many stocks that have news stories every day, I cannot simply use "presence of a single news story" as a condition – if I did, such trading strategy would never take a position in some of the more talked about stocks, which is not the effect I am trying to explore. Instead, I define a concept of "more news than usual" by the following:

- Take a set S of news categories, and aggregate their daily counts.
- Take the mean and standard deviation of the aggregated news counts over a 30 day period.
- If the number of news stories on that day is greater than n standard deviations above the mean, take no position in that stock for m days.

This is meant to be a priori plausible integration. In order to implement this strategy, there are three key parameters. S, the set of news categories, n, the number of standard deviations above the mean required to trigger the gating signal, and m, the number of days to override the trading rule learner.

5.8.5 A Very Simple Learner

I have a principled way to set S, the set of news categories – simply identify the categories that produced a positive Sharpe ratio difference on the training set. However, I do not have a principled way to set n and m, where n is the number of standard deviations away from a running mean needed to trigger a gating signal, and m is the time period, in trading days, to hold out of the market. This section explores the idea of learning them.

As mentioned above, I have little data from a temporal perspective, too little to engage in the cycle of algorithm design practiced in chapter 3. Therefore, I will take the basic conclusions from chapter 3 and apply them to build a reasonable learner. Those basic conclusions are:

- Keep representation simple
- Cut search off quickly
- Apply ensemble methods

Given that I have far less data here than in the explorations in chapter 3, I am even more paranoid about overfitting. In addition, I do not have enough data to spare for a search validation cutoff – the training set is 96 days long, splitting that in half again gives 48 – too small for comfort

Therefore, I eliminate search almost entirely; instead, for each iteration of the ensemble learner, I generate a small population of candidates and pick the best one instead of searching exhaustively over the space.

In addition, keeping with the theme of simplicity, I use the simplest ensemble method possible: what I called "simple committees" in section 3.6.1. Instead of explicit tinkering with the training set to produce variation in the learner, I simply re-run the algorithm over and over again, depending on the stochastic nature of the search itself to produce variation.

The exact algorithm is given below. Since I am integrating the news data with the trading rule learner developed in chapter 3, it looks complicated.

A note about data: I have fourteen months of news data, split into thirds. Since I am using the news data in conjunction with the trading rule learner, I have to train the trading rule learner as well. But since the trading rule learner does not require news data, I can use price data that precedes the news data period; specifically, I use use two years of preceding price data and then fix the trading rules induced by the trading rule learner. To be specific about the time periods involved:

- Trading Rule Training Set: March 1, 1998 to February 28, 2001
- News Training set: March 1, 2001 to July 19, 2001
- News Test set: July 20, 2001 to December 6, 2001
- News Holdout set: December 7, 2001 to April 30, 2002

The full approach is described below. For each trial:

- Run the trading rule learner detailed in sections 3.6.3 and 3.5.3 over the trading rule learning set; fix the learned trading rules
- Apply each category as a gating signal to the learned trading rules as described above, assign those that produce a positive Sharpe ratio difference to set S.
- For $i = 1$ to 10
  - Generate a population of 20 candidates according to the following rules:
    * Pick n, the standard deviation cutoff, uniformly over the interval $[3, 6]$
    * Pick m, the holdout period, uniformly over the interval $[5, 20]$
    * For each candidate, evaluate its performance on the training set by using it as a gating signal over the learned trading rules according to the following rule: for each stock at day t, if the total news volume produced by news stories of set S on day t is greater than n standard deviations above the 30-day running mean, take no position in that stock for m trading days.
  - Select the best candidate out of the 20 by performance on the training set.
  - Generate a portfolio P_i over the test set by applying the above gating signal rule.
- Average portfolios $(P_1, P_2, ..., P_n)$ together, rebalance to ensure that aggregate long positions equal aggregate short positions, giving a final portfolio $P_{ensemble}$.

- Evaluate the performance of this $P_{ensemble}$.

One other caution should be made: since earnings periods are easy to avoid in real trading strategies, it is probably essential that I control for their effects here. To do so, I assume that the trading strategy stays out of the market for a given stock for 10 days after an "Earnings Report" news story is detected.

The results on the test set, averaged over twenty trials, are presented in table 5.6 over the test and holdout sets. In addition to using the "good" categories – categories that produce positive Sharpe ratio differences over the training set – as a check I also use the total news volume (a count of every news story, regardless of category). Numbers in parentheses represent p-values presented are from bootstrap hypothesis testing, where the bootstrap data sets are generated by randomly scrambling the counts of set S.

The total news volume produce strong Sharpe ratio differences in both the test and holdout sets, although the holdout set does not produce statistical significance. The results from the "Good" categories is more interesting: large positive Sharpe ratio differences in the test set, large negative Sharpe ratio differences in the holdout set. At this point, conclusions are hard to draw about the individual categories – they could be inconsistent, or it could happen to be a bad time; more data is needed to draw a strong conclusion.

5.8.6 News Category Sharpe Differences

Another interesting approach to the question of how much impact the news data can have is to directly examine the Sharpe ratio differences produced by each news category when used as a gating signal to the trading rule learner. In this section, I explore this category by category, both including and eliminating earnings periods from the trading strategy. Table 5.7 presents the Sharpe ratio differences produced by using each category as a gating signal to the trading rule learner, for the training set, test set, holdout set, and over the full set of data (it also presents the Sharpe ratio produced by the basic learned trading strategy alone, in the first row). This data is without any special exclusion of earnings periods; I repeat the experiment with earnings periods excluded later. I indicate bootstrap statistical significance with asterisks; a single asterisk represents statistical significance with a bootstrap p value of .05 or less. The results for using the raw news volume, trading volume, and message board volume (calculated as described in section 4.3.1) are presented in the last three rows. For the total news volume, message volume, and trading volume categories, and message board volume, the gating signal is triggered when the current volume (of news, messages, or shares traded) is four standard deviations above the 30-day moving average.

Note that despite the labeling of train, test, and holdout sets, there is no actual learning on the level of the news categories – I split the data up into sub-chunks merely to examine the temporal consistency of the performance of individual categories, and I keep the train/test/holdout set labels to maintain consistency with section 5.8.5.

From examining the rightmost column – the Sharpe differences produced over the entire time period – the first conclusion is that there are several categories that produce noticeable Sharpe differences of greater than .5 over the whole data set – "Analyst Downgrade", "Conference Call" (which usually happens at the same time as an earnings report), "Earnings Report", and "Earnings Outlook" all produce large, statistically significant Sharpe ratio differences. In addition, "Analyst Upgrade" produces notable improvements in Sharpe ratio that are not statistically significant. Unsurprisingly, these categories are related to earnings or analyst activity.

Intriguingly, the best aggregate results come from the total news volume and the total message volume, with Sharpe ratio differences of 1.04 and .57 over the entire time period, both statistically significant. Trading volume, on the other hand, produces lower results – .43 over the entire time period. The temporal patterns shown by the message volume and the trading volume are similar – negative Sharpe ratio differences in the first third, small but positive differences in the second, and strong positive differences in the third. In contrast, the news volume produces strong positive Sharpe ratio differences during all time periods.

It's likely that message volume and trading volume are all echoing the same underlying phenomena – I suspect message volume is driven strongly by trading volume, although intriguingly it does a better job of translating into improved Sharpe ratios.

Is it Just Earnings?

As discussed above in section 5.8.5, one doesn't need text classification to identify earnings reports – they're on the calendar. It is possible that many of the effects produced by the relevant categories are artifacts of earnings reports, and if so, there is no need to examine news to account for them.

To investigate this possibility, table 5.8 presents the Sharpe ratio differences for relevant categories, with the effects of earnings report periods excluded. In order to do this, I assume that the bare trading rule learner automatically stays out of the market for 15 days after earnings report news stories – the best proxy I have for earnings report periods. Then, each individual news category adds another set of potential gating signals to stay out of the market.

This approach yields minor, if intriguing differences. As to be expected, the "Earnings Report" category now produces no differences – its effects are now integrated into the basic strategy, strongly improving its Sharpe ratio to 2.39 over the full time period. "Conference Call" drops as well, to -.17, which is unsurprising, given its connection to earnings reports. The Sharpe ratio difference produced by "Analyst Downgrade" stays strong, although it does not produce statistical significance. Some categories that failed to show a positive Sharpe ratio differences improve – "Coverage Initiated" and "Dividend" show Sharpe ratio improvements of greater than .2, with only "Dividend" being statistically significant. The category "Earnings Outlook", improves its Sharpe ratio differences notably, and is statistically significant.

For the aggregate categories, the total news volume category produces a smaller improvement in Sharpe ratio, although the temporal pattern seems similar. The Sharpe ratio differences produced by message volume and trading volume produce both drop lower; intriguingly, they no longer produce similar temporal behavior, with trading volume producing strong results in both the first and final thirds of the data.

The total news volume is clearly the most useful indicator here; if I were building trading strategies for real, using the total news volume is clearly the best approach. Comparing the total news volume results with those produced in table 5.6, the learning approach seems to be slightly superior, producing higher Sharpe ratios in both the test and holdout periods; however, more data would definitely be needed before a strong conclusion could be reached.

5.8.7 Are the Categories Worthless?

The primary conclusion of the results presented above is that using the total amount of news as the signal is superior to using individual categories – either alone, or in some combination. This leads to the following question: is there any use in the category breakdowns at all?

Operationally, it doesn't seem so: if I just wanted to build a trading strategy, clearly, using the total news would be the best solution. However, I believe the categorizations are interesting for a couple of reasons.

