2007

On building predictive models with company annual reports

Xin Ying Qiu
University of Iowa

ON BUILDING PREDICTIVE MODELS WITH COMPANY ANNUAL REPORTS

by

Xin Ying Qiu

An Abstract

Of a thesis submitted in partial fulfillment of the
requirements for the Doctor of Philosophy
degree in Business Administration in the
Graduate College of The
University of Iowa

July 2007

Thesis Supervisor: Professor Padmini Srinivasan

ABSTRACT

Text mining and machine learning methodologies have been applied to biomedicine and business domains for new relationship and knowledge discovery. Company annual reports (or 10K filings), as one of the most important mandatory information disclosures, have remained untapped by the text mining and machine learning community. Previous research indicates that the narrative disclosures in company annual reports can be used to assess the company's short-term financial prospects. In this study, we apply text classification methods to 10K filings to systematically assess the predictive potential of company annual reports. We specify our research problem along five dimensions: financial performance indicators, choice of predictions, evaluation criteria, document representation, and experiment design. Different combinations of the choices we made along the five dimensions provide us with different perspectives and insights into the feasibility of using annual reports to predict company future performance. Our results confirm that predictive models can be successfully built using the textual content of annual reports. Mock portfolios constructed with firms predicted by the text-based model are shown to produce positive average stock return. Sub-sample experiments and post-hoc analysis further confirm that the text-based model is able to catch the textual differences among firms with different financial characteristics. We see a rich set of research questions with the promise of further insight in this research area.

Abstract Approved: _____
Thesis Supervisor

_____
Title and Department

_____
Date

ON BUILDING PREDICTIVE MODELS WITH COMPANY ANNUAL REPORTS

by

Xin Ying Qiu

A thesis submitted in partial fulfillment of the
requirements for the Doctor of Philosophy
degree in Business Administration in the
Graduate College of The
University of Iowa

July 2007

Thesis Supervisor: Professor Padmini Srinivasan

Graduate College
The University of Iowa
Iowa City, Iowa

CERTIFICATE OF APPROVAL

PH.D. THESIS

This is to certify that the Ph.D. thesis of

Xin Ying Qiu

has been approved by the Examining Committee
for the thesis requirement for the Doctor of
Philosophy degree in Business Administration
at the July 2007 graduation.

Thesis Committee: Padmini Srinivasan, Thesis Supervisor

Ramji Balakrishnan

Warren Boe

Mort Pincus

Nick Street

To my mother

ACKNOWLEDGMENTS

Finishing this Ph.D. dissertation would not be possible without the tremendous support from many people. I would like to first thank my advisor Professor Srinivasan, for admitting me into the Ph.D. program, providing years of research assistantship, opening a new world of wonderful projects to me, trusting me in finishing my study, and seeing it through. For all these and much more, I am truly grateful.

I would like to thank my committee members, Professor Street, Professor Balakrishnan, Professor Boe, and Professor Pincus, for their insights, support, and encouragement, during both my dissertation preparation and job seeking times.

I would like to thank my friends and ISOR lab mates, Kaan Ataman, Aditya Sehgal, Kyuseop Kwak, Thaddeus Sim, Chihlin Chi, Brian Almquist, and Chris Choi, whose warm friendship and genuine support fill my Ph.D. experience with fun and happiness.

I would like to thank my extended family, Aunt Xuezhang, Uncle Jiyao, Aunt Meizhang, Aunt Yuzhang, and Uncle Jijin, for always being there for me, cheering for me, and being more excited and enthusiastic about my progress than even myself.

Most importantly, no words can express my deepest gratitude to my mother, for her profound wisdom and the most relentless support. I dedicate this dissertation to my mother, Qianzhang Qiu, with all my heart.

ABSTRACT

Text mining and machine learning methodologies have been applied to biomedicine and business domains for new relationship and knowledge discovery. Company annual reports (or 10K filings), as one of the most important mandatory information disclosures, have remained untapped by the text mining and machine learning community. Previous research indicates that the narrative disclosures in company annual reports can be used to assess the company's short-term financial prospects. In this study, we apply text classification methods to 10K filings to systematically assess the predictive potential of company annual reports. We specify our research problem along five dimensions: financial performance indicators, choice of predictions, evaluation criteria, document representation, and experiment design. Different combinations of the choices we made along the five dimensions provide us with different perspectives and insights into the feasibility of using annual reports to predict company future performance. Our results confirm that predictive models can be successfully built using the textual content of annual reports. Mock portfolios constructed with firms predicted by the text-based model are shown to produce positive average stock return. Sub-sample experiments and post-hoc analysis further confirm that the text-based model is able to catch the textual differences among firms with different financial characteristics. We see a rich set of research questions with the promise of further insight in this research area.

TABLE OF CONTENTS

LIST OF TABLES

Table

LIST OF FIGURES

Figure

CHAPTER I
INTRODUCTION

1.1 Background

Publicly traded companies on the stock exchange market are required by the Securities and Exchange Commission (SEC) to regularly disclose information to the marketplace through mandatory reports. These reports are filed to the SEC's EDGAR database[1] and freely available to the public. Annual reports are one of the most important and valuable information disclosures from the perspectives of financial analysts, investors, and regulators. The mandatory disclosures in annual reports include performance-related information, such as reasons for price and sales changes, reasons for sales revenue and cost changes, planned expenditures, known trends, future liquidity position, and view of past year performance and future prospects. Readers of annual reports intuitively expect to deduce insights from the disclosures about the company's current and/or future performance, strategies and profitability. Previous research has shown that the narrative discussions in the annual reports are important when assessing firm value. For example, Rogers and Grant [84] found that the Management Discussion & Analysis (MD&A) section, a major narrative section in annual reports, constituted the largest proportion of information cited by financial analysts. A survey conducted by Association for Investment Management and Research (AIMR) [2] in 2000 found that the management's discussion of corporate performance was an extremely important factor to analysts when assessing firm value.

These narrative disclosures serve a similar purpose to that of the numerical financial data which is to disseminate information useful for maintaining market efficiency. Firms' financial data are available to the general public through the firms'

[1] http://www.sec.gov/edgar.shtml

disclosure of financial statements such as balance sheet, income statement, and cash flow statement. Investors can also follow financial market's activity through the media such as Yahoo Finance and the Wall Street Journal. Academic and research institutions can access Wharton Research Data Service for a wide variety of financial, economic, and market data. Interestingly, previous research on forecasting company's future performance has mainly focused on the historical numerical data. The goal of the predictive models built with financial data is generally to identify prominent financial ratios with good predictive power, or good classification algorithms such as neural networks [100, 112]. For example, Saad et al. [87] compared three neural network approaches for predicting short term stock trends based on historical pricing data. They found all three methods to be feasible with each offering distinct advantages.

Annual reports have been studied as a marketing and communication tool that corporations use to convey an image or messages to its stakeholders [39]. More recent studies on the relationship between the reports and firm performance have focused on special sections of the reports, such as the chairman's statement [95], management discussion and analysis (MD&A) [10], president's letter [1] and on the general writing style and readability [97]. The methods these studies employ are generally semi-automatic, including content analysis, readability measurements, manual annotation and categorization, linear discriminant analysis, logit model and other statistical analysis. Their main contributions are that the researchers were able to identify special features of the writing in general, or special disclosure variables, that correlate with certain performance ratio or general profitability. For example, Subramanian et al. [97] found that good performers used 'strong' writing in their reports while poor performers' reports contained significantly more jargon or modifiers and were hard to read. Smith et al. [95] identified thematic keywords from chairman's statements and generated discriminant functions to predict company failure. Bryan [10] showed that

the discussion of future operations and planned capital expenditures were associated with one-period-ahead changes in sales, earnings per share, and capital expenditure. Kohut et al. [54] studied president's letters in annual reports and suggested that poor performing firms tend to emphasize future opportunities over poor past financial performance as a communication strategy. These studies emanate from the intuitive recognition of a link between the textual report content and corporate performance. Their findings suggest that combining the textual analysis of the reports with the quantitative data in the financial statements may assist in predicting company performance and even specific outcomes such as failure and bankruptcy.

We can also find interesting research in the accounting domain which considers "non-accounting" information of a firm into the projection on the firm's future value. The Ohlson model formulated firm value as a linear function of current book value and future abnormal earnings which does not rely on observed or forecasted dividends [7, 26, 65, 72]. One important assumption in the Ohlson model is to define the stochastic process for abnormal earnings and nonaccounting information v_t as

$$x^a_{t+1} = w * x^a_t + v_t + \varepsilon_{1t+1}$$

$$v_{t+1} = \gamma * v_t + \varepsilon_{2t+1}$$

where x^a_{t+1} is the abnormal earnings at time $t+1$; v_t is the partially forecastable nonaccounting information at time t; ε_{1t+1} is the completely nonforecastable nonaccounting information at time $t+1$; and w and γ are know parameters. The forecastable nonaccounting information v_t at time t provides a shock to the abnormal earnings in time $t+1$ in an autoregressive process. We speculate that in the company annual reports there exists this forecastable nonaccounting information v_t which will contribute to the company's future value.

1.2 Motivation

Text classification and other text mining techniques have been used to build knowledge discovery applications in different domains. The goal of these applications is in general to discover hidden patterns, implicit relations, or new ideas from the vast amounts of document data. The document collections from different domains share a feature, that is they are unstructured or at best semi-structured. Examples include biomedical bibliography databases, news stories, and the ever-growing set of web pages. Because of the lack of structure and the many natural language features, it is both challenging and interesting as a research problem to extract and discover novel and implicit knowledge from document collections.

Several text mining systems have been built especially using document records in MEDLINE as a knowledge source. Perez-Iratxeta et al. [74] provide a system that could rank the candidate genes potentially associated with diseases. The connection between gene and disease is generated by exploring document co-occurrence among disease terms, chemical terms, and Gene Ontology [2] annotation terms. Wilkinson and Huberman's system [107] aims at identifying a community of genes that share common co-occurrence relations with certain diseases in MEDLINE. Other general purpose systems such as Manjal [96] provide options for mining MEDLINE to discover novel connections between pairs of topics where the topic type is unrestricted. We also see text mining research applied increasingly in the web domain, such as for extracting symbolic knowledge [17], generating new research connections from web pages [34], and identifying online communities [55]. For example, Gordon et al. [34] demonstrated that literature-based discovery could be applied to the web for finding new research problems in areas other than medicine.

One standard machine-learning approach in data mining from structured data

[2] http://www.geneontology.org/

is classification. When applied to documents, the goal of text classification is to automatically assign a predefined class (or topical category) label to a document. In biomedicine, text classification has been used in various ways such as to extract sentences from documents [16] and to explore gene document annotation with ontology terms [78]. For example, using statistical text classification, Craven [16] identified specific semantic relations between protein and subcellular structures from sentences extracted with the classifiers. Rice et al. [81] used term-based support vector machine classification to identify evidential passages that support the assignment of Gene Ontology terms to human proteins. These previous studies illustrate the variety of text classification-based applications in biomedicine.

In the business domain, which is the focus of this paper, text mining and text classification have been applied to business news stories and to information collected from the web. For example, Berstein et al. [8] explored relationships among firms and industries using information extracted from news stories. Online product reviews, opinions and discussions have also been studied by several researchers for opinion extraction and classification [18, 33, 67]. One important data source in the business domain is the mandatory information disclosure from companies by way of annual reports and quarterly reports. Interestingly, this data source remains largely untapped by the machine learning and text mining community.

Researchers rarely utilize the textual content of annual reports to build predictive models. This is despite findings that these reports have the potential to serve as indicators of company future prospects. The most relevant work in this direction is that of Kloptchenko et al. [51, 52] and Visa et al. [103]. In Visa et al. [103], paragraphs in annual reports were projected into paragraph maps and histograms were generated. Their goal was to discriminate paragraphs with similar words but distinct content. They were able to categorize paragraphs and cluster paragraphs with similar content. In the Kloptchenko et al. 2002 study [51], company quarterly

reports and corresponding financial ratios were clustered separately with prototype matching clustering and Self Organization Map (SOM) clustering respectively. Although the two sets of clusters did not coincide, the authors found that changes in textual reports tended to occur ahead of changes in financial performance. Li studied how annual reports' fog index correlates with earnings prospects and persistence [61]. In another research [62], Li examined the use of words of "risk" and "uncertain" in annual reports and their association with future earnings and stock return. Several other studies [19, 38] explored the language features in earnings press releases such as document length, tone, textual complexity, and optimistic/perssimistic languages. Other than these studies relevant to narrative disclosure, to the extent of our knowledge, no others exploit the text of annual reports for automatic predictions.

1.3 Research Questions

Set in this background literature, we see a well-justified opportunity to see if we can build classification models to predict company financial performance from annual reports. The goal of our research is to assess whether these annual reports can be used to predict the *change* in company financial performance. Note that, 'change' in performance is a temporal notion measured by comparing performance over different years. We propose to address this research goal with predictive models built using a text classification approach.

Our first step in achieving our research goal is problem specification. We find that we need to make decisions along several critical research dimensions, such as the choice of financial performance indicators, the choice of evaluation criteria, and experimental design. In other words, we need to define the shape, i.e. the parameters, of the problem being addressed. Our final predictive system reflects one reasonable combination of the alternatives that we could have used. By discussing and exploring these alternative options (in the next chapter), we achieve an additional goal in this

dissertation: to illustrate how classification procedures deriving from real world goals may be shaped in different ways. The challenge is to shape it in a way that is both realistic as well as approachable computationally. More specifically, we hope to contribute to research that considers real-world applications, especially in the business and finance areas.

We observe that the typical research paper in text classification involves pre-specified goals, domains, collections, class definitions, document formats, and sometimes even data distribution features. For example, several 'standard' text collections representing different domains have been used by researchers for testing classification algorithms. The 20 newsgroups collection that has generated much research (e.g. [45, 71, 69]) contains 20,000 articles from 20 different UseNet discussion groups each about a different topic. Some topics are closely related while others are highly unrelated. Also some newsgroup documents are known to belong to more than one group. This data set in effect defines a multi-class (topic) text classification problem. Another widely used data set is Reuters 21578 which consists of 12,902 news articles annotated with 90 topics from Reuters newswire. Researchers have used this collection extensively to evaluate a variety of information retrieval and machine learning techniques [32, 49, 58]. In biomedicine, the OHSUMED collection [40] is composed of 348,566 records from MEDLINE. Here classes are defined by MeSH (Medical Subject Heading) terms. Several papers have studied classification methods with this data set as for example papers exploring hierarchical algorithms (e.g. [50, 86]) This data set was also the basis of a TREC track on filtering strategies[3].

Our research goal, motivated by real world objectives in business and finance, is to predict change in company performance from their reports. We find that the universe of companies to consider, report types, company performance measures, notions of change in performance, baselines etc. need to be defined. In other words,

[3]http://trec.nist.gov/data/t9_filtering.html

we first need to identify key dimensions that relate to our research problem and select between alternatives within each dimension. We do this via the key questions presented next.

1. How should one assess a firm's performance?
2. What should one forecast?
3. How do we evaluate the performance of our predictive models?
4. How do we represent documents?
5. How do we design our experiments to ensure the validity of the results?