First, As more data is accumulated, allowing for a finer resolution of analysis, undoubtedly interesting trends will pop out. Perhaps it will turn out that high-performing categories will prove consistent over time, and superior to the total amount of news. Perhaps the proper approach is to take the total news signal, and remove only a few categories. There are many other ways to try and utilize the news signals for trading strategies besides what I've worked on here. There are many possibilities; keeping the exploration of individual categories open is key to finding them.

Also, the categories are interesting in and of themselves, in the tradition of event studies. Exploring how the market reacts, for example, to analyst actions is interesting – the data above clearly suggests that analyst downgrades are in some operational sense more important than analyst upgrades.

The next section, 5.9, examines this question question with event study style results.

5.9 Event Studies

In this section, I borrow the methodology of the classic economics event study to examine market reaction to the categories of news stories. In contrast to previous section, this examines the impact of the news story alone, divorced from a linkage to any notion of a trading rule learner.

The methodology behind event studies is straightforward:

- Identify all occurrences of a specific type of event
- For each event occurrence, take the daily excess returns (the daily stock return minus the returns of the S&P 500) for n days before and after the event
- This gives an excess return series $r_{t-10}, r_{t-9} ... r_t, r_{t+1}, ... r_t + 10$
- Calculate the forward cumulative excess returns for days $t, ..., t+10$ by summing forwards: $cr_n = \sum_{i=t}^{n} r_i$

- Calculate the backward cumulative excess returns for days $t-10, ..., t-1$ by summing backwards: $cr_n = \sum_{i=n}^{t} r_i$
- Average both series of returns over all instances of the events.

The result is a profile of the average reaction to an event, both leading up to the event, and after the event occurrence. Of course, arguments about market inefficiency can only use the after return profile, but examining the before return profile can occasionally be illuminating.

Table 5.9 presents the results of such an analysis on the 39 classified categories as well as news volume and trading volume, where an event is defined as the occurrence of a classified news story, so long as no event has happened less than ten trading days previously. This exclusion condition is designed to prevent news stories concerning the same event that occur on subsequent days as counting as separate events.

The leftmost column presents the number of times each news category triggers and event. Moving rightwards, the columns present the average cumulative returns at times $t-10$, $t-5$, t, $t+5$, and $t+10$ respectively. The cumulative returns are all referenced against time t; that is, they are set to zero at time t.

Note that these results need to be interpreted gingerly. Traditional statistical significance tests of event studies depend on the events not overlapping in time; unfortunately, this is not possible here, as many of the news stories occur so frequently that they occur for multiple stocks on the same day, preventing traditional analysis of statistical significance.

The news categories that show the largest absolute effects also have few event occurrences. On the positive side, "Price Increases" and "Debt Repayment" show positive returns of 3.61% and 2.09% respectively, with 80 and 146 event occurrences. On the negative side, "Options", and "Earnings Restatement" produced -6.2% and -2.29% returns with 61 and 63 events, respectively. Given the low event occurrence frequencies for these categories, these results should not be taken as strong conclusions.

Intriguingly, the news categories that produce positive Sharpe difference ratios in conjunction with the trading rule do not product outstanding cumulative returns one way or another. Total news volume produces .24%, the three earnings categories produce insignificant positive returns in the range of .15% to .30%; only "Analyst Upgrade" produced absolute returns of greater than 1%, at 1.18%.

However, if one examines the pre-event return patterns, most of these categories show strong absolute returns in the days leading up to the event (note that negative pre-event cumulative returns correspond to prices rising before the event; positive pre-event cumulative returns correspond to prices declining before the event). "Analyst Upgrade", "Analyst Downgrade", "Earnings Warning" and "Earnings Outlook" all have absolute cumulative returns at $t-01$ of greater than 2%.

One additional caution about these results: The news stories can occur anytime before market close, but the returns are derived from close to close prices. Immediate price reaction to the news, before the current day's market close, is included in the day $t-1$ returns. Since price reaction to significant news is likely largest immediately after the news, the crude temporal resolution of price data used here likely obscures a number of interesting effects.

5.10 Conclusions & Future Work

The core conclusions of this chapter resolve around the results of sections 5.8.5 and 5.8.6. They can be summarized as follows:

- Hand-built classifiers can be built to identify successfully identify certain classes of news stories from headlines, providing roughly 90% precision and 70% recall.
- Using the presence of a news story as a gating signal to the trading rule learner developed in chapter 3 produces an increase in Sharpe ratio, under several different circumstances.
- A simple learning algorithm, using total news volume, produces strong differences in Sharpe ratio over the last two thirds of the data set. Using a collection of categories whose individual performance was good in the training set produced mixed results: good on the middle third of data, bad on the final third.
- Some individual news categories produce positive Sharpe ratio differences over the full 14 month dataset; these tend to center around earnings and analyst actions.
- The performance of the various individual news categories is often inconsistent over time; when divided into three sub-periods, most categories produce negative Sharpe ratio differences in at least one sub-period. However, given that the

sub-periods involved are less than five months each, this fact must be treated gingerly.

- The signal produced by the aggregate volume of news produces positive Sharpe differences over the entire data set. The aggregate volume of message traffic, and trading volume, produce lesser results.

- Clearly, there is useful information about the aggregate volume of news that is not simply a function of individual categories.

- Adjusting the algorithm so that time periods around earnings announcement news stories are automatically excluded does not dramatically change these results. However, for total message volume and trading volume, performance is dramatically reduced..

- Event study style analysis fails to show strong post-event cumulative returns, positive or negative, except for news stories that occur extremely infrequently.

These individual points reinforce the basic conclusion that was the primary goal of this chapter: news can be used to augment a trading rule learner to enhance performance.

5.10.1 Future Work

The work in this chapter asks far more new questions than it answers; the avenues for future work are almost too numerous to mention.

Overwhelmingly, the most important step forward is simply more data. Fourteen months worth of news is enough to show promise, but not enough to draw firm conclusions. Unfortunately, this cannot be rushed – it is a matter of time, waiting for news to accumulate. In addition to a simple lengthening of the time period, acquiring more densely sampled price data would allow for far more sensitive examinations of price reaction to news stories. Undoubtedly, the market reacts quickly to material news events, and increasing the temporal resolution of the analysis performed here from daily to hourly or every ten minutes would undoubtedly uncover intriguing trends.

One interesting project would be to compare the approach presented by Lavrenko [95] and his coauthors; they use a purely operational method of classification – classifying news stories purely in terms of their expected effect – positive or negative – on prices.

In addition to the temporal dimension, expanding beyond simple headlines is also crucial. Although headlines invariably contain the most important piece of information in a news story, there is undoubtedly more information to be gained from expanding the scope of the data to – at the bare minimum – the first paragraph. This would hopefully produce not only better classification accuracy, but also give a richer view of how each news story fits in the ontology.

Expanding beyond headlines will present challenges to the simple classification methodology used here. While the hand built classifiers used in this section work well enough to produce interesting results, clearly they can be augmented with other methods.

The obvious path to potential improvement is to take the hand built classifiers and integrate them with more sophisticated statistical methods. The existing classifiers make it easier to construct large data sets of sparsely occurring categories, a key advantage in constructing statistical methods.

Farther down the line, expanding the approach into a domain-specific classifier based on information extraction might produce higher accuracy, although the existing evidence for the superiority of that approach is not overwhelming.

The method of integrating the trading rule learner and the news data I use here is only one of many, many possible methodologies. Here, I use news stories as a gating signal to keep the trading rule learner away from certain stocks; it is certainly possible that some news stories might warrant increased, rather than reduced, positions.

Other interesting avenues of future research will no doubt be found in different breakdowns of the data. For this work, I have essentially treated all stocks identically, creating a split only by the volume of news produced. Certainly, the idea that stock prices would react differently to news stories according to sector classifications or market capitalization is intuitively appealing. Performing this kind of analysis could uncover powerful operational distinctions between sectors useful for real world implementation.