To summarize, we find that each of the above five aspects defines a dimension along which we need to make choices. A different set of choices along these five dimensions may reflect different assumptions and possibly result in different predictive models. These may in turn induce different analysis from the accounting and financial perspectives. In the next Chapter, we discuss the choices that we have to make. We hope to outline the potential for further studies, while also illustrating the complexity of working with research problems motivated by real world applications and goals.

CHAPTER II
RESEARCH APPROACHES AND LITERATURE REVIEW

In this chapter, we address the five research questions presented above by examining the choices available along each of these five research dimensions: financial performance indicators, choice of prediction, evaluation criteria, document representation, and experiment design.

2.1 Financial Performance Indicators

In the accounting and financial research domains, there are many variables and criteria for measuring a firm's performance. These include accounting measures such as operating earnings, net income [97], and current ratio, and market response measures such as stock return [87]. Kohut and Segars [54] studied the content of president's letters in annual reports in relation to firm performance. They used return on equity (ROE) to rank and select high-performing and low-performing firms. ROE is a ratio of net income over shareholder's equity. It measures the earning power of a share owner's equity. Zhang et al. [112] compared neural network models and a variety of linear statistical models in forecasting Earnings per Share (EPS). EPS measures the amount of earnings per share of stock. Smith and Taffler [95] employed a UK accounting ratio based z-score to define company failure. Kloptchenko et al. [51] selected 7 ratios to characterize a firm's financial performance. These include three profitability ratios, one liquidity ratio, two solvency ratios, and one efficiency ratio. From the accounting research perspective, how to measure a firm's performance is a research question about accounting choices and their consequences [27]. In this dissertation, we consider both the accounting measure (i.e. return on equity (ROE) and earnings per share (EPS)) and the market response measure (i.e. stock return).

We denote year $t + 1$ to be the year for which we want to predict a company's

change of performance, and year t to be the given year of which the annual reports are used for prediction. We design three financial measures for assessing a firm's performance using two accounting measures for operation (i.e. ROE and EPS) and one market response measure (i.e. size-adjusted cumulative return (SAR)).

Return on Equity (ROE) is the ratio of net income over shareholders' equity. It shows how much income was earned for every dollar invested by owners. We denote year $t+1$ to be the year for which we want to predict a firm's performance relative to its performance in year t. Year t is also the year of which the annual reports are used for prediction. We have available to us the firm f_i's ROE ratio value for year t (denoted as $ROE_t^{f_i}$) and for year $t+1$ (denoted as $ROE_{(t,t+1)}^{f_i}$). We define a growth rate of ROE at year $t+1$ relative to year t as:

$$\Delta ROE_{(t,t+1)}^{f_i} = (ROE_{t+1}^{f_i} - ROE_t^{f_i})/|ROE_t^{f_i}|$$

$$where$$

$$ROE_{t+1}^{f_i} = NetIncome_{t+1}^{f_i}/Equity_t^{f_i}$$

Similarly, we define a growth rate in EPS as a second option to assess a firm's operating performance. Earnings per Share (EPS) is total earnings divided by the number of shares outstanding. This ratio shows how much of a firm's earnings are available for distribution as dividends to each share of common stock. It helps the investors decide on the potential of future dividends and the firm's ability to finance its growth internally. We denote a firm's EPS in year t as $EPS_t^{f_i}$, and the growth rate of EPS at year $t+1$ relative to year t is defined as:

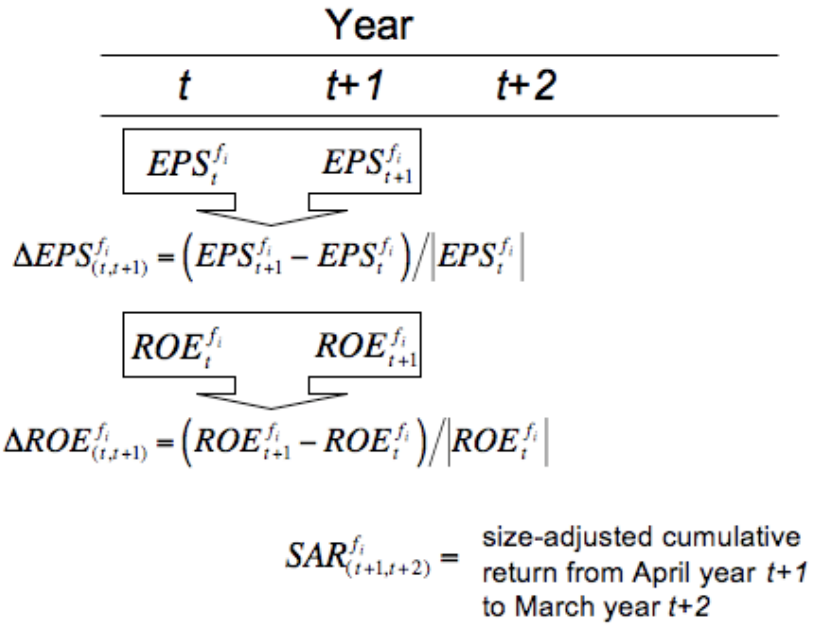
$$\Delta EPS_{(t,t+1)}^{f_i} = (EPS_{t+1}^{f_i} - EPS_t^{f_i})/|EPS_t^{f_i}|$$

A third performance measure is SAR, the size-adjusted cumulative return, which is inherently temporal. $SAR_{(t+1,t+2)}^{f_i}$ is the cumulative return from April of year $t+1$ to March of year $t+2$, minus the return over the same period, for the corresponding market decile. The decile adjustment, which removes the average return for all firms with similar size, accounts for the market risk from investing in the sample firm. We

measure the return for a year to be consistent with our use of annual data. The SAR tells us the incremental return we may expect in a year if we invest in the firm at the end of March of year $t+1$, based on the predictions made for year $t+1$.

Figure 2.1 demonstrates the constructions of these three financial performance measures. The definitions of ROE and EPS growth rates at year $t+1$ relative to year t indicate both the direction and the magnitude of a firm's changes in ROE and EPS ratio. SAR measure reflects both the company's profitability and market response information. Thus, ranking all firms by these three performance measures of all the firms in a certain year allows us to compare firms in terms of the direction and the magnitude of changes in their operating performance.

Figure 2.1: **Timeline for Constructing $\Delta EPS^{f_i}_{(t,t+1)}$, $\Delta ROE^{f_i}_{(t,t+1)}$ and $SAR^{f_i}_{(t+1,t+2)}$ for Predicting Year $t+1$**



2.2 Choice of Prediction

After we select a performance measure, we also need to decide on what aspect of this measure to predict. For example, if we decide to evaluate a firm's performance with its EPS, should we forecast its EPS value, or the firm's rank with respect to other firms in terms of EPS value, or some category decided by EPS value? This decision on what to forecast implies a choice among a real value prediction, or a categorization of firm performance into specific classes, or firm ranking (which is itself a special case of categorization).

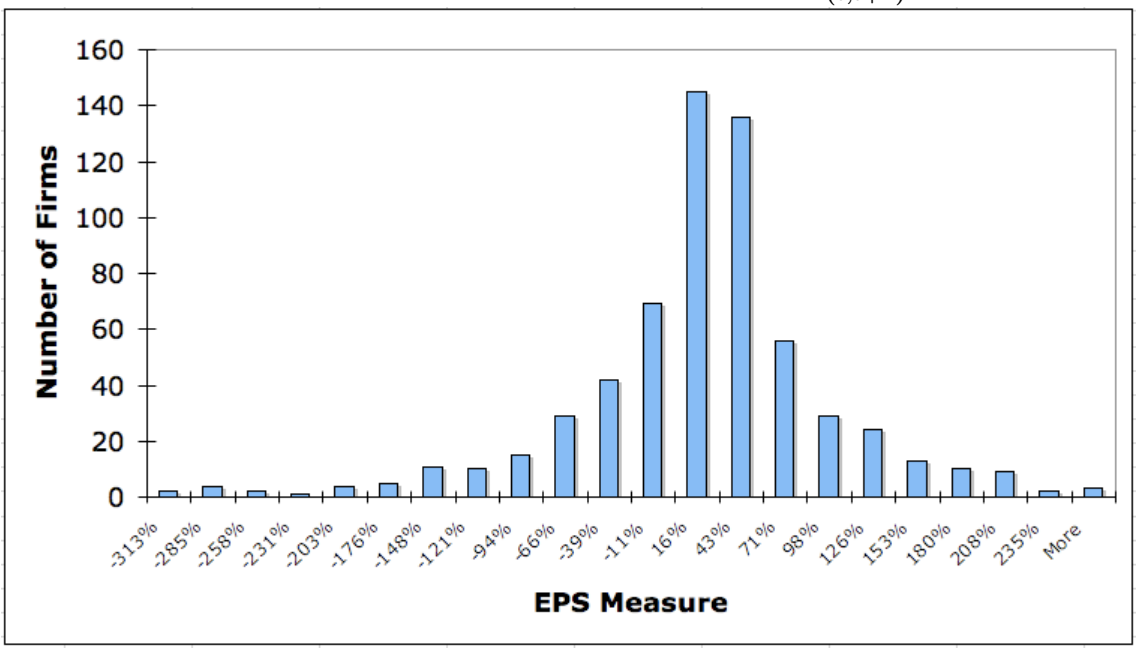
Within the categorization framework, we also face the question of how to define the classes for firm performance. How many classes should we have? Should we go to the extreme and place each firm in its own class in which case we have a 'ranking' problem? Also, what are the appropriate criteria for categorizing firm performance into different classes? These questions certainly do not have single answers, but they definitely require well-reasoned decisions.

Since we are at the initial stage where we seek to explore the feasibility of building predictive models from the textual content of annual reports, we would prefer to start with a methodology that forecasts coarse-grained results. In other words, we would prefer to forecast a performance category rather than a real value of a performance measure. We then face the question of how many performance classes we should define and what criteria we should use to define the classes.

We begin by exploring the characteristics of the data distribution, i.e. the distribution of firms' performance based on each measure. Figure 2.2 and 2.3 are examples of the data distribution. Here we see that firm performances measured with $\Delta EPS^{f_i}_{(t,t+1)}$ for year 2002 and 2003 follow a near normal distribution. This observation holds in general for data of other years as well as for $\Delta ROE^{f_i}_{(t,t+1)}$ and

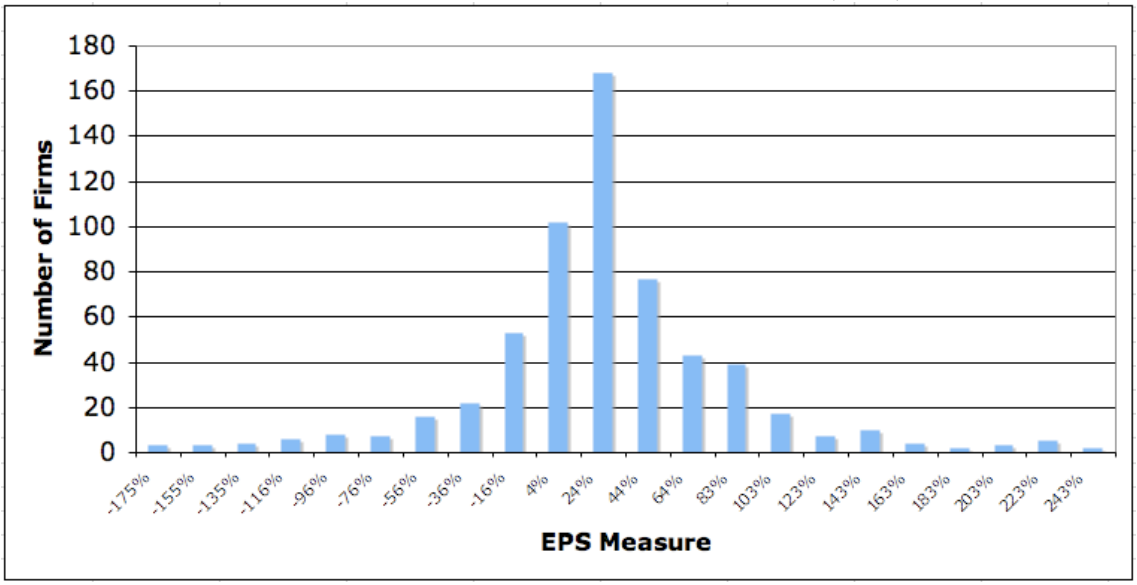
$SAR^{f_i}_{(t+1,t+2)}$. Appendix A provides more examples on data distribution.[1] The observation on near normal distribution allows us to decide on a minimum of three classes of performance. (Note that as the number of classes is increased, the problem approaches to a strict ranking problem.) For the data of each firm in a year corresponding to a measure, we categorize the top 25% firms as out-performing, the middle 50% as average-performing, and the bottom 25% as under-performing. Given these three known classes of firms, our goal is to predict the correct classes using models built with reports from the previous year. More generally, we apply a three-class document classification approach to predict the future performance of firms.

Figure 2.2: **Histogram for EPS Measure ($\Delta EPS^{f_i}_{(t,t+1)}$) for 2002**



[1]We later found that for $SAR^{f_i}_{(t+1,t+2)}$ measure, not all years' data follow a near normal distribution, i.e. years 1999, 2000, and 2001. The majority year/measure combination of data is near normal, and we prefer a consistent class definition for all year/measure combination of data. An alternate way to handle the inconsistency in data distribution is to formulate the prediction problem as a ranking problem.

Figure 2.3: **Histogram for EPS Measure ($\Delta EPS^{f_i}_{(t,t+1)}$) for 2003**



2.3 Evaluation Criteria

2.3.1 Measures and Baselines

The effectiveness of a classifier is defined as its capability to make the right categorization decisions. The standard effectiveness measures include: precision, recall, accuracy, error rate, break-even point, E-measure, F-measure, macro-averaging, and micro-averaging. We would like to evaluate our predictive models using standard measures from text classification research. We would also like to compare model performance with meaningful benchmarks. Our predictive models will give each firm/year data point a label of out-performing, average-performing, and under-performing. One standard evaluation measure is accuracy rate (or 1 − error rate) which is defined as the proportion of correctly classified samples out of all samples. This accuracy rate of the predictive model will be compared with two baselines generated by: 1) majority vote, and 2) analyst forecasts.

Majority vote baseline: Since we decide to categorize our sample set for each year into 25% out-performing, 50% average-performing, and 25% under-performing

subsets, the majority vote decision with any of the three financial performance measure is to assign all firms' performance as average-performing. This is our first baseline with default 50% accuracy. Thus we would like to see if our predictive model's accuracy will be higher than 50%.