Table 5.3: Accuracy of hand-built classifiers

Classifier	Precision 3/12-3/18	Recall 3/12-3/18	F-Measure 3/12-3/18	Precision 5/1-5/31
Coverage Initiated	100% (179/179)	93.7% (179/191)	.97	99.0% (534/539)
Analyst Upgrade	95.7% (72/84)	100% (72/72)	.98	97.9% (319/326)
Analyst Downgrade	100% (102/102)	100% (102/102)	1.00	100% (443/443)
Conference Call	98.0% (49/50)	92.5% (49/53)	.95	99.2% (612/617)
Merger	92.9% (26/28)	83.9% (26/31)	.88	100% (255/255)
Acquisition	93.5% (86/92)	64.9% (86/131)	.77	96.4% (505/524)
IPO	100% (12/12)	81.3% (12/16)	.90	89.8% (97/108)
Minority Investment	16.7% (1/6)	14.3% (1/6)	.15	93.7% (89/95)
Shareholder Meeting	100% (9/9)	90.0% (9/10)	.95	97.9% (231/236)
Bankruptcy	100% (5/5)	71.4% (5/7)	.83	92.9% (39/42)
Lawsuit	100% (144/144)	82.6% (144/172)	.90	98.8% (749/758)
Regulatory Action	86.7% (13/15)	21.0% (13/62)	.34	95.0% (132/139)
SEC 10-K	100% (126/126)	100% (126/126)	1.00	100% (47/47)
SEC 10-Q	100% (57/57)	100% (57/57)	1.00	100% (1807/1807)
SEC 8-K	100% (145/145)	100% (145/145)	1.00	100% (655/655)
Earnings Warning	94.8% (91/96)	100% (91/91)	.97	88.4% (183/207)
Earnings Report	92.1% (175/190)	80.3% (175/218)	.86	97.9% (1650/1685)
Earnings Outlook	88.2% (82/93)	48.5% (99/169)	.63	85.9% (208/242)
Earnings Restatement	100% (2/2)	50% (2/2)	.67	100% (16/16)
Earnings Charge	100% (4/4)	100% (4/4)	1.00	100% (34/34)
New Financing	93.3%(42/45)	71.2% (42/59)	.81	92.0% (414/450)
Stock Buyback	100% (13/13)	68.4% (13/19)	.81	97.1% (67/69)
Debt Repayment	100% (2/2)	40% (2/5)	.57	73% (27/37)
Stock Split	100% (3/3)	100% (3/3)	1.00	100% (40/40)
Options	100% (5/5)	83.3% (5/6)	.91	100% (7/7)
Recall	N/A (0/0)	0% (0/3)	N/A	88% (22/25)
Product Research	100% (16/16)	28.6% (16/56)	.44	93.9% (108/115)
Price Cuts	100% (5/5)	50% (5/10)	.67	100% (50/50)
Price Increases	100% (2/2)	66.7% (2/3)	.80	88% (22/25)
Marketing	85.2% (23/27)	34.3% (23/67)	.49	80.8% (118/146)
CEO Comments	87.5% (7/8)	16.1% (7/31)	.27	89.3% (67/75)
Charity	100% (2/2)	6.3% (2/16)	.12	92.9% (26/28)
Award	93.1% (54/58)	62.4% (54/85)	.75	96.2% (201/209)
Belt Tightening	95.4% (41/43)	83.6% (41/55)	.89	98.3% (288/293)
Management Change	98% (149/152)	67.1% (149/222)	.80	97.0% (842/860)
Reorganization	100% (2/2)	12.5% (2/16)	.22	96.6% (56/58)
Executive Comp	100% (6/6)	60% (6/10)	.75	100% (20/20)
Labor Issues	33.3% (2/6)	33.3% (2/6)	.33	93.9% (46/49)
Dividend	98.8% (81/82)	96.4% (81/84)	.98	99.8% (443/444)

Table 5.4: Frequency of News Occurrence

Classifier	Actual 3/5-3/18	Classified 3/1/01-2/28/02	Estimate (Measured Recall)	Estimate (Recall = 1)
Coverage Initiated	314	4439	4690	4395
Analyst Upgrade	128	5707	5587	5587
Analyst Downgrade	295	6317	6317	6317
Conference Call	117	11916	12779	11821
Merger	63	2219	2645	2219
Acquisition	274	5396	8015	5202
IPO	28	881	973	791
Minority Investment	25	614	4023	575
Shareholder Meeting	28	723	786	708
Bankruptcy	14	757	985	703
Lawsuit	304	6492	7765	6414
Regulatory Action	118	1396	6315	1326
SEC 10-K	205	1652	1652	1652
SEC 10-Q	68	4998	4998	4998
SEC 8-K	294	5030	5030	5030
Earnings Warning	289	4322	3821	3821
Earnings Report	447	21924	26729	21464
Earnings Outlook	360	4724	8367	4058
Earnings Restatement	6	138	276	138
Earnings Charge	10	713	713	713
New Financing	103	3815	4929	3510
Stock Buyback	41	899	1276	873
Debt Repayment	11	296	541	216
Stock Split	4	208	208	208
Options	11	78	94	78
Recall	4	261	N/A	230
Product Research	88	864	2837	811
Price Cuts	21	392	784	392
Price Increases	15	155	205	137
Marketing	124	1438	3387	1162
CEO Comments	49	770	4271	688
Charity	42	420	6193	390
Award	152	1767	2724	1700
Belt Tightening	111	3937	4629	3870
Management Change	493	7015	10141	6805
Reorganization	34	773	5974	747
Executive Comp	12	201	335	201
Labor Issues	14	423	1193	397
Dividend	154	4271	4422	4262

Table 5.5: Augmenting Trading Rule Learner with Earnings News

Classifier	Sharpe Ratio	Difference
None	1.74	N/A
Earnings Warning	1.43	-0.32
Earning Report	2.12	0.38
Earning Outlook	2.10	0.36
Earning Restatement	1.61	-0.13
Earnings Charge	1.51	-0.23

Table 5.6: Sharpe Differences on Test and Holdout Sets

Method	Test Set Sharpe Difference	Holdout Set Sharpe Difference
"Good" Categories	1.92 ($p < .01$)	-2.01
Total News Volume	1.42 ($p < .05$)	1.10

Table 5.7: Differences in Sharpe ratio

Classifier	Train Set Sharpe Diff	Test Set Sharpe Diff	Holdout Set Sharpe Diff	Full Set Sharpe Diff
Trading Strategy	1.74	1.58	1.19	1.51
Coverage Initiated	-0.12	-0.02	0.67	0.14
Analyst Upgrade	0.38	0.25	-0.01	0.24
Analyst Downgrade	0.34	1.04 *	2.51 *	1.16 *
Conference Call	-0.65	0.60	1.85 *	0.51 *
Merger	0.20	-0.11	-0.06	0.02
Acquisition	0.33	-0.41	-0.08	-0.08
IPO	-0.03	-0.06	-0.02	-0.05
Minority Investment	0.14	0.06	-0.31	-0.03
Shareholder Meeting	-0.10	0.02	-0.05	-0.03
Bankruptcy	-0.13	-0.15	0.55	0.03
Lawsuit	-0.36	0.24	-0.48	-0.19
Regulatory Action	0.03	0.23	-0.09	0.05
SEC 10-K	0.04	0.04	-0.44	-0.12
SEC 10-Q	-0.42	-0.84	-0.36	-0.53
SEC 8-K	0.03	-0.33	0.15	-0.02
Earnings Warning	-0.32	0.48	-0.28	-0.06
Earnings Report	0.38	0.24	1.93 *	0.87 *
Earnings Outlook	0.36	1.26 *	0.06	0.55 *
Earnings Restatement	-0.13	0.03	0.16	0.01
Earnings Charge	-0.23	0.07	-0.09	-0.09
New Financing	0.06	0.17	-0.17	0.02
Stock Buyback	0.13	0.02	-0.03	0.04
Debt Repayment	-0.09	-0.01	-0.27	-0.12
Stock Split	0.02	0.02	0.04	0.02
Options	0.14	-0.01	0.02	0.05
Recall	-0.02	0.01	-0.11	-0.04
Product Research	-0.17	0	0.31	0.04
Price Cuts	-0.10	0.19	-0.03	0.01
Price Increases	0.03	0.02	0.01	0.02
Marketing	-0.17	-0.29	0.52	-0.01
CEO Comments	0.09	0.01	0.03	0.08
Charity	0	0	0.03	0.01
Award	-0.01	0.25	-0.16	0.01
Belt Tightening	-0.07	-0.29	-0.20	-0.18
Management Change	-0.20	0.55	-0.09	0.12
Reorganization	-0.01	-0.17	0.22	0.01
Executive Comp	0	0	-0.09	-0.05
Labor Issues	0.03	-0.08	-0.17	-0.05
Dividend	-0.15	0.36	0.22	0.14
Total News Volume	1.30 *	0.59 *	1.68 *	1.04 *
Total Message Volume	-0.30	0.38	2.11 *	0.57 *
Trading Volume	-0.09	0.08	1.62 *	0.43 *

Table 5.8: Differences in Sharpe Ratio, Excluding Earnings Report Periods

Classifier	Train Set Sharpe Diff	Test Set Sharpe Diff	Holdout Set Sharpe Diff	Full Set Sharpe Diff
Trading Strategy	2.12	1.82	3.12	2.39
Coverage Initiated	0.05	0.22	1.23	0.40
Analyst Upgrade	0.37	0.10	-0.80	0.01
Analyst Downgrade	-0.14	0.04	2.65 *	0.53
Conference Call	-0.88	-0.46	1.76	-0.17
Merger	0.40	-0.03	-0.28	0.06
Acquisition	0.51	-0.22	-0.60	-0.14
IPO	0.03	-0.10	0.11	-0.07
Minority Investment	0.09	0.06	-0.58	-0.10
Shareholder Meeting	-0.11	0.08	-0.14	0.04
Bankruptcy	-0.20	0	0.87	0.05
Lawsuit	-0.22	0.02	-0.36	-0.16
Regulatory Action	-0.03	0.24	0.21	0.14
SEC 10-K	0.47	0.02	-0.92	-0.10
SEC 10-Q	-0.09	-1.16	-0.68	-0.73
SEC 8-K	0.61	0.21	-0.02	0.31
Earnings Warning	-0.10	0.55	-0.17	0.13
Earnings Report	0	0	0	0
Earnings Outlook	1.05 *	0.57 *	0.95	0.92 *
Earnings Restatement	0.10	0.01	-0.14	0.01
Earnings Charge	-0.13	0.13	0.22	0.07
New Financing	-0.04	0.47	0.05	0.16
Stock Buyback	0.26	0.11	-0.06	0.13
Debt Repayment	0.04	0.01	-0.20	-0.03
Stock Split	0.03	0.02	0.02	0.03
Options	0.17	0.02	0.24	0.13
Recall	0.01	0.01	-0.09	-0.01
Product Research	-0.23	-0.10	0.15	-0.08
Price Cuts	-0.12	0.27	-0.03	0.05
Price Increases	0.04	0.01	0.01	0.02
Marketing	-0.35	-0.36	0.49	-0.17
CEO Comments	0.02	-0.07	0.53	0.11
Charity	0	-0.12	0.02	-0.05
Award	0.08	0.24	-0.11	0.09
Belt Tightening	-0.12	-0.18	0.18	-0.05
Management Change	-0.01	0.18	-0.11	0.08
Reorganization	-0.12	-0.27	0.13	-0.12
Executive Comp	0	-0.01	-0.16	-0.05
Labor Issues	0.11	-0.05	-0.14	-0.02
Dividend	0.06	0.33	0.32	0.24 *
Total News Volume	1.05 *	0.95 *	1.20 *	0.82 *
Total Message Volume	-0.19	-0.14	1.00	0.15
Trading Volume	1.29 *	-0.91	0.85	0.14

Table 5.9: Event Studies, First Occurrence

Classifier	# events	t-10	t-5	t	t+5	t+10
Coverage Initiated	1524	-0.52%	-0.24%	0%	-0.19%	-0.23%
Analyst Upgrade	1938	-2.22%	-2.19%	0%	0.90%	1.18%
Analyst Downgrade	2016	5.36%	5.10%	0%	-0.13%	0.02%
Conference Call	2209	-0.30%	-0.37%	0%	0.29%	0.50%
Merger	714	-0.07%	0.07%	0%	-0.50%	-0.45%
Acquisition	1675	-0.87%	-0.75%	0%	0.12%	0.51%
IPO	405	-0.65%	0.22%	0%	0.20%	0.09%
Minority Investment	350	-1.09%	-0.59%	0%	0.79%	1.05%
Shareholder Meeting	238	-0.88%	-1.49%	0%	1.05%	1.97%
Bankruptcy	282	3.16%	2.54%	0%	-0.43%	-0.69%
Lawsuit	1126	2.15%	0.71%	0%	0.04%	0.13%
Regulatory Action	711	0.92%	0.82%	0%	-0.29%	-0.10%
SEC 10-K	333	-1.00%	-0.78%	0%	1.06%	1.89%
SEC 10-Q	896	0.35%	0.11%	0%	0.15%	1.13%
SEC 8-K	1289	0.64%	0.52%	0%	-0.12%	-0.04%
Earnings Warning	1350	3.96%	3.41%	0%	-0.45%	0.28%
Earnings Report	3102	0.24%	0.02%	0%	0%	0.16%
Earnings Outlook	1694	2.71%	2.18%	0%	0.20%	0.15%
Earnings Restatement	63	8.16%	8.47%	0%	-3.54%	-2.29%
Earnings Charge	342	1.89%	0.89%	0%	0.61%	1.21%
New Financing	1116	0.19%	0.52%	0%	0.25%	0.66%
Stock Buyback	212	-1.93%	-1.84%	0%	1.22%	1.81%
Debt Repayment	146	4.59%	2.31%	0%	1.27%	2.09%
Stock Split	50	-4.11%	-2.32%	0%	-0.16%	0.85%
Options	61	-2.36%	0.10%	0%	-2.73%	-6.24%
Recall	111	-3.09%	-1.14%	0%	1.16%	1.43%
Product Research	254	-0.24%	0.36%	0%	1.12%	1.38%
Price Cuts	215	1.83%	1.17%	0%	0.76%	0.18%
Price Increases	80	-1.60%	-2.72%	0%	1.42%	3.61%
Marketing	825	-1.13%	-0.46%	0%	0.15%	0.78%
CEO Comments	580	0.29%	0%	0%	0.16%	0.68%
Charity	241	-1.37%	-0.08%	0%	0.10%	0.42%
Award	1001	-1.03%	-0.47%	0%	0.44%	0.98%
Belt Tightening	1262	2.00%	1.51%	0%	0.28%	0.54%
Management Change	2279	-0.97%	-0.26%	0%	0.08%	0.07%
Reorganization	369	0.80%	-0.01%	0%	0.98%	1.73%
Executive Comp	218	1.02%	0.42%	0%	0.18%	0.92%
Labor Issues	224	0.62%	1.01%	0%	-0.72%	-1.66%
Dividend	1008	-1.57%	-1.12%	0%	-0.05%	0.67%
Total News Volume	2894	1.01%	0.85%	0%	-0.05%	0.25%
Trading Volume	2281	2.39%	2.39%	0%	0.43%	0.74%

chapterConclusions and a Look Forward

The goal of this thesis was to start with a systematic exploration of common technical analysis techniques, add machine learning techniques, and then develop ways to classify business news use it to augment existing trading strategies. As the results presented in previous sections show, those goals were met. I discuss them now on a chapter by chapter basis.

5.11 Individual Chapter Conclusions

5.11.1 Technical Analysis

The chapter on technical analysis (Chapter 2) demonstrated that many technical analysis techniques do produce excess returns. However, they almost always provided positive results when used in ways not discussed by the practitioner literature – in fact, they usually worked in nearly the exact *opposite* way that the practitioner literature suggested.

Furthermore, the technical analysis techniques that produced excess returns nearly always functioned better on illiquid stocks – most of the excess returns come from stocks in the most illiquid half of the stock universe, where illiquidity is measured by the mean of trading volume multiplied by the price. Furthermore, the performance of the technical analysis degrades markedly as transaction costs rise, raising concerns about the practical implementability of such approaches.

5.11.2 Machine Learning

The goal of this chapter (chapter 3) was to see if simple machine learning techniques could be used to augment technical analysis techniques discussed above. The results show that machine learning can help, although in interesting ways. First, at least within the bounds of the genetic learner I explore, learning simple strategies proves to be superior to trying to learn complex strategies. Ensemble methods improve performance as well. And finally, examination of the strategies learned by the algorithm leads to an interesting conclusion: the strategies that work well are all based on short term moving averages.

5.11.3 Message Boards

This chapter primarily covers qualitative descriptions of message board data; integrating the message board data with trading strategies is covers in chapter 5.

5.11.4 News

This chapter, chapter 5, is the heart of the thesis. I built a comprehensive ontology of financial news stories, and hand-wired classifiers for most of the categories of the ontology. I explored the integration of these classifications with the trading rule learner developed in earlier sections, and found that using the total amount of news generated – looking for a surprisingly large amount of news produced, by any source – was the most successful approach in augmenting the Sharpe ratios produced by the trading rule. Some individual categories, linked to earnings and analyst actions, also produced strong Sharpe ratio differences

5.12 Contributions

To summarize the primary contributions of this thesis:

- A cataloging of technical analysis indicators and their effectiveness in a trading strategy framework.
- The use of a large stock universe – literally two orders of magnitude larger than the time series used in some comparable work – for testing the usefulness of technical analysis, machine learning, and news data.
- A deeper appreciation for just how noisy financial data is compared to traditional machine learning datasets, and an understanding of the use of representation and search to fight noise.
- The introduction of ensemble methods to the problem of learning trading rules.
- A methodology for fighting overfitting at the algorithm level in applying machine learning to financial data.

- The introduction of a framework to start research in earnest on how to use news data in financial research, in a set-up compatible with either either event study methodology or a learning trading rule approach.

- A productive example of integration between numerical market data and text data in a machine learning trading rule system.

5.13 Looking Forward

This thesis started walking down two well-traveled roads: the economic analysis of market efficiency, and the quest of AI to predict the stock market. However, the emphasis on using news data has sent it off in directions relatively unexplored as of yet; there is a tremendous amount of work to be done to extend the results here.

One sure area of progress in future work that applies to every chapter here is the application of sector breakdowns to the stocks. The work in this thesis has treated all stocks equivalently. But intuitively, the idea that stocks in different sectors produce and react to news differently is very plausible, and an obvious route of exploration.

Given the relationship between between the work presented in Chapter 3 and Hellström's [46] work on predicting rank measures, it is clear that using mean reversion on short term moving average ratios is extremely promising. I would focus machine learning attention on that representation. I think the most promising approach would be to fix the representation to short term moving averages, and use genetic search or memory based methods such as locally weighted regression, together with ensemble methods. The most important concern here is clearly the noise in financial data, and machine learning methods have much to contribute.

Of course, the most promising doors are those opened by the news approach. In addition to the obvious methodological enhancements such as using more than just headlines and integrating statistical techniques into the classification, the real possibilities are the new kinds of experiments this methodology allows. My event study-style analysis presented in section 5.9 is just a beginning – now, event study-like analyses can be done on arbitrary text events – rumors of events, anything that can be captured in text classification, whether explicitly or implicitly through sets of examples. Now, the effects or more or less any kind of event (well, any kind of event reported by news agencies) can be studied, if imperfectly due to the limitations of text classification.

This is the real contribution of the thesis. This is the bottom line: there is a long tradition of both economists and practitioners trying to understand the interrelationships between real world events and market activity. Understanding how to transform news, our best proxy for these events, into a numerical signal that is integrable with traditional statistical approaches allows the joining of two domains: the disciplined domain of statistical analysis, and the messy but ultimately crucial domain of real events in the real world, in ways that can revolutionize how we study markets.

Appendix A

Classifier Rule Descriptions

This appendix presents the actual classifiers used in chapter 5. Below I list each category for which I built a classifier (I built one for 39 out of 70 categories), each with its corresponding classifier.

Under each category there is a numbered list of rules. If *any* rule matches a headline, then it is counted as a category match. Each rule consists of one or more regular expressions in quotes, and either AND or NAND logical operators. When two or more regular expressions are linked with the AND operator, they both must match for the rule to match. When a regular expression is followed by a NAND and another regular expression, the rule matches only if the first regular expression matches and the second does not (note that this is not strictly a NAND operators, because if the second matches and the first does not, the rule still doesn't match). Regular expressions [56] are a powerful technique for text matching ([40] for a practical introduction). Unfortunately, they are too complex to exhaustively describe here; however, in the list below, I give descriptions and examples of the regular expression rules needed to understand the classifiers presented.