Analysts forecast baseline: Analysts also study firm performance and make EPS forecasts and stock recommendations. Analysts use a wide variety of information sources (including annual reports) and techniques. For each firm/year, the aggregate consensus forecasts on EPS from analysts can also be evaluated against the firm's actual EPS. Thus we can also calculate the accuracy of analysts' forecast. By comparing our text classification model's accuracy with the analysts forecast accuracy, we may be able to better understand the value of our textual predictive models. This will also inform us about the extent to which we may automate the prediction decisions. We denote the analysts' EPS forecast in year t as $AnalystsEPS_t^{f_i}$, a firm's actual EPS in year t as $ActualEPS_t^{f_i}$. The analysts' predicted growth rate of EPS at year $t+1$ relative to year t is defined as:

$$\Delta AnalystsEPS_{(t,t+1)}^{f_i} = (AnalystsEPS_{t+1}^{f_i} - ActualEPS_t^{f_i})/|ActualEPS_t^{f_i}|$$

We then rank the firm/year data according to $\Delta AnalystsEPS_{(t,t+1)}^{f_i}$ and categorize the top 25% as analysts' forecasted out-performing, middle 50% as analysts' forecasted average-performing, and bottom 25% as under-performing groups. Then we are able to calculate the accuracy of analysts forecast and compare this accuracy with that of the predictive models.

We also use analysts' stock recommendations as another evaluation baseline for our predictive model with $SAR_{(t+1,t+2)}^{f_i}$. We use the consensus mean recommendations for year t+1 outstanding as of the end of the fourth month after year t's fiscal year end. These are scaled ranking from 1 to 5 with 1 referring to strong buy up to 5 referring to strong sell. These can also be grouped into three recommendation classes based on the distribution of the recommendation scores by the analysts. We denote the class

of analysts recommendation for company f_i in year $t+1$ as $AnalystsClassStock_{t+1}^{f_i}$. All companies' $AnalystClassStock_{t+1}^{f_i}$ and their true classes based on $SAR_{(t+1,t+2)}^{f_i}$ can be used to generate a predictive accuracy of analysts forecast for year $t+1$. This accuracy will be another evaluation bench for our text classification model.

Prediction Costs: The accuracy rate does not take into account the cost of making wrong decisions. Observe that the three classes of out-performing, average-performing and under-performing are essentially a coarse ranking of the companies. This kind of ranking is likely to influence investors' decision on investment. The cost of making wrong classifications is important to consider in the setting of predicting firms' future performance. In addition, the classification models to be discussed in Section 2.5.3 might differ in the types of errors they make. Thus we decide to design a count of cost of errors as another evaluation measure to compare the three classification models.

Since our majority vote baseline assigns all firms to be "average", we are more interested in our model's predictive performance for out-performing and under-performing firms. There are two types of errors a model may make in predicting out-performing or under-performing. One is to predict out-performing (or under-performing) as under-performing (or out-performing). The other is to predict out-performing (or under-performing) as average. The former error should have higher cost than the latter since the former error is a misclassification over 2 levels while the latter is over 1 level. Formally, $Cost$ of a predictive model is defined as:

$$Cost = \begin{cases} 2 \times C, & if \begin{cases} TrueClass = \textit{out-perform and } PredictedClass = \textit{under-perform or} \\ TrueClass = \textit{under-perform and } PredictedClass = \textit{out-perform} \end{cases} \\ C, & otherwise \end{cases}$$

2.3.2 Mock Portfolio

Aside from comparing the model's accuracy with the majority vote baseline, and with the analysts' forecast accuracy, we could also use "utility score". Schapire et al. [91] presented three utility measurements in which the last defined certain rewards for correct classification and punishment for incorrect classification. These intuitively resemble the risk and return generated from a portfolio that is used for trading stocks of out-performing and under-performing firms. The profitability of the portfolio based on model predictions can serve as an evaluation of the model's operation value.

To illustrate the mechanism of a mock portfolio, we assume that the year to predict is our current year is year $t+1$ and our transaction date is March 31st year $t+1$. Therefore we have available to us the annual reports of year t[2] but no financial performance data for current year $t+1$ is available to build a model. We assume that we want to construct a portfolio in year $t+1$ involving 10 million dollars and hope to make a profit in year $t+2$. Our portfolio will be constructed in such a way that we will keep the stocks of companies that we believe are going to perform better in year $t+2$ and sell the stocks of those that will perform worse in year $t+2$. Thus we need to apply a text classification model built with year $t-1$ document and year t financial performance to year t's documents, and predict the companies' SAR return for the period of March year $t+1$ to March year $t+1$.

More specifically, we use year $t-1$ documents paired with year t financial performance data to build a classifier. We apply this classifier to year t's documents and predict companies' performance in year $t+1$. Suppose our model identifies 5 companies to perform better in year $t+1$ and 10 to perform worse in year $t+1$ and the rest of companies's performance to stay the same. We sell the to-perform-worse companies' stocks at a total value of 10 million dollars and buy the to-perform-better

[2]The annual report for a given year t is required to be filed and available to the public by March 31st in year $t+1$.

companies' stocks with a total value of 10 million dollars. In both the buying and selling transactions, we will allocate equal values of stocks among the companies[3]. That is to say we will sell 1 million dollars of stocks for each of the 10 to-perform-worse companies, and buy 2 million dollars of stocks from each of the 5 to-perform-better companies.

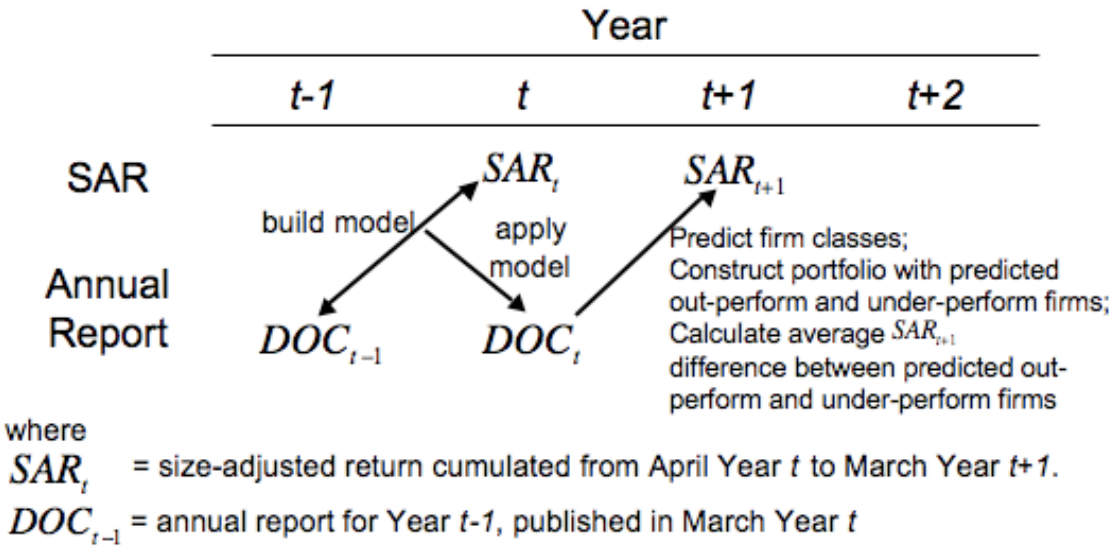
Since we also have available to us the stock price information from all companies on March 31st in current year $t+1$, we will thus construct a portfolio with a certain number of stocks from the 5 to-perform-better companies. On March 31st in year $t+2$, we will sell the stocks of the 5 to-perform-better companies and buy the stocks of the 10 to-perform-worse companies. If our prediction made in year $t+1$ is correct, this portfolio transaction will generate non-negative profit. Thus the profitability of this mock portfolio will be another way to evaluate the model's predictive accuracy. We can vary the current year $t+1$ systematically to generate multiple portfolios for different years and evaluate each portfolio's profitability. This analysis will shed light on the implementability of our text classification model. The model's portfolio return could also be compared with that of the analysts based on analysts stock recommendation as another evaluation method. Figure 4.21 illustrates the process to construct a portfolio for year $t+1$ using the predictions given by documents of year t after applying the model built with year $t-1$ documents and year t SARs to the year t documents.

2.3.3 Sub-sample Experiments

In the accounting field, there are cross-sectional factors that are known to influence the firms' information environment as well as the portfolio return [9]. These include the size of the firm (as measured with total assets), profit-making or loss (as

[3]Alternatively, we could allocate stock values for each company proportional to the company's share of the capital market

Figure 2.4: **Portfolio Construction**



measured with EPS), and growth firms or value firms (as measured with market-to-book ratio). We would like to see if our models and the reports could predict the differences in portfolio return among firms of different sizes, profits, and market-to-book ratios. Thus, we design a sub-sample portfolio experiment that applies the mock portfolio design to sub sets of firms.

More specifically, we have three approaches to partition firms into sub sets: by size, by EPS, or by market-to-book ratio. By size, for each year, firms are partitioned into large firms with total assets $\geq$ median vs small firms total assets $<$ median. By EPS, firms are partitioned to profitable firms with EPS ≥ 0 or loss firms with EPS < 0. By market-to-book ratio, there are glamour firms with market-to-book ratio $\geq$ median and value firms with market-to-book ratio $<$ median. Then, the mock portfolio experiment as described in Section 2.3.2 is run on each of these three partitions with each partition giving two sub sets. The goal is to assess whether, if we control firms' size, profitability and growth potential, we will still be able to produce excess return and if so, how the excess return is related to the control factors.

2.4 Document Representation

In information retrieval research, documents are typically represented as vectors of weighted terms. This is generally referred to as the vector representation model. There are three aspects to consider when building term based vector representation model: 1) how to define a term; 2) whether to use the full set of terms or a selected subset and if the latter, how to select a subset; 3) how to weight the terms in the vector model. We will consider each of these three aspects next.

2.4.1 Term/Feature Definition

The most widely-used "bag of words" approach is to use all the terms in the training corpus regardless of the order of the terms, to represent document vectors. This appears to be the default standard in text classification [90, 88]. Functional or connective words are considered as stop words and are generally removed since they are assumed to have no information content. Stemming is sometimes performed to remove the suffixes and to map words to their morphological forms [76]. This helps, to a limited extent, to keep the remaining feature terms to be statistically independent from each other. Researchers have explored other more complex textual representations such as n-grams, phrases, syntactic phrases, terms clustered as metatfeatures, or factors generated from original term spaces by way of latent semantic indexing. Several researchers have explored using syntactic phrases to represent documents [4, 24] [29, 57], but the results were not promising. "N-gram" has been used to denote either n-character terms or window of n words. Using n-grams to model documents and perform text classification has generated mixed results in different domains [11, 73, 99]. Due to inconsistencies in the corpus used for experiments and the learning algorithms employed, it is difficult to conclude if n-character term indexing and n-word window sequence indexing will perform better than single-word indexing with different

classifiers. Lewis in [57] studied representing documents with terms that were clustered with a nearest neighbor algorithm. His findings suggested that coarse-grained metafeatures from clustering failed to capture the semantics of the content. Furthermore, he suggested that phrases with or without syntactic restriction were not good content indicators. Similarly Li and Jain [63] showed that term-clustering produced only marginal improvement in classification accuracy. Bekkerman et al. [6] employed an approach to combine feature distributional clustering with support vector machine (SVM) classifiers and observed significant dependency of the results on the datasets. Recent research by Moschitti [68] suggests that the elementary textual representation based on words applied to SVM models is very effective in text classification. More complex linguistic features such as part-of-speech information and word senses did not contribute to the prediction accuracy of SVMs.

Latent semantic indexing (LSI) [20] is arguably a successful feature generation method that is aimed to reduce the term space dimension and capture the latent structure of the documents. LSI applies matrix decomposition to the term-by-document matrix to produce a large number of orthogonal factors. Singular value factors are truncated to generate a new smaller term-document matrix that approximates the original document semantic structure. The idea behind LSI is to find semantically related terms and define a smaller set of new features as functions of the original terms. The original larger dimensions are projected onto a lower-dimension feature space. LSI has proved to be effective in information retrieval [20], document routing [22], information filtering [28], and spam filtering [31]. However, when the training document set is large as for example in Text Retrieval Conference (TREC) collections, computing hundreds of LSI factors is very expensive. The new reduced dimension alone lacks direct interpretation. Moreover, research [94] shows that when documents are easily represented by a small set of content-representative terms, the basic term indexing performs better than LSI. When documents contain a large number of terms

that all semantically contribute to the labeling of the documents, LSI seems to be superior in capturing the implicit meaning of the documents.

2.4.2 Term/Feature Selection

Since the term space generated from our 10K report collection is very likely of extremely high dimension, we will need to reduce the term space and generate smaller global or local[4] vocabularies. The benefits of such a reduced term space include better generalization ability of the model, saving of computing time, and possibly better interpretation and understanding of the predictive features. Term selection refers to those methods that select a subset of the original term space to represent and index the documents. Most term selection methods either compute statistical feature scores to select terms or apply feature selection algorithms from machine learning to search for better textual features. Feature selection aims at achieving the optimal classification performance as measured for example with error rate. In addition, the statistical properties of the selected feature set could be used to augment the semantic characteristic of the features to provide better understanding of a class of documents. Feature selection approaches can be categorized into "wrapper" [53] and "filter" [48] methods with the latter being far more computationally efficient.

The main term selection functions that have been studied in text classification are: document frequency [111], information gain [56], mutual information [98], chi-square [110], correlation-coefficient [85], relevancy score [106], odds ratio [66], and simplified chi-square [30]. Yang and Pedersen [111] showed that document frequency, chi-square, and information gain performed very well and similarly to each other in text categorization. Ng [70] proposed a correlation coefficient as the square root of chi-square, and showed that it performed better than chi-square on a standard Reuters[5]

[4]Global vocabulary implies that the terms from all the documents are used for training and testing classifiers. Local vocabulary is class-specific vocabulary designated to represent documents of a certain class.

[5]Reuters collection is a set of newswire stories classified under categories related to economics.

22173 collection (one of the datasets used by several researchers [111]). Galavotti et al. [30] simplified Ng's correlation coefficient by eliminating the denominator and a redundant factor in the numerator. The reduced function called "simplified chi-square" was shown to perform better than correlation coefficient and chi-square. Their experimentation was done systematically on Reuters-21578. Although LSI feature extraction approach as discussed previously performs better than chi-square term selection method [94], it is yet to be determined how LSI compares with simplified chi-square or correlation coefficient.

In a pilot study [79], we experimented with document frequency thresholding, chi-square and our own proposed z-test feature selection method. The document frequency was used for global dimension reduction. The chi-square and z-test methods were applied for local-dimension reduction such that chi-square produced class-specific and mutually exclusive vocabularies, while z-test produced class-specific vocabularies with overlapping terms. Our results in the annual reports classification experiment showed that, with SVM linear kernel classifier, document frequency is efficient in dramatically reducing term space while maintaining the same classification accuracy. However, no class-specific feature information can be inferred with the chi-square methods. Z-test methods performed well in classifying positive and neutral classes of documents while chi-square methods achieve similar performance as SVM-only method in classifying positive class documents only.

2.4.3 Term/Feature Weighting

Researchers have used many ways to calculate term weights in document vectors, such as a boolean value representing the absence or presence of the term, term frequency (number of times a given term appears in a document), document frequency

It is publicly available as benchmark collection for the purpose of text categorization or document classification experiments.