- In general, a string of letters and numbers simply means "match that string".
  - For example, the regular expression "charge" would match any occurrence of the string "charge".
- The | character is a logical OR operator.
  - For example, the regular expression "no charge|ahead|to charge" matches either the string "no charge", the string "ahead", or the string "to charge"

- Parenthesis are used similarly in mathematical notation, as a grouping operator.
  - For example, the regular expression "initiator(e|Inge|es|ed)" would match the strings "initiate", "initiating", "initiates", or "initiated".
- A question mark means "match this expression 0 or 1 times"
  - For example, the regular expression "charge(d|s)?" would match the strings "charge, "charges, or "charged".
- The escaped character "\d" matches a single digit.
  - For example, the regular expression "\d" matches the strings "0", "1", "2", "3", "4", "5", "6", "7", "8", or "9".
- A plus sign ("+") means "match this expression 1 or more times"
  - For example, the regular expression "\d+" matches any string of digits
- Characters in brackets match any of the *individual* characters.
  - For example, [sd] matches the strings "s" or "d"; the regular expression "change[sd]?" matches the strings "change, "changes", "changed"

A.1 The Classifiers

Here, I present the rules for classifiers used for each of the 39 categories discussed in chapter 5.

- Coverage Initiated: Analyst initiates coverage of a publicly traded company. ("Coverage initiated on Sprint PCS by Dresdner Klnwrt Wasserstein", PCS, 3/5/2001)
  1. "coverage initiated"
  2. "open(s)?|expand(s|ed)?|initiat(ing|es|ed)?|announc(es|ed)?" AND "coverage"
  3. "open(s)?|initiat(ing|es|ed)?|announc(es|ed)?" AND "coverage"

- Analyst Upgrade: Report of an analyst upgrading their rating of a publicly traded company. ("Marsh McLennan upgraded by Bear Stearns", MMC, 3/9/2001)

  1. "upgrade(s|d|ing)" NAND "Moody's|S&P|debt"
  2. "upgrade" AND "surge(s)?|lead(s)?|lift(s)?|boost(s)?|rise(s)?|soar(s)?"

- Analyst Downgrade: Report of an analyst downgrading their rating of a publicly traded company. ("Motorola downgraded by Merrill Lynch", MOT, 3/9/2001)

  1. "downgrade(s|d|ing)?" NAND "Moody's|S&P|debt"

- Credit Upgrade: One of the major ratings agencies raises their credit rating of the company. ("Moody's Upgrades Senior Secured Bank Debt Ratings of Nextel Partners to B1", NXTP, 4/5/2001)

  1. "upgrade(s|d)?" AND "Moody's|S&P|debt"

- Credit Downgrade: One of the major ratings agencies lowers their credit rating of the company. ("Moody's lowers GMs credit outlook", GM, 4/7/2001")

  1. "downgrade(s|d)?|cut(s)?|lower(s)?" AND "Moody's|S&P"
  2. "downgrade(s|d)?|lower(s)?" AND "debt"
  3. "warning" AND "Moody's|S&P|debt"

- Conference Call: Mention of a company's scheduled analyst conference call. ("AppliedTheory Earnings Call scheduled for 10:00 am ET today", ATHY, 3/9/2001)

  1. "conference call(s)?|earnings call(s)?" NAND "merger|acquisition|shareholder(s)?"
  2. "quarter(ly)?" AND "call"

- Merger: News about a merger involving a publicly traded company. ("James River Bankshares to Merge With First Virginia Banks", FVB, 3/5/2001)

  1. "merger|merge(s)?" NAND "post|acquisition(s)?|charge(s)?|expenses|cost(s)|operations"
  2. "link-up"

- Acquisition: News about an acquisition involving a publicly traded company, either another company or specific assets of another company. (‘Convergys to Acquire U.K. Billing Software Maker”, CVG, 3/6/2001)

  1. “to acquire|will acquire|acquires|acquired|acquisition” NAND “$8-K$|minority|rights|license|stake|technology”
  2. “to acquire|will acquire|acquires|acquired|acquisition” AND “majority stake”
  3. “poison pill”
  4. “tender offer” NAND “bond(s)?”
  5. “takeover|buyout|take over” NAND “position”
  6. “bid(s)?|buy(s|ing)?” AND “$[\d]+” NAND “bond(s)?|note(s)?|upgrade|debt|coverage|$8 - K$|minority|rights|license|stake|technology”
  7. “pay(s|ing)?” AND “$[\d]+” NAND “bond(s)?|note(s)?|upgrade(s)?|debt|coverage|$8 - K$|minority|rights|license|stake|technology” AND “cash”

- IPO/Spinoff: News about a publicly traded company spinning off one of its divisions. (“Kraft Foods Files for $5 Billion IPO”, MO, 3/16/2001)

  1. “ipo|spinoff|initial public offering” NAND “post|since IPO|from IPO|class action(s)?|lawsuit”

- Minority Investment: News about a company taking, or selling a minority stake in another company. (“Motorola Inc. Anchors Second Round Financing in WatchPoint Media, Inc.”, MOT, 3/8/2001)

  1. “cut(s|ting)?|tak(e|es|ing)|buy(s|ing)?|acquir(e|es|ed|ing)|rais(es|ed|ing)|boost(s|ed|ing)?|increas(e|es|ed|ing)|lift(s|ed)|up(ped|s)?” AND “stake(s)?” NAND “majority”
  2. “dump(s|ed|ing)?|sell(ing|s)?|sold|liquidat(e|ed|ing)” AND “stake(s)?” NAND “majority”
  3. “purchase(s)?” AND “stake” NAND “majority|shareholder(s)?”
  4. “sale” AND “stake” NAND “majority|shareholder(s)?”
  5. “minority interest|minority stake”

- Shareholder Meeting: News about a shareholder meeting, or other shareholder issues. (“Maxim Pharmaceuticals Announces Results of 2001 Annual Stockholders Meeting”, MAXM, 3/9/2001)

  1. “shareholder(s|s')?|stockholder(s)?” AND “meeting”;
  2. “special meeting”
  3. “annual meeting”
  4. “shareholder vote”

- Bankruptcy: News of a company filing for bankruptcy. (“Finova Files Chapter 11”, FNV, 3/7/2001)

  1. “bankrupt|bankruptcy|chapter 11”

- Lawsuit: News about a lawsuit (potential or actual) or arbitration involving the company or one of its key employees. Most of these news stories are announcements of class action suits. (“Metromedia Trial Set for March; Shareholder Activist Seeks Right to Inspect Company’s Books and Records”, MMG, 3/5/2001)

  1. “class action(s)?|class period(s)?|arbitration|su(es|ed|ing)?”
  2. “patent(s)?” AND “dispute|settle(d)?|suit|protect”
  3. “court battle|restraining order|litigation|injunction|lawsuit(s)?”
  4. “fil(e|es|ing)|throw(s)? out|launch(es)?|dismiss(es)?|toss(es)?|win(s)?|lose(s)?|announce(s)?” AND “(law)?suit” NAND “class”
  5. “seek damages”

- Regulatory action: News about regulatory action – at the state, national, or potentially supranational level – involving the company. This includes investigations by regulatory agencies. (“SEC Probing Bezos Stock Sales, New York Times Reports”, AMZN, 3/9/2001)

  1. “legislators|regulators|lawmakers”
  2. “FDA|SEC|FTC|FCC” AND “approval(s)?|clearance|warning|go ahead|review|ok|ban|inquiry|request”

   3. “FDA|SEC|FTC|FCC” AND “approve(s|d)?|investigat(e|es|ed|ing)|prob(es|e|ed|ing)|cite(s|d)?|order(s|ed)?|question(s|ed)?”
   4. “FTC” AND “charge(s|d)?|accuse(s|d)?”
   5. “investigating|probing|probe(s)?”

- SEC Filing 10-K: Report of a SEC form 10-K filing. These reports are almost invariably from Edgar, and are straightforward. (“EARTHLINK INC - Annual Report (SEC form 10-K)”, ELNK, 3/9/2001)

   1. “(\()*(10-K)(\))*”

- SEC Filing 10-Q: Report of a SEC form 10-Q filing. (“LITTON INDUSTRIES INC - Quarterly Report (SEC form 10-Q)”, LIT, 3/8/2001)

   1. “(\()*10-Q(\))*”

- SEC Filing 8-K: Report of a SEC form 8-K filing. (“NEUBERGER BERMAN INC FILES (8-K) Disclosing Other Events”, NEU, 3/9/2001)

   1. “(\()*8-K(\))*”

- Earnings Warning: A company issues an earnings warning, lowers earnings estimates, or warns about coming revenue shortfalls. (“ON The Move: New Focus Clearly Warns”, NUFO, 3/6/2001)

   1. “warn(s|ing|ings)?” NAND “letter|FBI|sales|hoax|virus|worm”
   2. “miss estimates|miss numbers”
   3. “earnings warning(s)?”

- Earnings report: News about an earnings report (“Q4 2000 Guess Earnings Release - After Market Close”, GES, 3/8/2001)

   1. “earnings release” NAND “conference call”
   2. “release(s)?|releasing|report(s|ed)?|announce(s)?” AND “earnings|income|revenue(s)?” NAND “conference call|traffic|meeting|profit”
   3. “release(s)?|releasing|report(s|ed)?|announce(s)?” AND “results” NAND “sales|conference call|traffic|meeting|trial|study”

4. “beat(ing|s)?|top(s|ped)?|meet(s)?” AND “estimate(s)?|expectations|forecast(s)?|target|view(s)?” NAND “sales|to meet”
5. “in line with” AND “estimate(s)?|expectations|forecast(s)?|view(s)?” NAND “sales”
6. “beat(s)? (the)? street”
7. “miss(es|ed)|fall(s)? short” AND “estimate(s)?|expectations|forecast(s)?|view(s)?” NAND “sales”

- Earnings outlook: News about estimated future earnings, revenues, and profits. (“Inforte Corp. Remains Comfortable with March 2001 Quarter Guidance”, INFT, 3/8/2001)

1. “earnings” NAND “warn(s|ing|ings)?|conference|release|releases|releasing|report[s]?|announce[sd]?|call|post(s|ed)?”
2. “revise(s|ed)?|provide(s)?|update(s)|issue(s)?|adjust(s)?” AND “guidance|outlook|estimate(s)?|expectation(s)?|forecast(s)?|view(s)?|projection(s)?”
3. “lowered” AND “guidance|outlook|estimate(s)?|expectation(s)?|forecast(s)?|view(s)?|projection(s)?|earning(s)?|target(s)?” NAND “price target(s)?|sales|warn(s|ing)?|meet(s)?|matches|miss(es)|shipment|subscriber|beat(ing|s)?|top(ping|s)?|hit(ing|s)?|exceed(s|ing)?”
4. “lowers|cut(s|ting)?|reduce(s)?|chop(s|ped)?|slash(es)?” AND “guidance|outlook|estimate(s)?|expectation(s)?|forecast(s)?|view(s)?|projection(s)?|earning(s)?|target(s)?” NAND “price target(s)?|sales|warn(s|ing)?|meet(s)?|matches|miss(es)|shipment|subscriber”
5. “see(s)?|maintain(s)?|(re)?affirm(s)?|confirm(s)?|reiterate(s)?|give(s)?|stick(s)? to|offer(s)?” AND “guidance|outlook|estimate(s)?|expectation(s)?|forecast(s)?|view(s)?|projection(s)?|target(s)?” NAND “price target(s)?|sales”
6. “to exceed” AND “guidance|outlook|estimate(s)?|expectation(s)?|forecast(s)?|view(s)?|projection(s)?|target”
7. “comment(s)? on” AND “outlook”
8. “guide(s)? down”
9. “earnings guidance”

- Earnings Restatement: News about a restatement of past earnings. ("Kroger to Restate Earnings by 'Minor Amounts"', KR, 3/5/2001)

  1. "restate(s|d)?|restatement(s)?"

- Earnings Charge: News about a charge taken to earnings. ("NCR Anticipates Charge Related to ATM Equipment Distributor.", NCR, 3/6/2001)

  1. "charge(s)?" NAND "no charge|ahead|to charge|face(s)?|settle(s)?|higher|arrest(s|ed)?"
  2. "take" AND "charge(s)?" NAND "charge of"
  3. "post(s)?|see(s)?|set(s)|anticipate(d|s)?|predict(ed|s)?" AND "charge(s)?"
  4. "special|restructuring|nonrecurring" AND "charge(s)?" NAND "before|pre-charge|excluding"
  5. "quarter" AND "charge(s)?" NAND "before|pre-charge|excluding"

- New Financing: Discussion of issues of new debt, shares, other securities, or loans. ("Charter Communications files $4 billion shelf", CHTR, 3/9/2001)

  1. "offer(s)?|sell(s)?|issue(s)?|sale(s)?|pricing|price(s)?|launch(es)?|placement|raise(s)?|tap(s)?" AND "note(s)?|bonds|debt|debenture(s)?" NAND "pay down|research"
  2. "offer(s)?|issue(s)?|sale(s)?|pricing|launch(es)?|placement|raise(s)?|tap(s)?" AND "shares"
  3. "file(s)?|filing" AND "shelf"
  4. "public offering(s)?" NAND "initial|IPO(s)?"
  5. "secondary offering(s)?|equity offering(s)?|note offering(s)?"
  6. "offering(s)?" AND "note(s)?|stock|securities|debt"
  7. "shelf" AND "$[\d]+\w*"
  8. "credit" AND "facility"
  9. "equity" AND "financing"
  10. "securitization program"

- Stock Buyback: News of a stock buyback. ("Waddell & Reed Financial Announces Share Repurchase", WDR, 3/5/2001)

1. “share(s)?|stock” AND “repurchase(s)?”
2. “redeem|buy back|buyback|buy-back” AND “stock|share(s)?”
3. “stock|share(s)?” AND “redemption”

- Debt Repayment: News of debt repayment. (“NHI Further Reduces Indebtedness, Lowers Interest Rate”, NHI, 3/7/2001)

  1. “repay(s)?|buy(s)? back|buyback|trim(med|s)?|reduce(s)?” AND “debt|indebtedness”
  2. “reduce(s|ed)?|pay down|prepay|prepayment|repayment|cut(s)?” AND “debt|indebdetness”
  3. “redeem(s|ed|ing)?” AND “notes|bonds|debt”
  4. “call|redemption” AND “notes|debt|bonds”

- Stock Split: News about a stock split. (“Frontier Airlines’ 3-for-2 Stock Split to Take Effect on March 6, 2001”, FRNT, 3/5/2001)

  1. “stock split” NAND “reverse”
  2. “split(s)?” AND “stock” NAND “reverse”

- Options: News of a company’s treatment of employee options. (“TIBCO Software Announces Stock Option Exchange Program for Employees”, TIBX, 3/8/2001)

  1. “option(s)?” AND “employee(s)?|re-pricing|reprice|regrant|grant(s)?”
  2. “option(s)?” AND “exchange” AND “voluntary”
  3. “option plan”

- Recall/defect: News of a product recall or defect. (“CPSC, McDonald’s Announce Recall of ’Scooter Bug’ Happy Meal Toys”, MCD, 3/7/2001)

  1. “recall(s)?|defect(s)?”

- Product Research: Discussion of product research, primarily the granting of patents and clinical drug trials. (“AVANT Announces Changes to Its Phase IIb Trials of TP10 In Infants Undergoing Cardiac Surgery”, AVAN, 3/6/2001)

  1. “win(s)?|receive(s)?|issue(s|d)?|award(s|ed)?” AND “patent(s)?” NAND “lawsuit|dispute|suit|litigation”

2. “clinical|phase” AND “trial(s)?|test(s|ing)?|study”
3. “begin(s)|initiate(s)?|launch(es)?” AND “trial(s)?|test(ing|s)?|study”
4. “new study|comparison study|follow-up study”
5. “(study|trial) (data|results)”

- Price Cuts: News about cutting prices. (“Intel, AMD Unveil More Price Cuts”, AMD, 3/5/2001)

  1. “price cuts”
  2. “slash(es|ing)?|cut(s|ting)?|reduce(s)?” AND “price(s)?” NAND “target(s)?”
  3. “reduction(s)?” AND “price(s)?” NAND “target”
  4. “price war”
  5. “lower(s) prices”

- Price Increases: News about raising prices. (“Eastman Announces Price Increase for Oxo Alcohols”, EMR, 3/5/2001)

  1. “price increase(s)?|price hike(s)?” NAND “recind”
  2. “raise(s)?|rebound(s)?|increase(s)?” AND “price(s)?” NAND “target|IPO|recind”

- Marketing: News about a public company’s marketing or branding campaigns, including sponsorships of major events and conferences, and contests and sweepstakes. (“SunTrust Bank Launches New Branding Effort; Customer Service at Helm of ’Unexpected’ Campaign”, STI, 3/6/2001)

  1. “branding|co-branding|co-brand|new brand|re-branding” NAND “head”
  2. “(co-)?marketing” NAND “agreement|collaboration|alliance|partnership|pact|VP|vice-president(s)?|president(s)?|director|officer|chief|head|executive(s)?|solution|approval|lead|catalina|student loan|marketing company|marketing corporation”
  3. “advertising” AND “review”
  4. “official supplier”
  5. “sponsor(s|ship)?” NAND “401”

6. “marketing program”
7. “sweepstakes|consumer promotion|contest” NAND “no contest|to contest”
8. “debut(s|ed)?|launch(es|ed)?|unveil(s|ed)?” AND “advertising|ads|brand|marketing|ad campaign|promotion(s)?”

- CEO Comments: A story about the CEO publicly commenting on issues concerning the company. Occasionally comments are made by the CFO or COO, or other high ranking official instead. This does not include CEO presentations at conferences, which are counted under “Public Presentation”. (“Ameritrade chair doesn't see M&A in top e-brokers”, AMTD, 3/5/2001)

  1. “CEO|CFO|COO|chair” AND “promise(s|ed)|cite(s|ed)?|see(s)?|promise(s)?|eye(s)?|say(s)|tell(s)|address(es)?|attribute(s|ed)?|describe(d|s)?|outline(s|ed)?|speak(s)?|respond(ed|s)?|present(ed|s)?|applaud(ed|s)?|like(s)?|vow(s)?|highlight(s)?|hope(s|ed)?” NAND “conference|forum|symposium|meeting|event|resign(s|ed)?|name(s|d)?|to speak|to address|retire(s|d)?”
  2. “CEO|CFO|COO|chair” AND “interview|comments”

- Charity: News about charitable activities or sponsorships. (“BroadVision CEO And His Wife to Donate $15 Million to Fund Particle Astrophysics and Cosmology Institute”, BVSN, 3/6/2001)

  1. “donate(s|d)?” AND “$[\d]+\w?”
  2. “charity|charitable|donation|scholarship(s)?|fundraising”
  3. “raise(s|d)?” AND “$[\d]+\w?” AND “for” NAND “offer(ing)?”