(number of documents containing a given term), certain combinations of term frequency and document frequency, or entropy weighting [23]. TF×IDF[6] is the most commonly used weighting scheme for estimating the usefulness of a given term as a descriptor of a document. Its implication is that the best descriptive terms of a given document are those that occur very often in this document but not much in the other documents. For each of the factors in TF×IDF weighting scheme (i.e. term frequency, document frequency, document length normalization), we have many alternatives. In our preliminary experiments, two TF×IDF weighting schemes, i.e. "ltc" and "atc" have been experimented with to represent company annual reports in conjunction with Support Vector Machine (SVM) classifiers. They are computed as follows:

$$ltc = \frac{w_i}{\sqrt{\sum_i w_i^2}}$$
$$where \quad w_i = (\ln(tf) + 1.0) \times \ln(\frac{N}{n})$$
$$atc = \frac{w_i}{\sqrt{\sum_i w_i^2}}$$
$$where \quad w_i = (0.5 + 0.5 \times \frac{tf}{maxtf}) \times \ln(\frac{N}{n})$$

where tf is raw term frequency; $maxtf$ is the frequency of most frequent term in the document; N is the total number of documents in the collection; n is the number of documents containing the given term; w_i is the weight of term i (which is defined differently for "ltc" and "atc" as in the above equations).

In our preliminary study [79], these two weighting schemes produced similar results. In this dissertation, we propose to use "bag of words" document vectors with "atc" (i.e. TF×IDF, cosine normalized) weights as our baseline document representation model.

[6] A product of term frequency factor, document frequency factor, and a document length normalization.

2.5 Experiment Design

Text classification procedure involves two steps: training a model on a data set, and testing the model on a separate data set that was unseen during training. The performance of a model on the training set will not be a good indicator for the model's performance in real world application. Therefore, to test a model's performance, we need an independent testing set that preferably have no overlap with the training set. However, the data available to us is always a sample of the universe. With the limitation on data, our experiment designs need to facilitate assessing the validity and the statistical significance of our results. Cross validation and t-tests are standard techniques for this. Others include leave-one-out, bootstrapping, and loss functions. It would also be desirable if our research leads to a system that could be implemented in reality. Fortunately, our domain is characterized by an abundance of annual reports from companies as also by the availability of their historical performance figures. This allows us to use designs that simulate real-world application with historical data. We have two experiment designs for performing text classification each of which is capable of estimating the predictive model's performance.

2.5.1 Design One: Training and Testing with Documents from the Same Time Period

The key characteristic of this experiment is that both training and testing data are from the same time period. Training documents are from year t paired with class information built from change of financial performance in year $t+1$ relevant to t ($\Delta EPS^{f_i}_{(t,t+1)}$). Testing documents are also from year t and used to predict change in performance in year $t+1$ (again $\Delta EPS^{f_i}_{(t,t+1)}$).

We can vary year t systematically to perform a series of experiments all with the same design. Different years may generate different results. It would be interesting to determine if and why some years are easier to predict. From the research

perspective, this design can help us understand the feasibility of building such models and assessing the existence of predictive signals. Since training and testing data are from the same time period, we assume that we have "perfect knowledge" of the future environment where the model will be applied. This design is however not an implementable design because the real-world application requires applying the model to the data of a different time period. Estimate of model performance with this design will be optimistic.

In this design, we ensure that each class of documents is properly represented with minimum bias by conducting 10-fold cross-validation with stratification. Stratification is a procedure to ensure that the class distributions in the training sample and the test sample are identical. T-test significance tests are used to compare the models with the baselines and with each other. The average accuracy rate and average cost of errors from cross-validation and their corresponding p-values will support the conclusions we draw about our predictive models. Notice that cross-validation is done across companies for each year. Therefore, this experiment also gives us a sense of how well the predictive model could generalize across data from different firms.

Cross-validation with this design is described below using the EPS performance indicator as an example. The procedures are analogous for the other two indicators.

1. Split all documents in year t, $D_t^{f_i}$, into 10 equal random groups, according to the class of change of EPS in year $t+1$. Iteratively use 9 groups as training documents and 1 group as testing documents to form one fold, and thus obtain 10 folds of training and testing documents.

2. For each fold, build a classification model with the training documents and knowledge of their true classes ($TrueClassEPS_{(t,t+1)}^{f_i}$) defined with respect to corresponding $\Delta EPS_{(t,t+1)}^{f_i}$ values.

3. Apply the model to the testing documents and generate $PredictedClassEPS_{(t,t+1)}^{f_i}$

for each company of the testing documents.

4. For each fold, calculate model prediction accuracy by comparing $PredictedClassEPS^{f_i}_{(t,t+1)}$ and $TrueClassEPS^{f_i}_{(t,t+1)}$.

5. Average accuracy across all 10 folds.

6. Compare model's average accuracy with that of the two baselines. The baselines are as described in Section 2.3.1.

2.5.2 Design Two: Training and Testing with Documents from Adjacent Time Periods

The characteristic of this design is that the classification model is built with documents from one year (year t), but tested with documents from the immediately following year (year $t+1$) documents. This design uses historical data to simulate the real-world application of our predictive models for forecasting future performance. Thus it will give us an idea about how our classification models would perform if applied in reality. Again we vary t as models from different years may generate different accuracies. In contrast to design one, this design is implementable as the model is built using 'prior' data and tested on 'current' data for predicting future. Since the model will be tested on data set in a time period different from that of the training data, we expect the model's performance will not be as good as that of design one where both training and testing data come from the same time period. However, this design will provide us insights on how well the model is able to generalize across data from different years.

For example, the procedure involving EPS is as follows:

1. Use all documents in year t, $D^{f_i}_t$, as training documents. Build a classification model with their true classes $TrueClassEPS^{f_i}_{(t,t+1)}$ defined with respect to $\Delta EPS^{f_i}_{(t,t+1)}$.

2. Use all documents in year $t+1$, $D_{t+1}^{f_i}$, as testing documents. Apply the classification model and generate the predicted classes for year $t+2$ which is $PredictClassEPS_{t+2}^{f_i}$.

3. Calculate model prediction accuracy by comparing $PredictClassEPS_{t+2}^{f_i}$ and $TrueClassEPS_{t+2}^{f_i}$.

4. Compare model's average accuracy with that of the baselines.

The above procedure will be performed for each of the three performance indicators. The baselines are as described in Section 2.3.1.

2.5.3 Classification Algorithms

The above sections cover the five research dimensions identified in Chapter I. In addition, we need to choose an appropriate classification algorithm for our problem even though designing a better algorithm is not the focus of this paper.

Document indexing and information retrieval systems such as SMART are able to rank documents according to a similarity score with a given query. This similarity score function can be further extended or operationalized to produce a classifier. For example, Ittner et al. [41] converted the similarity scores into probabilities and applied logistic regression to estimate the parameters and perform categorization on new documents. The Rocchio algorithm is originally an information retrieval algorithm designed using relevance feedback[82, 83, 89]. The similarity scores from retrieval with the Rocchio algorithm have been used for categorization [46]. Knowledge engineering approaches such as expert system were also explored for document categorization [37].

Machine learning techniques and statistical methods have been extensively studied in the context of text classification. These include regression [108], naive Bayes classifiers [46, 59, 101], decision trees [14, 60], inductive rule learning [13], K nearest neighbors [109], neural networks [70, 85, 104], support vector machines [98, 110], and

classifier ensembles [92].

Previous research has shown that SVMs perform well as compared with classifiers such as naive Bayes, Rocchio, and K-NN [47]. In our pilot study [79], we used linear kernel function for SVM classification and achieved accuracy significantly better than baseline. We decide to use SVMs, specifically the SVM-Light[7] implementation of Support Vector Machines with default parameter settings and linear kernel function. Support vector machines (SVMs) [102] have been recognized as being able to efficiently handle high-dimensional problems. SVMs find a separating hyperplane in the high-dimensional space transformed from the original feature space through kernel function such that this hyperplane achieves the maximum margin in terms of the distance between the data points and the hyperplane. The data points thus separated are classified into different classes. Different kernel functions augment the feature space with different information to determine the distance between data points. The basic kernel functions used in most SVM implementations are linear function, polynomial function, radial basis function, and sigmoid function.

SVMs are designed mainly for solving binary or two-class classification problems. Studies on solving multi-classification problem have covered approaches to reduce multi-class to binary class problem [36], margin-based binary learning [3], using error-correcting output codes [21], and general multi-class prototype algorithm [105]. Since we have a three-class classification problem, we need to consider different options for this n (where $n = 3$) -class classification problem. First, we could perform one-against-rest classification for each class, and combine the results to make a final decision. Second, we could perform one-against-one classification for $n(n-1)/2$ pairs of classes, and combine the results to make a final decision. Third, we could use algorithms designed specifically for multi-class classification.

The disadvantage of using the second option (i.e. one-against-one) is that the

[7] http://svmlight.joachims.org/

number of classification models required will increase exponentially with the number of classes n. However, when n is small, such as $n = 3$, option one and two generate the same number (i.e. $n = 3$) of models. We decide to use option one and three for our current study. More specifically, for option one, we use linear SVM to produce a total of three one-against-rest models for the three classes. To combine the results of the three models, we experiment with two options: one is to use the highest predictive scores from each SVM model to assign a class label. We denote this multi-class classification model as *SVM-score*. Second, we transform the predictive scores of each SVM classifier into a probability of belonging to the positive class of that binary classification using Linn-Platt's method [64, 75]. Then we calibrate the results of the three models by picking the highest probability to assign the class label to the document. We denote this model as *SVM-prob*. For the third option of using the algorithm designed specifically for multi-class classification, we choose the package implemented by Joachims[8] based on Crammer and Singer's study [15]. We denote this model as *SVM-multi*.

As discussed in Section 2.2, we define the three-class problem based on a 25-50-25% data partition for each year. Alternatively, we could also define a 10-80-10% class definition where the top 10% firms are out-performing, the bottom 10% are under-performing and the middle 80% are average. We explore the effects of this alternative class definition with the mock portfolio experiments discussed in Section 2.3.2. This 10-80-10% class definition implies that the three binary classifiers will use data with highly-skewed distribution. For example, to predict the out-performing class, the data distribution is 10-90% positive vs. negative. The lack of data points for the minority class will unavoidably impact training the binary classifier.

There are different approaches to overcome the problem of highly-skewed data

[8] http://www.cs.cornell.edu/People/tj/svm_light/svm_multiclass.html/

as studied in [12, 42, 77]. One is to modify the cost parameter in the SVMs objective functions so that the cost of misclassifying minority class is higher than that of misclassifying the majority class. Another method is to duplicate the samples of the minority class when forming the training set, so that the two classes have approximately the same distributions. We choose this second approach to train each binary classifier with 10-80-10% class definition. For example, to build a binary classifier for predicting out-performing class where the distribution is 10-90% positive (i.e. out-performing) vs negative (i.e. average and under-performing), the positive data points are replicated 9 times in the training set so that the distribution of positives vs. negatives is forced to be 50-50%. By emphasizing the presence of minority data points, the classifier is trained to better distinguish the two classes. This approach of duplicating minority class data to overcome the skewed data distribution problem is applied to SVM-Score and SVM-Prob models.

To summarize, we consider three measures of firm performance: ROE, EPS, and SAR. Specifically, we measure change in performance from one year to the next (i.e. $\Delta ROE^{f_i}_{(t,t+1)}$, $\Delta EPS^{f_i}_{(t,t+1)}$ and $SAR^{f_i}_{(t+1,t+2)}$). We consider the problem as a three-class classification problem where a firm needs to be categorized as out-performing, average, or under-performing. We will use SVMs based classification algorithms, specifically three types of classifiers (i.e. *SVM-score*, *SVM-prob*, *SVM-multi*) will be used. Classification models will be built using TF×IDF weighted features extracted from annual reports. We will study two experimental designs one of which is 'implementable'. We will measure both accuracy and cost of error of our models and compare with the majority vote baseline as well as with analysts forecasts on EPS and stock recommendations. By shaping the problem in this way, we believe we will have a reasonably comprehensive understanding of our model's predictive potential in theory and in practice.

CHAPTER III
DATA COLLECTION

3.1 Financial Data Collection

We restrict our sample of firms to the manufacturing industry (SIC codes from 2000 to 3999) in the US, having December as the month defining the end of fiscal year. We also restrict our experiments to data from 1997 to 2003. These restrictions make our experiments tractable and also ensure some degree of sample homogeneity.

Our research goal is to predict change in financial performance using previous year's annual reports. Based on our definition of the three measures ($\Delta ROE^{f_i}_{(t,t+1)}$, $\Delta EPS^{f_i}_{(t,t+1)}$ and $SAR^{f_i}_{(t+1,t+2)}$), we need to collect the corresponding financial ratios.

For the experiments with $\Delta EPS^{f_i}_{(t,t+1)}$, and for each firm/year, we retrieve from I\B\E\S database[1] actual EPS, and analysts consensus mean for EPS forecast made in April of the fiscal year with forecast ending period in December of the fiscal year. The retrieved data ranges from 1997 to 2003, giving us $\Delta EPS^{f_i}_{(t,t+1)}$ data from 1998 to 2003.

For the experiments with $\Delta ROE^{f_i}_{(t,t+1)}$, we retrieve from COMPUSTAT database[2] data 18 (Income Before Extraordinary Items) from 1996 to 2003, data 216 (Stockholders' Equity) from 1996 to 2003. ROE is calculated as:

$ROE^{f_i}_{t+1} = Income^{f_i}_{t+1}/Equity^{f_i}_{t}$

Thus the $ROE^{f_i}_{t+1}$ data from 1997 to 2003 give us $\Delta ROE^{f_i}_{(t,t+1)}$ from 1998 to 2003.

For the experiments with $SAR^{f_i}_{(t+1,t+2)}$, we retrieve from CRSP database[3] the

[1] http://wrds.wharton.upenn.edu/
[2] http://wrds.wharton.upenn.edu/
[3] http://wrds.wharton.upenn.edu/

monthly return and decile monthly return for each firm/year. We calculate size-adjusted cumulative return as the size-adjusted buy-and-hold return cumulated for 12 months from April 1 of the fiscal year to the next April. The SAR data ranges from 1997 to 2003.

3.2 Document Collection

The data collected for the above three financial measures gives us a total of 1809 firms. Next, we retrieve the annual reports for these 1809 firms by first manually downloading the accession codes for these firms' annual reports from MergentOnline[4]. The accession codes are the unique identifiers of the annual reports. We then use these accession codes to automatically retrieve the reports from EdgarScan database [5]. Out of the 1809 firms with financial data, we are able to retrieve at least one annual report for 1519 firms. In total, we have 12564 annual reports from 1519 firms for year 1996 to 2004.

There are 10 different submission types for annual reports: 10K (10K filings), 10K405 (10K filings where regulation S-K Item 405 box on the cover page is checked), 10K405A (amendments to 10K405), 10KA (amendments to 10K filings), 10KSB (10K filings for small business), 10KSBA (amendments to 10KSB), 10KSB40 (optional form for small business where regulation S-B Item 405 box on the cover page is checked), 10KSB40A (amendment to 10KSB40), 10KT (10K transition re-port), 10KTA (amendment to 10K transition report). Table 3.1 shows the types of reports in this set and their distribution. We focus on the major submission types of 10K and 10K405. This sub-sample comprises 9,616 documents (or 76.54% of total retrieved valid documents). Our final useable documents with matching financial performance measure values are 5,421 annual reports from 1,276 firms published for years 1996 to 2002. Each firm in our data set does not have to have financial data or document for

[4]http://www.mergentonline.com
[5]http://edgarscan.pwcglobal.com/servlets/edgarscan

every year.