- Award: News about awards won, records set, or special recognition of the company the company, or of employees of the company (“eXcelon e-Business Products Win Best of EnterpriseVision Award”, EXLN, 3/5/2001)

  1. “world record|honor(ed|s)?” NAND “honor system”
  2. “award” NAND “award-winning|damages|jury|to award|damage|award winner|finalist(s)?|annual”
  3. “win(s)?|receive(s|d)?|capture(s|d)?|earn(s|ed)?|garner(s|ed)?|bestow(s|ed)?|acheive(d|s)?” AND “award(s)?” NAND “damage|jury|contract”

4. “win(s)” AND “best”
5. “recognized”
6. “cite(s|d)?|recognize(s|d)?|recognition” AND “excellence”
7. “best of show|product of the year”
8. “named” AND “top”

- Belt Tightening: News about layoffs and restructuring, production cuts (“Copper Mountain cuts jobs, VIPs resign”, CMTN, 3/7/2001)

1. “laid off|lay(s)? off|layoff(s)?|job cuts”
2. “ax(es|ed|ing)?|slash(ing|es)?|cut(ting|s)?|trim(s|ed|ming)?|pare(s)?|slice(s|d)?|lay(s) off|shed(s|ding)?|reduc(es|ed|ing)” AND “job(s)?|staff(ing)?|workers|work force|workforce| employees|production”
3. “elimination|reduction(s)?” AND “job(s)?|staff(ing)?|workers|work force|workforce| employees”
4. “cost cutting|cost-cutting|hiring freeze”
5. “cuts back”
6. “slash(ing|es)?|cut(ing|s)?|trim(s|ed|ming)?|pare(s)?|slice(s|d)?” AND “expenses|costs| pay|spending”
7. : “clos(e|ing)|shut(s|ting)?” AND “plant”

- Management Change: News concerning turnover in key management position, such as CEO or other corporate officer. (“Nolan’s Tech Tidbits: Wal-Mart.com CEO May Be Leaving”, WMT, 3/8/2001)

1. “promote[sd]?|name[sd]?|appoint(s|ed)?|join(s|ed)?|tap(s|ped)?|succeed(s|ed)?| add(s|ed)?” AND “VP|COO|CFO|CEO|CIO| Chairman|Chief|president(s)?|vice-president(s)?| chairman|director(s)?|head|exec|board|management|executive(s)?”
2. “announce(s|d)?” AND “VP|COO|CFO|CEO|CIO|Chairman|Chief|president(s)?|vice-president(s)?|chairman|director(s)?|head|exec|board|management|executive(s)?” AND “new|appointment(s)?|change|promotion(s)?|retirement|election|resignation”
3. “step(s|ped)? down|step(s|ped)? aside|resign(s|ed)?” AND “VP|COO|CFO|CEO|CIO| Chairman|Chief|president(s)?|vice-president(s)?|chairman|director(s)?|head|exec|board| management|executive(s)?”

4. “search” AND “VP|COO|CFO|CEO|CIO|Chairman|Chief|president(s)?|vice-president(s)?|chairman|director(s)?|head|exec|board|management|executive(s)?”
5. “executive shake-up|executive departure(s)?|executive resignation(s)?”

- Reorganization: News of corporate reorganization. Sometimes this is the straightforward reorganization of business units; often it is a cover for job cuts (“AOL Time Warner Combines TV Networks Under Turner Broadcasting”, AOL, 3/6/2001)

  1. “reorganization|reorganiz(es|ed|ing)|restructur(e|es|ing)” NAND “debt|chap(ter)? 11|charge(s)?|financial|file(s|d)?|agreement(s)?|relationship|acquisition|rate(s)?”
  2. “new organization(al)?|organizational change(s)?”

- Executive Compensation: News about the compensation of key executives. (“Whirlpool CEO’s bonus cut in half”, WHR, 3/9/2001)

  1. “executive(s)?|exec(s)?|CEO|brass|chief|chairman” AND “bonus(es)?|pay(s)?|compensation|salary|pay hike”

- Labor Issues: News about strikes and other labor issues. (“Machinists union opposes United, US Airways merger”, 3/7/2001)

  1. “strik(ing|es)” NAND “deal(s)?|three|two|out|gold|partnership|balance|back|alliance”
  2. “union(s)?” NAND “union planters|credit union(s)?|first union|union pacific|union carbide|western union|union bank|european union|southern union|union state bank|stake(s)?”

- Dividend: News relating to a stock’s dividend. (“Wal-Mart Increases Dividend By 18 Percent”, WMT, 3/9/2001)

  1. “dividend(s)?” NAND “dividend growth|dividend miles”
  2. “distribution policy”

Bibliography

[1] Business Wire. `http://www.businesswire.com`.

[2] CBS Marketwatch. `http://www.cbsmarketwatch.com`.

[3] Forbes.com. `http://www.forbes.com`.

[4] Reuters. `http://www.reuters.com`.

[5] Silicon Investor's Discussion Forums. `http://www.siliconinvestor.com/stocktalk`.

[6] The Motley Fool's Discussion Forums. `http://boards.fool.com`.

[7] The Motley Fool's Index Center. `http://www.fool.com/school/indices/introduction.htm`.

[8] The Wall Street Journal Online. `http://www.wsj.com`.

[9] Yahoo! Finance. `http://finance.yahoo.com/`.

[10] Ragingbull.com. `http://ragingbull.lycos.com/`, 1999.

[11] Yahoo! Stock Specific Message Boards. `http://messages.yahoo.com/yahoo/Business___Finance/Investments`, 1999.

[12] SEC brings charges in internet manipulation scheme. SEC press release 2000-135, SEC, September 2000.

[13] Steven B. Achelis. Technical Analysis A to Z. `http://www.equis.com/free/taaz/`.

[14] J.T. Alander. On the optimal population size of genetic algorithms. In *Proceedings of CompEuro 92*, pages 65–70. IEEE Computer Society Press, 1992.

[15] Franklin Allen and Risto Karjalainen. Using genetic algorithms to find technical trading rules. *Journal of Financial Economics*, 51(2):245–271, 1999.

[16] Harry Mamaysky Andrew W. Lo and Jiang Wang. Foundations of technical analysis: Computational algorithms, statistical inference, and emprical implementation. *Journal of Finance*, 40(4), August 2000.

[17] Chidanand Apte, Fred Damerau, and Sholom M. Weiss. Text mining with decision rule and decision trees. In *Working Notes of Learning from Text and The Web, Conf. Automated Learning and Discovery (CONALD-98).* Pittsburgh, PA: Carniegie Mellon University, 1998.

[18] P. Asquith and D. Mullins. The returns to acquiring firms in tender' offers: Evidence from three decades. *Journal of Financial Economics*, 15:61–69, 1986.

[19] P. W. Munro B. Parmanto and H. R. Doyle. Improving committee diagnosis with resampling techniques. In *Advances in Neural Information Processing Systems*, volume 8, pages 832–888, 1996.

[20] Tobias Blickle and Lothar Thiele. A comparison of selection schemes used in evolutionary algorithms. *Evolutionary Computation*, 4(4):361–394, 1996.

[21] Leo Breiman. Bagging predictors. *Machine Learning*, 24:123–140, 1996.

[22] William Brock, Josef Lakonishok, and Blake LeBaron. Simple technical trading rules and the stochastic properties of stock returns. *Journal of Finance*, 47(5):1731–1764, 1992.

[23] John L. Casti. *Searching for Certainty*. William Morrow, 1990.

[24] C. W. Cleverdon. Progress in documentation. evaluation of information retrieval systems. *Journal of Documentation*, 26:55–67, 1970.

[25] C. W. Cleverdon. On the inverse relationship of recall and precision. *Journal of Documentation*, 28:195–201, 1972.

[26] D.Challet and Y.-C. Zhang. Emergence of cooperation and organization in an evolutionary game. *Physica A*, (246), 1997.

[27] D.Challet and Y.-C. Zhang. On the minority game: Analytical and numerical studies. *Physica A*, (256):514–532, 1998.

[28] T. G. Dietterich. Ensemble methods in machine learning. *Machine Learning*, 24:1–15, 2000.

[29] Thomas G. Dietterich. An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization. *Machine Learning*, 40(2):139–157, 2000.

[30] Darrell Duffie. *Dynamic Asset Pricing Theory*. Princeton University Press, 1996.

[31] Bradley Efron and Robert J. Tibshrani. *An Introduction to the Bootstrap*. Chapman & Hall, 1993.

[32] Christos Faloutsos and Douglas W. Oard. A survey of information retrieval and filtering methods. Technical Report CS-TR-3514, University of Maryland, 1995.

[33] Eugene F. Fama. Efficient capital markets: A review of theory and empirical work. *Journal of Finance*, (25):383–423, 1970.

[34] Eugene F. Fama. Efficient capital markets: II. *Journal of Finance*, (46):1575–1617, 1991.

[35] Tom Fawcett and Foster Provost. Activity monitoring: Noticing interesting changes in behavior. In Chaudhuri and Madigan, editors, *Proceedings on the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 53–62, San Diego, CA, 1999.

[36] Tom Fawcett and Foster J. Provost. Adaptive fraud detection. *Data Mining and Knowledge Discovery*, 1(3):291–316, 1997.

[37] D. B. Fogel. *Evolutionary Computation: Toward a New Philosphy of Machine Intelligence*. IEEE Press, 1995.

[38] Yoav Freund and Robert E. Shapire. A decision-theoretic generalization of on-line learning and an application to boosting. Technical report, AT&T Bell Laboratories, Murray Hill, NJ., 1995.

[39] Yoav Freund and Robert E. Shapire. Experiments with a new boosting algorithm. In *Proceedings of the 13th International Conference on Machine Learning*, pages 146–148, 1996.

[40] Jeffrey E. F. Friedl. *Mastering Regular Expressions*. O'Reilly and Associates, 2002.

[41] William Fung and David Hsieh. The risk in hedge fund strategies. *Review of Financial Studies*, 14:313–341, 2001.

[42] David Goldberg. *Genetic Algorithms in Search, Optimization, and Machine Learning*. Addison-Wesley, 1989.

[43] C.A.E. Goodhart and Margaret O'Hara. High frequency data in financial markets: Issues and applications. *Journal of Empirical Finance*, (4):73–114, 1997.

[44] Sanford J. Grossman and Joseph E. Stiglitz. The impossibility of informationally efficient markets. *American Economic Review*, 70, 1980.

[45] L. Harris. Transaction data study of weekly and indradaily patterns in stock returns. *Journal of Financial Economics*, 16(1):99–117, 1986.

[46] Thomas Hellström. Predicting a rank measure for stock returns. *Theory of Stochastic Processes*, 3:64–83, 2000.

[47] Thomas Hellström. Optimizing the sharpe ratio for a rank based trading system. 2001.

[48] Thomas Hellström and Kenneth Holmström. Predicting the stock market, 1998.

[49] Arthur Hill. Stockchart.com's Chart School. `http://www.stockcharts.com/education/`.

[50] John Holland. *Adaptation in Natural and Aritificial Systems*. Ann Arbor: University of Michigan Press.

[51] John Holland. Genetic algorithms and the optimal allocation of trials. *SIAM Journal of Computing*, 2(2):88–105, 1973.

[52] G. Jarrell and A. Poulsen. The returns to acquiring firms in tender' offers: Evidence from three decades. *Financial Management*, 18:12–19, 1989.

[53] Paul Jefferies, Michael L. Hart, P. M. Hui, and Ned F Johnson. From market games to real-world markets. 2000.

[54] Torsten Joachims. Categorization with support vector machines: Learning with many relevant features. Technical Report LS8- Report 23, Universität Dortmund, 1997.

[55] Ned F Johnson, David Lamper, Paul Jefferies, Michael L. Hart, and Sam Howison. Application of multi-agent games to the prediction of financial time series. *Physica A*, (299):222–227, 2001.

[56] Stephen Kleene. Representation of events in nerve nets and finite automata. In *Automata Studies*, pages 3–42. Princeton University Press, Princeton, N.J.

[57] John R. Koza. *Genetic Programming: On the Programming of Computers by Means of Natural Selection*. MIT Press, 1992.

[58] Blake LeBaron. An evolutionary bootstrap method for selecting dynamic trading strategies. In *Decision Technologies for Computational Finance*, pages 141–160. Amsterdam: Kluwer Academic Publishers, 1998.

[59] Blake LeBaron. Technical trading rule profitability and foreign exchange intervention. *Journal of International Economics*, forthcoming 1998.

[60] Y. LeCun, J. Denker, S. Solla, R. E. Howard, and L. D. Jackel. Optimal brain damage. In D. S. Touretzky, editor, *Advances in Neural Information Processing Systems II*, San Mateo, CA, 1990. Morgan Kauffman.

[61] D. B. Lenat and R. V. Guha. *Building Large Knowledge-Based Systems*. Addison-Wesley: Reading, MA, 1990.

[62] Douglas B. Lenat. Cyc: A large-scale investment in knowledge infrastructure. *Communications of the ACM*, 38(11):33–38, 1995.

[63] M. E. Lesk and G. Salton. Relevance assessments and retrieval system evaluation. *Information Storage and Retrieval*, 4:343–359, 1969.

[64] David D. Lewis and Marc Ringuette. A comparison of two learning algorithms for text categorization. In *Proc. Symposium on Document Analysis and Information Retrieval SDAIR-94*, pages 81–93, 1994.

[65] Michael Lewis. Jonathan Lebed: Stock manipulator, S.E.C. nemesis – and 15. *New York Times Magazine*, February 2001.

[66] Andrew W. Lo. Finance: A selective survey. *Journal of the American Statistical Association*, (95), 2000.

[67] Brad L. Miller and David E. Goldberg. Genetic algorithms, selection schemes, and the varying effects of noise. *Evolutionary Computation*, 4(2):113–131, 1996.

[68] R. G. Miller. *Simulaneous Statistical Inference*. Addison-Wesley, 1966.

[69] Dunja Mladenic. Text-learning and related intelligent agents. *IEEE Expert Special Issue on Applications of Intelligent Information Retrieval*, pages 44–54, July-August 1999 1999.

[70] Chris Neely and Paul Weller. Technical analysis and central bank intervention. Technical report, Federal Reserve Bank of St. Louis, 1997.

[71] Chris Neely, Paul Weller, and Rob Dittmar. Is technical analysis in the foreign exchange market profitable? a genetic programming approach. Technical report, Federal Reserve Bank of St. Louis, 1997.

[72] Kamal Nigam, John Lafferty, and Andrew McCallum. Using maximum entropy for text classification. In *IJCAI-99 Workshop on Machine Learning for Information Filtering*, pages 61–67, 1999.

[73] C. L. Osler. Support for resistance: Technical analysis and intraday exchange rates. *FRBNY Economic Review*, July 2000.

[74] C. L. Osler and P. H. Kevin Chang. Head and shoulders: Not just a flaky pattern. Technical report, 1995.

[75] M.F. Porter. An algorithm for suffix stripping. *Program*, 14(3):130–137, July 1980.

[76] Martin J Pring. *Technical Analysis Explained*. McGraw Hill, 1991.

[77] J. Ross Quinlan. Induction of decision trees. *Machine Learning*, 1:81–106, 1986.

[78] J. Ross Quinlan. *C4.5: Programs for Machine Learning*. San Mateo, CA: Morgan Kaufman, 1993.

[79] Y. Raviv and N. Intrator. Bootstrapping with noise: An effective regularization technique. *Connection Science*, 8(3–4):355–372, 1996.

[80] Ingo Rechenberg. *Evolutionstrategie: Optimierung Technisher Systeme nach Prinzipien des Biologischen Evolution*. Stuttgart: Fromann-Holzboog Verlag, 1973.

[81] Ingo Rechenberg. *Evolutionstrategie '94*. Stuttgart: Fromann-Holzboog Verlag, 1994.

[82] Colin Reeves. Using genetic algorithms with small populations. In Stephanie Forrest, editor, *Proceedings of the Fifth International Conference on Genetic Algorithms*, San Mateo, CA, 1993. Morgan Kaufman.

[83] Ellen Riloff and Wendy Lehnert. Information extraction as a basis for high-precision text classification. *ACM Transactions on Information Systems*, (3):296–333, July 1994.

[84] H. Roberts. Statistical versus clinical prediction of the stock market". Technical report, Center for Research in Security Prices, University of Chicago, 1967.

[85] Gerald Salton. Automatic text analysis. *Science*, 168:335–343, 1970.

[86] Gerald Salton and Christopher Buckley. Term-weighting approaches in automatic text retrieval. *Information Processing and Management*, 24(5):513–523, 1998.

[87] Paul Samuelson. Proofthat properly anticipated prices fluctuate randomly. *Industrial Management Review*, (6):41–49, 1965.

[88] Hans-Paul Schwefel. *Numerische Optimierung von Computer-Modellen mittels der Evolutionsstrategie*. Basel, Stuttgart: Birkhuser, 1977.

[89] George W. Furnas Thomas K. Landauer Scott Deerwester, Susan T. Dumais and Richard Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407, 1990.

[90] Robert E. Shapire. The strength of weak learnability. *Machine Learning*, 5(2):197–227, 1990.

[91] Robert E. Shapire. A brief introduction to boosting. In *Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence*, pages 1–5, 1999.

[92] William Gibson & Bruce Sterling. *The Difference Engine.* Addison-Wesley, 1992.

[93] Ryan Sullivan, Allan Timmermann, and Halbert White. Data-snooping, technical trading rule performance, and the bootstrap. *Journal of Finance*, 54:1647–1692, 1998.

[94] B. Wüthrich V. Cho and J. Zhang. Text processing for classification. *Journal of Computational Intelligence in Finance*, 7(2):6–22, 1999.

[95] Dawn Lawrie Paul Oglivie David Jensen Victor Lavrenko, Matt Schmill and James Allan. Language models for financial news recommendation. In *Proceedings of the Ninth International Conference on Information and Knowledge Management*, pages 389–396, 2000.

[96] Halbert White. A reality check for data snooping. *Econometrica*, 68:1097–1126.

[97] J. Welles Wilder. *New Concepts in Technical Trading Systems*. Addison-Wesley, 1992.

[98] B. Wüthrich, D. Permunetilleke, S. Leung, V. Cho, J. Zhang, and W. Lam. Daily prediction of major stock indices from textual www data. In *Proceedings of the 4th International Conference on Knowledge Discovery and Data Mining - KDD-98*, 1998.

[99] Peter Wysocki. Cheap talk on the web: The determinants of postings on stock message boards. Technical report, University of Michigan Business School.

[100] Yiming Yang. Expert network: Effective and efficient learning from human decisions in text categorizatin and retrieval. In *17th Ann Int ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'94)*, pages 13–22, 1994.

[101] Yiming Yang. An evaluation of statistical approaches to text categorization. *Information Retrieval*, 1(1/2):69–90, 1999.

[102] Yiming Yang and Jan O. Pedersen. A comparative study on feature selection in text categorization. In *Proc. 14th International Conference on Machine Learning*, pages 412–420, 1997.

[103] Y.-C. Zhang. Modeling market mechanisms with evolutionary games. *Europhys. News*, (51), 1998.