Filing Type	Number of Docs	Percentage
10K	7293	58.05%
10K405	2323	18.49%
10K405A	387	3.08%
10KA	2076	16.52%
10KSB	278	2.21%
10KSBA	92	0.73%
10KSB40	71	0.57%
10KSB40A	25	0.2%
10KT	13	0.1%
10KTA	6	0.05%

Table 3.1: **Filing Types Distribution of 12564 Annual Reports from 1519 Firms**

Table 3.2 identifies the number of documents used for each experiment. The numbers differ as each experiment (for a year/measure combination) depends on the availability of the corresponding performance data. Thus we see that certain experiments could only be run for particular years.

Table 3.3 provides descriptive data for the sample firms. The average firm has a mean sales of $2,749 million and median sales of $324 million, indicating the presence of several large firms in the sample. The average ROE is 4.86%, while the median is 7.23%. The mean and median values for the market to book ratio are 1.97 and 1.19 respectively. Table 3.4 shows the industry breakdown for the sample firms. We do not find any significant clustering of industries specific to our sample. The spread of firms is similar to the distribution of all firms in COMPUSTAT database from the relevant SIC codes. Therefore, we expect our results to be generalizable, albeit only

	Number of Documents			
Year	**Total**	**$\Delta ROE^{f_i}_{(t,t+1)}$ Experiments**	**$\Delta EPS^{f_i}_{(t,t+1)}$ Experiments**	**$SAR^{f_i}_{(t+1,t+2)}$ Experiments**
1997	782	–	–	782
1998	857	–	746	765
1999	804	612	714	719
2000	771	584	663	726
2001	798	566	649	758
2002	743	564	674	710
2003	666	601	658	–
Total	5421	4138	4104	4460

Table 3.2: **Distribution of Documents Used for Experiments**

to manufacturing firms.

Item	No. of Firm/Year	Mean	Median	25th Percentile	75th Percentile
Sales (millions)	5196	$2,749.16	$324.27	$64.03	$1582.93
Net Assets (millions)	5193	$3,219.29	$368.47	$95.82	$1682.87
ROE	4132	4.86%	7.23%	-12.34%	17.88%
EPS	5372	$0.2557	$0.58	$-0.25	$1.36
Size Adjusted Return	4504	3.12%	-12.62%	-41.95%	20.54%
Market-to-book Ratio	5189	1.97	1.19	0.64	2.39

Table 3.3: **Data Descriptive Statistics**

As discussed in Section 2.2, we divide the data for each firm/year into three classes: out-performing, average, and under-performing. More specifically, we cut off 2% of data points at each tail before we categorize the three classes. Tables 3.5, 3.7, and 3.6 show the data statistics for the three classes using three financial measures.

In addition, as discussed in Section2.3.3, we also conduct several experiments

SIC Code (Description)	**No. of Firms**
20 (Food And Kindred Products)	40
21 (Tobacco Products)	5
22 (Textile Mill Products)	19
23 (Apparel And Other Finished Products Made From Fabrics And Similar Materials)	13
24 (Lumber And Wood Products, Except Furniture)	16
25 (Furniture And Fixtures)	11
26 (Paper and Allied Products)	39
27 (Printing, Publishing, And Allied Industries)	34
28 (Chemicals And Allied Products)	284
29 (Petroleum Refining And Related Industries)	18
30 (Rubber And Miscellaneous Plastics Products)	31
31 (Leather And Leather Products)	9
32 (Stone, Clay, Glass, And Concrete Products)	20
33 (Primary Metal Industries)	43
34 (Fabricated Metal Products, Except Machinery And Transportation Equipment)	40
35 (Industrial And Commercial Machinery And Computer Equipment)	177
36 (Electronic And Other Electrical Equipment And Components, Except Computer Equipment)	193
37 (Transportation Equipment)	55
38 (Measuring, Analyzing, And Controlling Instruments; Photographic, Medical And Optical Goods; Watches And Clocks)	206
39 (Miscellaneous Manufacturing Industries)	23
Total	1,276

Table 3.4: **Industry Composition**

to examine the predictive ability of annual reports of sub-sample firms, i.e. large vs small, loss vs profit, and growth vs glamour firms. The detailed class definitions are provided in Appendix B. Further, to test the model robustness, we also define a different three-class for the data of SAR measure experiments. In stead of a 25-50-25% definition, we divide the data into a 10-80-10% category and study the models performance. The data detail is provided in Table 3.8.

	Before 2% Cutoff Each Tail				**After 2% Cutoff Each Tail**					
	Bottom 2%		**Top 2%**		**Bottom 25%**		**Middle 50%**		**Top 25%**	
Year (Total size, Size After Cutoff)	**EPS Measure Threshold**	**Sample Lost**	**EPS Measure Threshold**	**Sample Lost**	**EPS Measure Threshold**	**Sample Size**	**EPS Measure Threshold**	**Sample Size**	**EPS Measure Threshold**	**Sample Size**
1998 (746, 717)	<-528.57%	14	>257.14%	15	<-33.33%	179	≥-33.33%, <22.22%	358	≥22.22%	180
1999 (714, 686)	<-400%	13	>375%	15	<-19.51%	171	≥-19.51%, <36%	342	≥36%	173
2000 (663, 636)	<-335.85%	13	>798.21%	14	<-15.22%	158	≥-15.22%, <41.67%	317	≥41.67%	161
2001 (649, 624)	<-572.73%	12	>150%	13	<-77.59%	155	≥-77.59%, <6.9%	312	≥6.9%	157
2002 (674, 647)	<-975%	13	>563.16%	14	<-27.52%	161	≥-27.52%, <42%	324	≥42%	162
2003 (658, 631)	<-500%	13	>513.25%	14	<-11.11%	155	≥-11.11%, <40.72%	318	≥40.72%	158

Table 3.5: **Class Definition with $\boldsymbol{\Delta EPS^{f_i}_{(t,t+1)}}$**

	Before 2% Cutoff Each Tail				After 2% Cutoff Each Tail					
	Bottom 2%		Top 2%		Bottom 25%		Middle 50%		Top 25%	
Year (Total size, Size After Cutoff)	**ROE Measure Threshold**	**Sample Lost**	**ROE Measure Threshold**	**Sample Lost**	**ROE Measure Threshold**	**Sample Size**	**ROE Measure Threshold**	**Sample Size**	**ROE Measure Threshold**	**Sample Size**
1999 (612, 587)	<-708.33%	12	>1017.61%	13	<-43.55%	146	≥-43.55%, <60.78%	294	≥60.78%	147
2000 (584, 561)	<-894.61%	11	>1040.88%	12	<-45.96%	140	≥-45.96%, <39.10%	280	≥39.10%	141
2001 (566, 543)	<-1112.91%	11	>459.37%	12	<-101.83%	135	≥-101.83%, <8.37%	272	≥8.37%	136
2002 (564, 541)	<-987.17%	11	>835.18%	12	<-48.98%	135	≥-48.98%, <65.18%	270	≥65.18%	136
2003 (601, 577)	<-876.16%	11	>824.33%	13	<-33.29%	144	≥-33.29%, <56.60%	288	≥56.60%	145

Table 3.6: **Class Definition with $\Delta ROE^{f_i}_{(t,t+1)}$**

	Before 2% Cutoff Each Tail				After 2% Cutoff Each Tail					
	Bottom 2%		Top 2%		Bottom 25%		Middle 50%		Top 25%	
Year (Total size, Size After Cutoff)	**SAR Measure Threshold**	**Sample Lost**	**SAR Measure Threshold**	**Sample Lost**	**SAR Measure Threshold**	**Sample Size**	**SAR Measure Threshold**	**Sample Size**	**SAR Measure Threshold**	**Sample Size**
1997 (782, 751)	<-98.44%	15	>130.84%	16	<-42.23%	187	≥-42.23%, <11.22%	376	≥11.22%	188
1998 (765, 734)	<-68.97%	15	>122.96%	16	<-39.62%	183	≥-39.62%, <4.54%	367	≥4.54%	184
1999 (719, 690)	<-127.65%	14	>770.07%	15	<-61.50%	172	≥-61.50%, <41.69%	345	≥41.69%	173
2000 (726, 697)	<-77.23%	14	>116.01%	15	<-36.17%	174	≥-36.17%, <34.26%	348	≥34.26%	175
2001 (758, 727)	<-93.30%	15	>179.07%	16	<-30.84%	181	≥-30.84%, <27.63%	364	≥27.63%	182
2002 (710, 681)	<-68.36%	14	>64.38%	15	<-34.47%	170	≥34.47%, <9.19%	340	≥9.19%	171

Table 3.7: **Class Definition with $SAR^{f_i}_{(t+1,t+2)}$**

	Before 2% Cutoff Each Tail				After 2% Cutoff Each Tail					
	Bottom 2%		Top 2%		Bottom 10%		Middle 80%		Top 10%	
Year (Total size, Size After Cutoff)	**SAR Measure Threshold**	**Sample Lost**	**SAR Measure Threshold**	**Sample Lost**	**SAR Measure Threshold**	**Sample Size**	**SAR Measure Threshold**	**Sample Size**	**SAR Measure Threshold**	**Sample Size**
1997 (782, 751)	<-98.44%	15	>130.84%	16	<-67.14%	75	≥-67.14%, <45.78%	600	≥45.78%	76
1998 (765, 734)	<-68.97%	15	>122.96%	16	<-52.38%	73	≥-52.38%, <29.53%	587	≥29.53%	74
1999 (719, 690)	<-127.65%	14	>770.07%	15	<-92.05%	68	≥-92.05%, <228.65%	552	≥228.65%	70
2000 (726, 697)	<-77.23%	14	>116.01%	15	<-57.35%	69	≥-57.35%, <56.67%	558	≥56.67%	70
2001 (758, 727)	<-93.30%	15	>179.07%	16	<-63.41%	72	≥-63.41%, <56.89%	582	≥56.89%	73
2002 (710, 681)	<-68.36%	14	>64.38%	15	<-52.38%	68	≥-52.38%, <26.07%	544	≥26.07%	69

Table 3.8: **10-80-10 Class Definition with $SAR^{f_i}_{(t+1,t+2)}$**

CHAPTER IV
RESULTS

In this Chapter, we present the results of our experiments. Each experiment represents a combination of the choices we made with respect to the five research dimensions: financial performance indicators, choice of prediction, evaluation criteria, document representation method, and experiment design. Since our goal is to explore the potential of building predictive models with annual reports, we first compare our three models (i.e. SVM-score, SVM-prob, and SVM-multi) using both experiment design one and two. Then, we look into what the best model has to say about predicting companies' future financial performances.

4.1 Experiment Design One: Cross Validation Design

4.1.1 Comparing SVM Models

We first find out how the three models (i.e. SVM-prob, SVM-score, and SVM-multi) perform with experiment design one. As discussed in Section 2.5.1, design one uses the data of the same year to both train and test the classifiers with cross validation. Table 4.1 compares predictors on their rankings in terms of accuracy. Ranking is given based on the predictive accuracy values of the three models evaluated with the same year and measure. Rank 1 is given to the model with the highest accuracy. We see that SVM-prob has the best average rank (reported in the second to last row) compared to SVM-score and SVM-multi. This holds for all three performance measures. The last row of the table presents comparisons with accuracies obtained using the majority vote baseline (defined in Section 3.3.1). For example, we obtain 4/6 for SVM-prob under the EPS measure. This implies that for four of the six EPS and SVM-prob experiments (1998 through 2003), SVM-prob was significantly better than

the majority vote baseline. Thus we see that for 60 to 80% of the runs SVM-prob is better than baseline depending on measure. Interestingly, SVM-score and SVM-multi face serious problems with the ROE based measure.

	$\Delta EPS^{f_i}_{(t,t+1)}$			$\Delta ROE^{f_i}_{(t,t+1)}$			$SAR^{f_i}_{(t+1,t+2)}$		
Year	**SVM-score**	***SVM-prob***	**SVM-multi**	**SVM-score**	***SVM-prob***	**SVM-multi**	**SVM-score**	***SVM-prob***	**SVM-multi**
1997							3 (0.507, 0.001)	1 (0.525, 0.001)	2 (0.521, 0.001)
1998	3 (0.50, 0.002)	1 (0.511, <0.001)	2 (0.509, 0.001)				1 (0.544, 0.001)	2 (0.538, 0.001)	3 (0.521, 0.001)
1999	1 (0.542, 0.001)	2 (0.541, 0.003)	3 (0.512, 0.001)	1 (0.51, <0.001)	2 (0.49, 0.002)	3 (0.48, <0.001)	2 (0.615, 0.003)	1 (0.628, 0.004)	3 (0.591, 0.002)
2000	1 (0.566, 0.002)	2 (0.563, 0.002)	3 (0.55, 0.001)	3 (0.497, 0.003)	1 (0.52, 0.002)	2 (0.51, 0.001)	3 (0.539, 0.002)	1 (0.565, 0.001)	2 (0.543, 0.002)
2001	3 (0.529, 0.003)	1 (0.546, 0.002)	2 (0.54, 0.003)	3 (0.52, 0.002)	1 (0.53, 0.001)	2 (0.526, 0.003)	3 (0.519, 0.004)	1 (0.528, 0.003)	2 (0.524, 0.001)
2002	2 (0.547, 0.004)	1 (0.55, 0.004)	3 (0.529, 0.001)	2 (0.497, 0.002)	1 (0.516, <0.001)	3 (0.47, <0.001)	2 (0.523, 0.002)	1 (0.541, 0.003)	3 (0.515, 0.001)
2003	2 (0.531, 0.003)	3 (0.523, 0.001)	1 (0.534, 0.001)	2 (0.506, 0.003)	1 (0.53, 0.001)	3 (0.50, 0.001)			
Average Rank (Accu., Var.)	2 (0.5367, <0.001)	**1.67 (0.5389, <0.001)**	2.33 (0.5289, <0.001)	2.2 (0.506, <0.001)	**1.2 (0.519, <0.001)**	2.6 (0.498, <0.001)	2.33 (0.541, 0.001)	**1.17 (0.554, 0.001)**	2.5 (0.537, 0.001)

Cell notation: Rank (Average Accuracy, Variance)

Table 4.1: **Comparing Models: Ranking Based on Predictive Accuracy (Design One)**

Table 4.2 compares the three types of classifiers across performance measures and across years evaluated with cost of errors (as defined in Section 3.3.4). Comparisons are done by ranking the classifiers in terms of cost of errors with 1 representing the best peformance. That is rank 1 is given to the classifier that incurs the least cost of errors. For example, in 1998 SVM-multi had the lowest error cost followed by SVM-prob and then SVM-score when using the EPS based performance measure. The last row of the table gives the average rank. We see that SVM-prob has the best

rank for the ROE and SAR measures. While SVM-prob has the second average rank for EPS, it is still very close to the average rank of the best classifier by cost (i.e. SVM-multi). From this table it is reasonable to conclude that SVM-prob is the best approach of the three overall.

	$\Delta EPS^{f_i}_{(t,t+1)}$			$\Delta ROE^{f_i}_{(t,t+1)}$			$SAR^{f_i}_{(t+1,t+2)}$		
Year	**SVM-score**	**SVM-prob**	***SVM-multi***	**SVM-score**	***SVM-prob***	**SVM-multi**	**SVM-score**	***SVM-prob***	**SVM-multi**
1997							3 (41.1, 11.2)	1 (38.3, 11.3)	2 (38.5, 12)
1998	3 (40.4, 13.4)	2 (38, 5.8)	1 (36.6, 5.8)				3 (39.2, 12.8)	2 (37.7, 20)	1 (37.4, 10.9)
1999	2 (35.4, 11.6)	1 (34, 18)	3 (36.2, 10.2)	2 (31.9, 2.9)	1 (31, 11.5)	3 (32.8, 1.7)	2 (28.8, 11.7)	1 (27.6, 16.5)	3 (29.4, 11.2)
2000	3 (31.6, 10.7)	1 (30.2, 7.3)	2 (30.7, 6.7)	3 (31.8, 19.9)	1 (28.1, 5.2)	2 (30.1, 8.9)	3 (34.1, 11)	1 (31.6, 9.4)	2 (33.6, 17.6)
2001	3 (31.4, 18.5)	2 (30.2, 11.3)	1 (29.8, 6.6)	3 (29.5, 5.4)	1 (27.3, 4.5)	2 (27.9, 7.2)	2 (39.3, 18)	1 (36.7, 8.5)	1 (36.7, 9.6)
2002	3 (33.3, 20.5)	2 (32.4, 17.8)	1 (31.9, 3.9)	2 (30.3, 13.3)	1 (26.9, 8.3)	3 (30.8, 3.3)	3 (35, 24.7)	1 (32.7, 29.8)	2 (34.4, 8.9)
2003	3 (33.1, 19.6)	2 (32.9, 5.6)	1 (31.1, 7.4)	3 (32.1, 14.5)	1 (28.9, 5.2)	2 (31, 6.4)			
Average Rank (Cost, Var.)	2.83 (34.2, 11.3)	1.67 (32.95, 8.4)	**1.5 (32.7, 8.6)**	2.6 (31.1, 1.3)	**1 (28.4, 2.6)**	2.4 (30.5, 3.1)	2.67 (36.3, 20.6)	**1.17 (34.1, 17.6)**	1.83 (35, 10.9)

Cell Notation: Rank (Average Cost, Variance)

Table 4.2: **Comparing Models: Ranking Based on Cost of Errors (Design One)**

Table 4.3 shows the consolidated comparisons with baseline for three models for each year. We can see that SVM-prob is able to predict significantly better than baseline in 12 out of 17 (or 71%) measure/year predictions, with the rest the same as baseline. SVM-score can predict significantly better than baseline in 6 out of 17 (or 35%) measure/year predictions, with the rest the same as baseline. SVM-multi

predicts significantly better in 8 out of 17 (or 47%) measure/year predictions. However, it gives 2 out of 17 (or 12%) significantly worse prediction, and 7 out of 17 (or 41%) same as the baseline. Consistent with the observations in Table 4.1, we find that all three models have most difficulty in giving better-than-baseline predictions for $\Delta ROE^{f_i}_{t+1,t}$, compared with predicting the other two measures.

	SVM-score			***SVM-prob***			**SVM-multi**		
Year	**EPS**	**ROE**	**SAR**	**EPS**	**ROE**	**SAR**	**EPS**	**ROE**	**SAR**
1997	–	–	S	–	–	***B***	–	–	S
1998	S	–	***B***	S	–	***B***	S	–	***B***
1999	***B***	S	***B***	***B***	S	***B***	S	W	***B***
2000	***B***	S	***B***	***B***	S	***B***	***B***	S	***B***
2001	S	S	S	***B***	***B***	S	***B***	S	***B***
2002	***B***	S	S	***B***	***B***	***B***	***B***	W	S
2003	S	S	–	S	***B***	–	***B***	S	–
Summary [B-S-W]	[3-3-0]	[0-5-0]	[3-3-0]	[4-2-0]	[3-2-0]	[5-1-0]	[4-2-0]	[0-3-2]	[4-2-0]

Notes: B, W and S: significantly better, worse or the same as baseline respectively.

Table 4.3: **Comparison of Three Models and Majority Vote Baseline**

Table 4.4 summarizes the comparison of the three classifier models with the three financial performance measures for design one. As stated before, ranking is done by comparing the average predictive accuracies among the three models with 1 being the best of the three. Rank 1 for accuracy implies the highest accuracy among the three models. Rank 1 for cost means lowest cost of errors in three models. The table shows the average rankings based on predictive accuracy and cost of errors for three financial performance measures. As shown in the last row of this table, SVM-prob is overall the better performer when evaluated with accuracy and cost of errors.

The only one exception when SVM-prob is the second best is for EPS measure evaluated with cost of errors.

	Average Cost Ranking			**Average Accuracy Ranking**		
Performance Measure	**SVM-score**	***SVM-prob***	**SVM-multi**	**SVM-score**	***SVM-prob***	**SVM-multi**
$\boldsymbol{\Delta EPS^{f_i}_{(t,t+1)}}$	2.8	1.67	1.5	2.0	1.67	2.33
$\boldsymbol{\Delta ROE^{f_i}_{(t,t+1)}}$	2.6	1.0	2.4	2.2	1.2	2.6
$\boldsymbol{SAR^{f_i}_{(t+1,t+2)}}$	2.67	1.17	1.83	2.33	1.17	2.5
Average Ranking Over Measures	2.7	**1.28**	1.91	2.18	**1.34**	2.48

Table 4.4: **Comparing Models with Design One and All Three Measures from All Years**

Tables 4.5, 4.6, and 4.7 are the classification contingency tables of all three models for predicting $\Delta EPS^{f_i}_{(t,t+1)}$ with design one, averaged over years. Since our baseline is majority vote which assigns every test data point as "Average", we are interested in particular in the models' true predictive rates for the out-performing and the under-performing firms. In addition, there are two types of errors that we would pay special attention to: predicting out-performing as under-performing, and predicting under-performing as out-performing. The former is loss of opportunity. The latter is loss with high cost. Comparing the three classification contingency tables for predicting $\Delta EPS^{f_i}_{(t,t+1)}$, we observe that SVM-score has the highest true predictive rates for both out-perform and under-perform, as well as the highest error rates of both types. SVM-prob model has the second highest true predictive rates and error rates of both types. SVM-multi model is the lowest in both true predictive rates and error rates.

SVM-prob is our best model overall in terms of predictive accuracy and cost of

error as discussed above. It seems that SVM-prob is able to achieve this performance by balancing the trade-off between giving correct prediction and taking risks for error. SVM-score seems to be the "boldest" in risking for errors in order to achieve higher true predictive rate. SVM-multi seems to be the most "conservative" in sacrificing true predictive rate for lowest error rates. Similar observations can be made for the three models in predicting $SAR^{f_i}_{(t+1,t+2)}$ and $\Delta ROE^{f_i}_{(t,t+1)}$ with design one. The classification contingency tables for these two measures are provided in Appendix C, Tables C.1, C.2, C.3, C.4, C.5, and C.6.

Design One	**SVM-score Prediction on $\Delta EPS^{f_i}_{(t,t+1)}$**			
25-50-25% Class Definition	**Out-Perform**	**Average**	**Under-Perform**	**True Distribution**
True Out-Perform	27.47%	62.31%	10.23%	25.15%
True Average	9.75%	81.38%	8.86%	50.02%
True Under-Perform	12.38%	63.25%	24.37%	24.84%
Prediction Distribution	14.85%	72.09%	13.06%	100.00%

Table 4.5: **Classification Contingency Table of SVM-score for Predicting $\Delta EPS^{f_i}_{(t,t+1)}$ with Design One**

Design One	**SVM-prob Prediction on $\Delta EPS^{f_i}_{(t,t+1)}$**			
25-50-25% Class Definition	**Out-Perform**	**Average**	**Under-Perform**	**True Distribution**
True Out-Perform	21.10%	71.26%	7.66%	25.15%
True Average	7.13%	87.32%	5.54%	50.02%
True Under-Perform	8.37%	71.85%	19.79%	24.84%
Prediction Distribution	10.94%	79.46%	9.61%	100.00%

Table 4.6: **Classification Contingency Table of SVM-prob for Predicting $\Delta EPS^{f_i}_{(t,t+1)}$ with Design One**

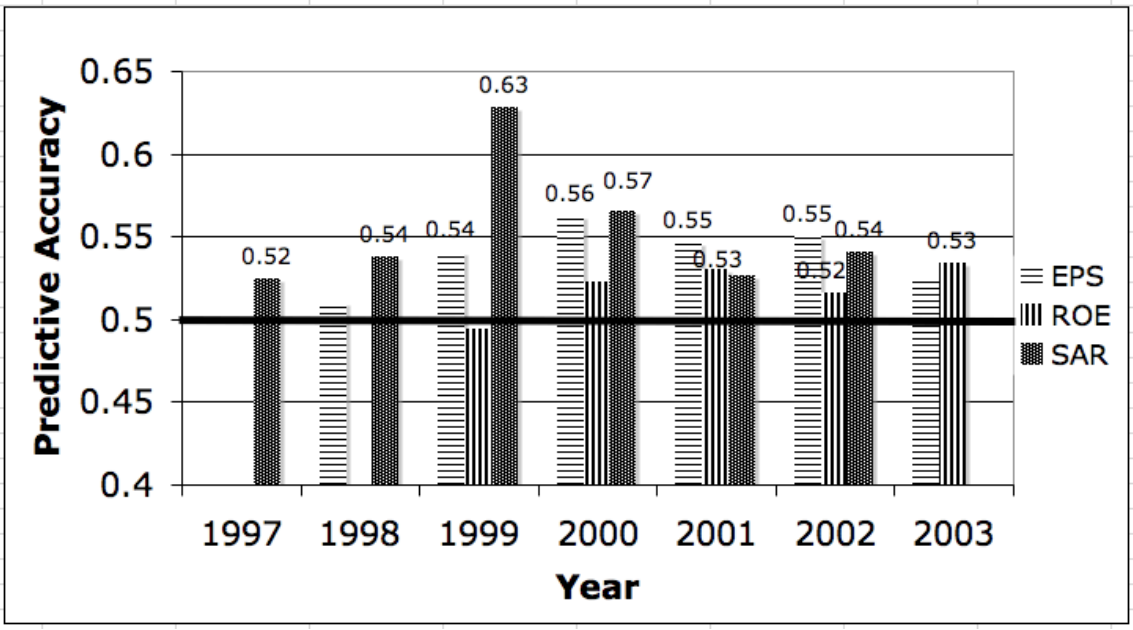
Design One	**SVM-multi Prediction on $\Delta EPS^{f_i}_{(t,t+1)}$**			
25-50-25% Class Definition	**Out-Perform**	**Average**	**Under-Perform**	**True Distribution**
True Out-Perform	14.04%	81.39%	4.57%	25.15%
True Average	3.72%	92.60%	3.70%	50.02%
True Under-Perform	5.99%	81.73%	12.28%s	24.84%
Prediction Distribution	6.87%	87.08%	6.06%	100.00%

Table 4.7: **Classification Contingency Table of SVM-multi for Predicting $\Delta EPS^{f_i}_{(t,t+1)}$ with Design One**

4.1.2 Results with SVM-prob Model

Finally we close this section by plotting the actual scores obtained using the SVM-prob model for predicting the three financial measures. These are plotted in Figure 4.1. We see that SVM-prob is able to predict significantly better than baseline in 12 out of 17 (71%) sets of financial measure/year prediction. Among the three financial performance measures, $SAR^{f_i}_{(t+1,t+2)}$ has the best predictability with 5 out of 6 (83%) predictions as significantly better than baseline, followed by $\Delta EPS^{f_i}_{(t,t+1)}$ with 67% and $\Delta ROE^{f_i}_{(t,t+1)}$ with 60% significantly better prediction. This is further confirmed by comparing the average predictive accuracies by financial measures as shown in Table 4.8. $SAR^{f_i}_{(t+1,t+2)}$ has the highest average accuracy over years, followed by $\Delta EPS^{f_i}_{(t,t+1)}$ and $\Delta ROE^{f_i}_{(t,t+1)}$. As discussed in previous section, $\Delta ROE^{f_i}_{(t,t+1)}$ as a financial performance measure does not seem to possess prominent predictive power as the other two measures. Looking at the predictive accuracies by year as shown in Table 4.9, among the years with results from all three financial measures (i.e. 1999 to 2002), we find that 1999 is the year with the highest average accuracy across all three measures. In other words, year 1999 seems the easiest to predict by our best model.

Figure 4.1: **SVM-prob Average Accuracy with Design One for All Three Financial Measures**



Note: Model's significant accuracies are shown.

Design One	$\Delta EPS^{f_i}_{(t,t+1)}$	$\Delta ROE^{f_i}_{(t,t+1)}$	$SAR^{f_i}_{(t+1,t+2)}$
Accuracy Averaged Over Years	0.5389	0.5193	0.5597

Table 4.8: **Average Predictive Accuracy by Financial Measures with Design One, SVM-prob Model**

Design One	**1999**	**2000**	**2001**	**2002**
Accuracy Averaged Over Measures	0.5544	0.5501	0.5344	0.5357

Table 4.9: **Average Predictive Accuracy by Years with Design One, SVM-prob Model**

4.2 Experiment Design Two: Implementable Design

4.2.1 Comparing Models

Experiments of design two use the documents from year t to train a model and the documents from year t+1 to test a model. This design estimates the potential of applying the models in real-world prediction. We first present the comparison of the three models with design two, evaluated with accuracy and cost of error. Then we show the results of our best model. Table 4.10 compares the ranks by predictive accuracy for all three models with design two. We can see that SVM-prob ranks the best for all three financial measures (i.e. $\Delta EPS^{f_i}_{(t,t+1)}$, $\Delta ROE^{f_i}_{(t,t+1)}$, and $SAR^{f_i}_{(t+1,t+2)}$). Although for $\Delta EPS^{f_i}_{(t,t+1)}$, SVM-mult comes in very close in second place to SVM-prob, the rank differences between SVM-score/SVM-multi and SVM-prob for the other two financial measure (i.e. $\Delta ROE^{f_i}_{(t,t+1)}$ and $SAR^{f_i}_{(t+1,t+2)}$) are fairly large.

Table 4.11 gives us the ranking by cost of errors for all three models. SVM-prob ranks the best on average for $\Delta ROE^{f_i}_{(t,t+1)}$ and $SAR^{f_i}_{(t+1,t+2)}$. While SVM-multi performs the best in terms of costing the lowest errors for $\Delta EPS^{f_i}_{(t,t+1)}$ experiments, the second best model (i.e. SVM-prob) comes in very close to SVM-multi in average rank.

Table 4.12 summarizes the comparison of three models with design two and three financial performance measures. The last row shows that on average, SVM-prob stands out as the best in terms of both predictive accuracy and cost of errors. The only case when SVM-prob model ranks second with a very small margin is its cost ranking for predicting $\Delta EPS^{f_i}_{(t,t+1)}$. SVM-score is consistently the worst of the three when evaluated with accuracy and cost of errors.

Tables 4.13, 4.14, and 4.15 are the classification contingency tables of SVM-score, SVM-prob, and SVM-multi respectively for predicting $SAR^{f_i}_{(t+1,t+2)}$, averaged over years. We can observe that SVM-prob is not able to predict out-performing firms

	$\Delta EPS^{f_i}_{(t,t+1)}$			$\Delta ROE^{f_i}_{(t,t+1)}$			$SAR^{f_i}_{(t+1,t+2)}$		
Year	**SVM-score**	***SVM-prob***	**SVM-multi**	**SVM-score**	***SVM-prob***	**SVM-multi**	**SVM-score**	***SVM-prob***	**SVM-multi**
1998							3 (0.4428)	1 (0.5)	2 (0.4632)
1999	2 (0.4956)	1 (0.5073)	3 (0.4869)				3 (0.4478)	1 (0.5)	2 (0.4898)
2000	2 (0.5173)	3 (0.5)	1 (0.533)	3 (0.4831)	2 (0.4866)	1 (0.508)	3 (0.33)	1 (0.5237)	2 (0.4146)
2001	3 (0.4744)	1 (0.5337)	2 (0.4808)	3 (0.4954)	1 (0.523)	2 (0.4972)	3 (0.4677)	1 (0.5048)	2 (0.5007)
2002	3 (0.4683)	1 (0.524)	2 (0.5116)	3 (0.4695)	1 (0.512)	2 (0.4806)	3 (0.4449)	1 (0.5037)	2 (0.489)
2003	3 (0.4802)	2 (0.4913)	1 (0.4992)	2 (0.4714)	1 (0.487)	3 (0.4593)			
Average Rank (Accu., Var.)	2.6 (0.4872, <0.001)	**1.6 (0.5112, <0.001)**	1.8 (0.5023, <0.001)	2.75 (0.4798, <0.001)	**1.25 (0.5022, <0.001)**	2 (0.4863, <0.001)	3 (0.4266, 0.003)	**1 (0.5064, <0.001)**	2 (0.4715, 0.001)

Cell notation: Rank (Accuracy)

Table 4.10: **Comparing Models: Ranking Based on Predictive Accuracy (Design Two)**

while the other two models are able to. The same observation can be made of these models for predicting $\Delta EPS^{f_i}_{(t,t+1)}$ and $\Delta ROE^{f_i}_{(t,t+1)}$, as illustrated in the Appendix D, Tables D.1, D.2, D.3, D.4, D.5, and D.6. Another interesting observation is that in classifying out-performing and under-performing firms, SVM-score and SVM-multi appear to have different patterns. SVM-score's correct prediction for out-performing and under-performing is higher in percentage than SVM-multi (17.96% vs 8.86%, and 24.33% vs 16%). On the other hand, SVM-score's error rates of misclassifying out-performing as under-performing and vice versa are also higher than SVM-multi (20.5% vs 15.07%, and 17.52% vs 10.59%). This observation is also consistent in predicting the other two financial measures as illustrated in Appendix D. We observe that, similar to design one, SVM-score risks higher cost of errors for achieving higher

	$\Delta EPS^{f_i}_{(t,t+1)}$ Rank (Cost)			$\Delta ROE^{f_i}_{(t,t+1)}$ Rank (Cost)			$SAR^{f_i}_{(t+1,t+2)}$ Rank (Cost)		
Year	**SVM-score**	**SVM-prob**	***SVM-multi***	**SVM-score**	***SVM-prob***	**SVM-multi**	**SVM-score**	***SVM-prob***	**SVM-multi**
1998							3 (470)	1 (367)	2 (425)
1999	3 (430)	2 (390)	1 (388)				3 (423)	1 (345)	2 (379)
2000	2 (379)	3 (399)	1 (333)	3 (365)	2 (330)	1 (313)	3 (615)	1 (334)	2 (515)
2001	3 (426)	1 (332)	2 (391)	3 (360)	1 (307)	2 (319)	3 (434)	1 (389)	2 (403)
2002	3 (412)	1 (353)	2 (360)	3 (355)	1 (305)	2 (333)	3 (415)	1 (357)	2 (369)
2003	3 (401)	2 (360)	1 (355)	2 (371)	1 (337)	3 (379)			
Average Rank (Cost, Var.)	2.8 (409.6, 425.3)	1.8 (366.8, 755.7)	**1.4 (365.4, 588.3)**	2.75 (365.75, 46.92)	**1.25 (319.75, 260.92)**	2 (329, 1144)	3 (471.4, 6886.3)	**1 (358.4, 446.8)**	2 (418.2, 3401.2)

Table 4.11: **Comparing Models: Ranking Based on Cost of Errors (Design Two)**

	Average Cost Ranking			**Average Accuracy Ranking**		
Performance Measure	**SVM-score**	***SVM-prob***	**SVM-multi**	**SVM-score**	***SVM-prob***	**SVM-multi**
$\Delta EPS^{f_i}_{(t,t+1)}$	2.8	1.8	1.4	2.6	1.6	1.8
$\Delta ROE^{f_i}_{(t,t+1)}$	2.75	1.25	2	2.75	1.25	2
$SAR^{f_i}_{(t+1,t+2)}$	3	1	2	3	1	2
Average Ranking Over Measures	2.85	**1.35**	1.8	2.78	**1.28**	1.93

Table 4.12: **Comparing Models with Design Two and All Three Measures from All Years**

predictive accuracies when predicting out-performing and under-performing. SVM-multi on the other hand achieves lower error rates at the cost of lower predictive accuracies when predicting out-performing and under-performing.

Design Two	SVM-score Prediction on $SAR^{f_i}_{(t+1,t+2)}$			
25-50-25% Class Definition	**Out-Perform**	**Average**	**Under-Perform**	**True Distribution**
True Out-Perform	17.96%	64.52%	17.52%	25.08%
True Average	18.96%	64.21%	16.83%	49.99%
True Under-Perform	20.50%	55.17%	24.33%	24.94%
Prediction Distribution	19.09%	62.04%	18.87%	100.00%

Table 4.13: **Classification Contingency Table of SVM-score for Predicting $SAR^{f_i}_{(t+1,t+2)}$ with 25-50-25% Class Definition**

Design Two	SVM-prob Prediction on $SAR^{f_i}_{(t+1,t+2)}$			
25-50-25% Class Definition	**Out-Perform**	**Average**	**Under-Perform**	**True Distribution**
True Out-Perform	0%	94.36%	5.64%	25.08%
True Average	0%	94.09%	5.91%	49.99%
True Under-Perform	0%	85.53%	14.47%	24.94%
Prediction Distribution	0%	92.03%	7.97%	100.00%

Table 4.14: **Classification Contingency Table of SVM-prob for Predicting $SAR^{f_i}_{(t+1,t+2)}$ with 25-50-25% Class Definition**

Design Two	SVM-multi Prediction on $SAR^{f_i}_{(t+1,t+2)}$			
25-50-25% Class Definition	**Out-Perform**	**Average**	**Under-Perform**	**True Distribution**
True Out-Perform	8.86%	80.54%	10.59%	25.08%
True Average	8.83%	81.90%	9.27%	49.99%
True Under-Perform	15.07%	68.93%	16.00%	24.94%
Prediction Distribution	10.40%	78.33%	11.28%	100.00%

Table 4.15: **Classification Contingency Table of SVM-multi for Predicting $SAR^{f_i}_{(t+1,t+2)}$ with 25-50-25% Class Definition**

4.2.2 Results with SVM-prob Model

Finally, we present the predictive results by our best performer, the SVM-prob model. Figure 4.2 shows the model's accuracy as compared with the baseline. The

baseline accuracy is 50%. The experiment design determines that we will not be able to perform t-test significance test. However, we can see that out of the 14 measure/year predictions, the model is able to predict better than the baseline for 6 measure/years, and the same (within 0.5%) as baseline in 5 measure/years. Table 4.16 shows that $\Delta ROE^{f_i}_{(t,t+1)}$ again is the hardest to predict with the lowest average accuracy across years. Similar to the observations made for design one, $\Delta EPS^{f_i}_{(t,t+1)}$ is the easiest to predict with the highest average predictive accuracy over years, although the differences in average accuracy are very small compared with the other two financial measures. As shown in Table 4.17, out of the three years (2000 to 2002) for which we have predictions for all three measures, year 2001 has the highest average accuracy.

Figure 4.2: **SVM-prob Average Accuracy with Design Two for All Three Financial Measures.**

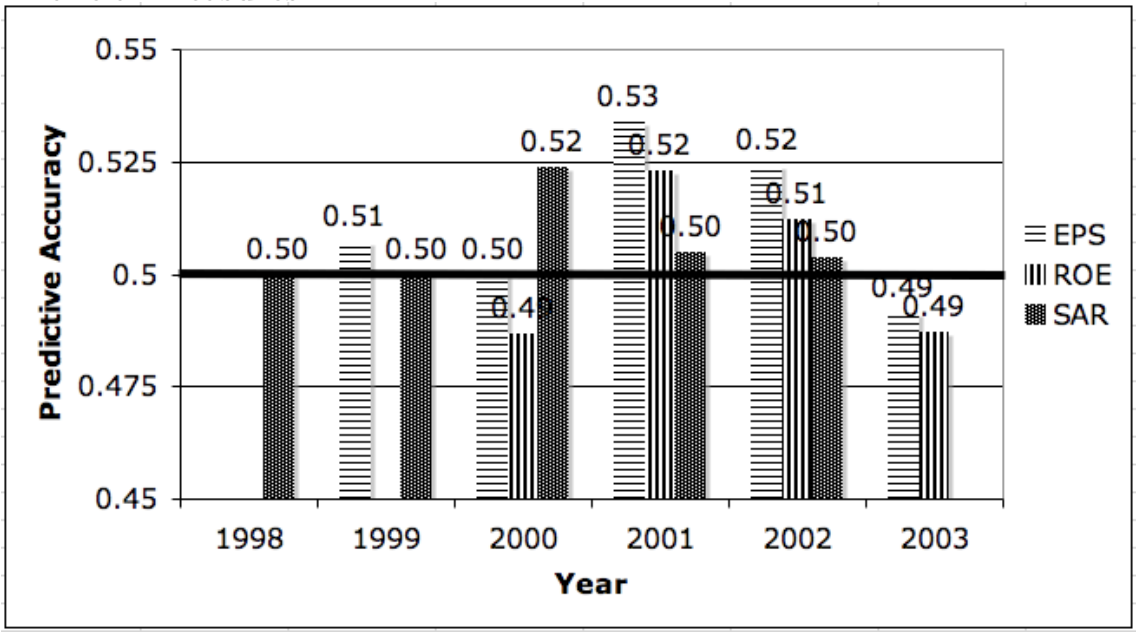


Figure 4.3 presents the cost of errors by SVM-prob model for all measure/year with design two. Although there is no fixed baseline cost of errors, the lower the cost of errors by the predictive models the better. We can see that experiments for

Design Two	$\boldsymbol{\Delta EPS^{f_i}_{(t,t+1)}}$	$\boldsymbol{\Delta ROE^{f_i}_{(t,t+1)}}$	$\boldsymbol{SAR^{f_i}_{(t+1,t+2)}}$
Accuracy Averaged Over Years	0.5112	0.5022	0.5064

Table 4.16: **Average Predictive Accuracy by Financial Measures with Design Two, SVM-prob Model**

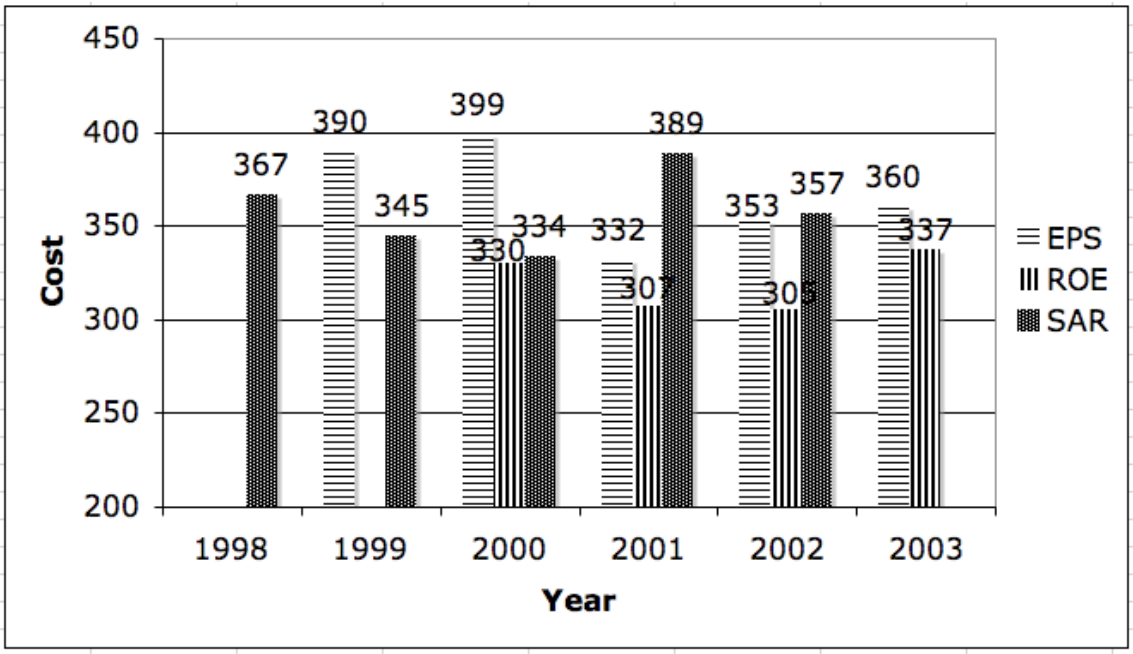
Design Two	**2000**	**2001**	**2002**
Accuracy Averaged Over Measures	0.5034	0.5205	0.5132

Table 4.17: **Average Predictive Accuracy by Years with Design Two, SVM-prob Model**

$\Delta ROE^{f_i}_{(t,t+1)}$ gives on average the lowest cost of errors, and the other two financial measures perform similarly in terms of cost.

This implementable design is an estimate of the model's real-world performance as it uses data from different time periods to build and test the model. With experiment design one, we apply our model to the test data that come from the same time period as the training data. Thus, we assume perfect knowledge of future environment where the model is going to predict. Design two however, tests our model with data from a "future" time period different from where the training data come from. Therefore, there may be factors about temporal changes in the test data that the model may not capture with the training data. We expect that with design two, the same model may perform worse than with design one. The results in Figure 4.2 and Figure 4.1 confirm our expections. However, to further examine SVM-prob model's performance with design two, we employ Kappa statistic, a measure to evaluate inter-rater reliability, to test the agreement between the model prediction and the true class

Figure 4.3: **SVM-prob Average Cost of Errors with Design Two for All Three Financial Measures**



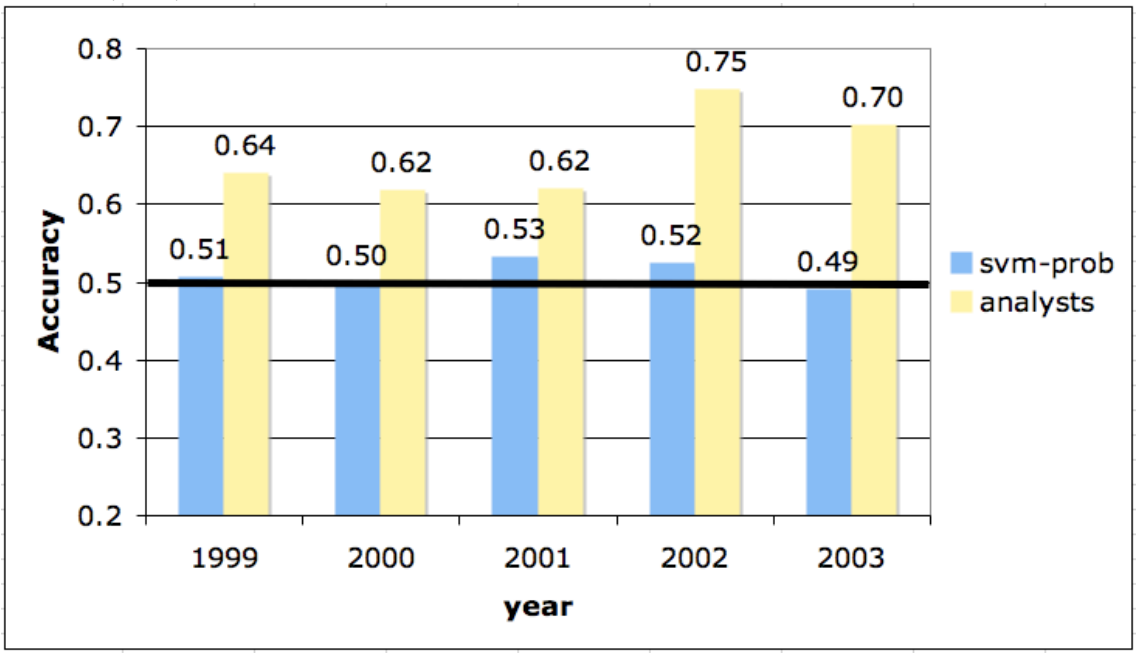
labels of the test data points. We obtain 0.0509 for Kappa statistic and 0.0670 for weighted Kappa statistic, indicating the model's significant agreement with the test data's true distribution at 0.01% level.

4.2.3 Comparing SVM-prob with Analysts Forecast

Besides evaluating our model with predictive accuracy and cost of error, we also have analysts forecast as another benchmark to asses our model's real-world prediction. This comparison with analysts forecast can be conducted for $\Delta EPS^{f_i}_{(t,t+1)}$ and $SAR^{f_i}_{(t+1,t+2)}$ as discussed in Section 3.3.2 and 3.3.3 respectively. Figure 4.4 presents the analysts forecast accuracy and SVM-prob model's accuracy with design two for predicting $\Delta EPS^{f_i}_{(t,t+1)}$. We see that analysts achieve accuracies consistently higher than SVM-prob by at least 20%. This result coincides with another study that suggests that "analysts have consistently been shown to forecast earnings more accurately than do mechanical models" [93]. The reason is most probably that analysts "have

access to more information to project future earnings than the accounting system has to produce the earnings number" [93]. However, as shown in Figure 4.5, analysts stock recommendation as evaluated with predictive accuracy are lower than the SVM-prob model as well as the baseline for all years. The large difference between the analysts stock forecast and the majority vote baseline as well as the model confirms the findings from other studies that analyst recommendations are biased because of the economic incentives faced by sell-side brokerage firms [43].

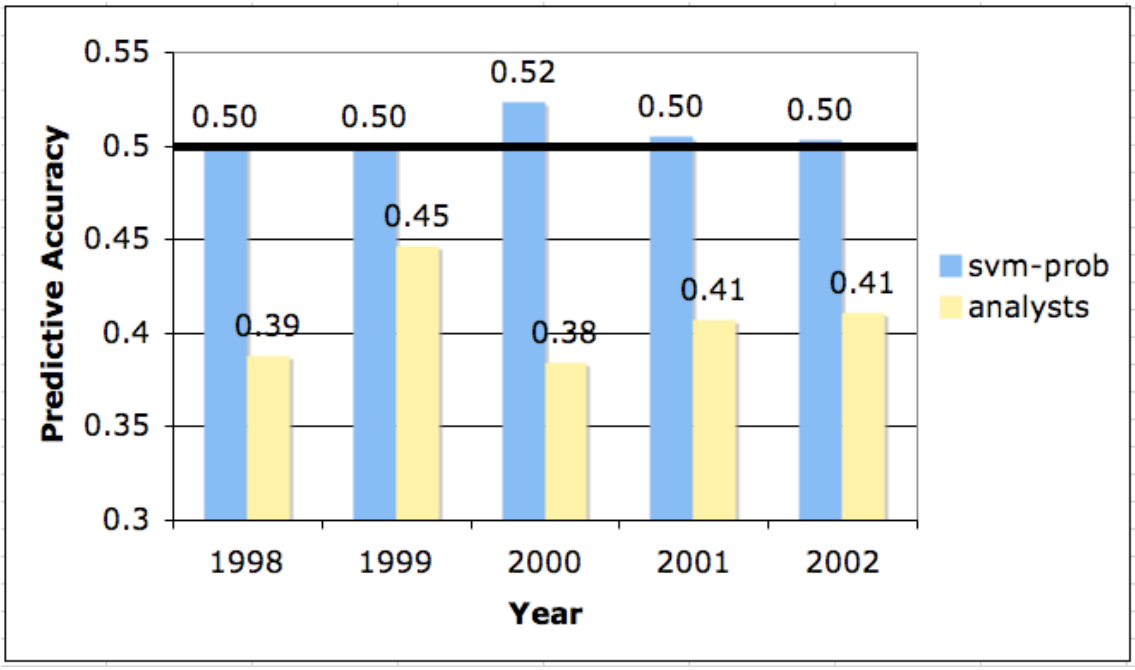
Figure 4.4: **Comparing SVM-prob Accuracy with Design Two for $\Delta EPS^{f_i}_{(t,t+1)}$ and Analysts Forecast on EPS**



4.2.4 Portfolio Return

The mock portfolio is designed to both evaluate the predictive model's cost of error, and test the trading strategy based on model predictions. As detailed in 2.3.2, we construct a portfolio consisting of the stocks of predicted out-performing firms bought with the money from selling the predicted under-performing firms. Then we

Figure 4.5: **Comparing SVM-prob Accuracy with Design Two for $SAR^{f_i}_{(t+1,t+2)}$ and Analysts Stock Recommendation**



sell the stocks of the predicted out-performing and buy the under-performing a year later. The difference in the average $SAR^{f_i}_{(t+1,t+2)}$ of the predicted out-performing and the average $SAR^{f_i}_{(t+1,t+2)}$ of the predicted under-performing firms is the portfolio return. Since $SAR^{f_i}_{(t+1,t+2)}$ is size-adjusted and thus normalized with market return, we hope for a positive portfolio return to prove that the model predictions are catching the excess return and therefore the true out-performers and under-performers. We experiment with both a 25-50-25% class definition and a 10-80-10% class definition for testing model robustness. As discussed in Section 2.5.3, to overcome the problem of skewed data distribution with 10-80-10% class definition, the classification method using 10-80-10% class definition duplicates the minority class data points when training the three binary classifiers. We also experiment with portfolio return for sub-sample firms: profit vs loss firms, large vs small firms, and glamor vs. value firms.

The construction of the portfolio implies that the models need to predict both out-performing firms and under-performing firms, using design two. However, as

discussed in Section 4.2.1, there is an interesting observation that the SVM-prob with SAR did not predict any out-performing firms with design two, and 25-50-25% class definition. The other two models (i.e. SVM-score and SVM-multi) are able to predict all three classes. This observation applies to the prediction on EPS measure and ROE measure with design two as well. In other words, SVM-prob model, although evaluated as the best with predictive accuracy and cost of error, in both design one and design two, does not predict any out-performing firms using design two (i.e. the implementable design).

We also experimented with a 10-80-10% class definition to predict $SAR^{f_i}_{(t+1,t+2)}$. Interestingly, SVM-score is able to give prediction for all three classes, while SVM-multi predicts every data point as "Average", and SVM-prob predicts almost all as "Average". Tables 4.18, 4.19, and 4.20 illustrate this observation.

Design Two	**SVM-score Prediction on $\boldsymbol{SAR^{f_i}_{(t+1,t+2)}}$**			
10-80-10% Class Definition	**Out-Perform**	**Average**	**Under-Perform**	**True Distribution**
True Out-Perform	8.72%	84.08%	7.22%	10.09%
True Average	8.16%	85.91%	5.93%	79.99%
True Under-Perform	9.23%	81.94%	8.83%	9.92%
Prediction Distribution	8.33%	85.32%	6.35%	100.00%

Table 4.18: **Classification Contingency Table of SVM-score for Predicting $\boldsymbol{SAR^{f_i}_{(t+1,t+2)}}$ with 10-80-10% Class Definition**

Given these results comparing the classification rates of the three models, we select SVM-score as it consistently predicts all three classes to construct a mock portfolio. As discussed in Section 2.3, we could use the distribution of analysts stock recommendations to define analysts forecasts for the three classes. Using analysts forecasts, we could also calculate portfolio returns and compare with that of the models as another evaluation method. Table 4.21 demonstrates the results. We are happy to see positive average portfolio returns given by the model as shown in the

Design Two	SVM-prob Prediction on $SAR^{f_i}_{(t+1,t+2)}$			
10-80-10% Class Definition	Out-Perform	Average	Under-Perform	True Distribution
True Out-Perform	0%	100.00%	0.00%	10.09%
True Average	0%	99.86%	0.14%	79.99%
True Under-Perform	0%	100.00%	0%	9.92%
Prediction Distribution	0%	99.89%	0.11%	100.00%

Table 4.19: **Classification Contingency Table of SVM-prob for Predicting $SAR^{f_i}_{(t+1,t+2)}$ with 10-80-10% Class Definition**

Design Two	SVM-multi Prediction on $SAR^{f_i}_{(t+1,t+2)}$			
10-80-10% Class Definition	Out-Perform	Average	Under-Perform	True Distribution
True Out-Perform	0%	100.00%	0%	10.09%
True Average	0%	100.00%	0%	79.99%
True Under-Perform	0%	100.00%	0%	9.92%
Prediction Distribution	0%	100.00%	0%	100.00%

Table 4.20: **Classification Contingency Table of SVM-multi for Predicting $SAR^{f_i}_{(t+1,t+2)}$ with 10-80-10% Class Definition**

last row. Our model's returns are also higher on average than those of the analysts. This implies that our predictive model is able to identify firms that perform better or worse than the market average. We also notice that year 2000 is the worst year for generating return, for both the predictive models and the analysts, and for both class definitions. One reason could be the considerable turbulence experienced by financial markets in year 2000 [5].

4.2.4.1 Sub-sample Portfolio Return

Firms' cross-sectional differences such as size, profitability, and growth potential could influence the informativeness of narrative disclosures [9]. We subdivide the data used for $SAR^{f_i}_{(t+1,t+2)}$ into sub samples of large firms vs small firms, profit vs loss firms,

	With 25-50-25% Definition		With 10-80-10% Definition	
Year	**SVM-score**	**Analysts**	**SVM-score**	**Analysts**
1998	4.27%	7.31%	5.09%	2.43%
1999	44.50%	47.69%	56.37%	58.27%
2000	-38.39%	-21.38%	-43.44%	-9.80%
2001	18.82%	-7.70%	14.15%	-9.51%
2002	15.69%	-4.75%	10.10%	-7.26%
Average	8.98 %	4.14%	8.45%	6.83%

Table 4.21: **Portfolio Return by SVM-score Model and Analysts Recommendation with Different Class Definitions**

and glamor vs value firms as discussed in Section 2.3.3, and reconstruct portfolio for each sub sample data set. Our goal is to verify if the model can pick up the differences in information content between sub sample firms. We continue to use the SVM-score model for portfolio construction for the reasons discussed above.

First we present the classification contingency tables as in Table 4.22 of SVM-score model with design two for each sub sample set. Again, we focus on the true predictive rates of the out-performing and under-performing firms, and the misclassification rates of out-performing to under-performing and vice versa. We observe that glamour firms have higher true predictive rate and lower false misclassification rate for out-performing firms (22.75% and 10.41%) than value firms (10.21% and 19.05%). These differences suggest that glamour firms may focus more on the upside in their narrative disclosure than the value firms. In contrast, glamour firms have higher misclassification rate and lower true predictive rate for under-performing firms (22.96% and 15.43%) than value firms (15.52% and 23.37%). This indicates that the value firms may relate more to identifying downside risks in their narrative discussion.

Big firms seem to have higher true predictive rate for both out-performing

firms and under-performing firms (21.36% and 24.04%) than small firms (14.89% and 15.18%). The misclassification rates of big firms are similar to those of the small firms. This suggests that big firms seem to provide more comprehensive and accurate information than small firms. With profit firms and loss firms, we observe that predicting out-performing firms seems to give similar true predictive and misclassification rates. In predicting under-performing firms, the profit firms seem to have higher true predictive and misclassification rates than the loss firms. These suggest that the profit firms and loss firms do not seem to significantly differ in presenting information on their upsides and downsides.

Next, we present the average portfolio returns by sub sample firms based on the predictions given by SVM-score model. Firms partition is done with previous year's criteria value. For example, firms in 1998 are determined by their market-to-book ratio in 1997 as glamour firms or value firms. Then predictive models are built with these firms' 1997 documents and 1998 financial performance. Models are then applied to 1998 documents to predict 1999 performance. Since we have available to us data about market-to-book ratio, assets, and EPS from 1997 to 2002, and we use experiment design two (i.e. the implementable design) to predict and construct portfolio, the sub sample portfolio results range from 1999 to 2002. Table 4.23 presents the results. We are happy to see distinctive differences in average returns between glamour and value firms (24.95% vs -3.99%), small and big firms (11.65% vs 1.03%), and profit and loss firms (0.199% vs -16.78%). This further confirms that the information content as disclosed in annual reports differs between firms of different cross-sectional features. This difference in information content can be successfully and automatically detected by predictive models built with textual data only.

Design Two		SVM-score Prediction on $SAR^{f_i}_{(t+1,t+2)}$			
25-50-25% Class Definition	**True Distribution**	**Out-Perform**	**Average**	**Under-Perform**	**Total No. of Firm/Year**
Glamour Firm	Out-perform (25%)	22.75%	66.85%	10.41%	372
	Average (50%)	21.58%	67.23%	11.19%	737
	Under-perform (25%)	22.96%	61.59%	15.43%	364
Value Firm	Out-perform (25%)	10.21%	70.74%	19.05%	370
	Average (50%)	13.73%	73.02%	13.25%	737
	Under-perform (25%)	15.52%	61.12%	23.37%	364
Big Firm	Out-perform (25%)	21.36%	59.76%	18.86%	372
	Average (50%)	19.86%	63.31%	16.85%	737
	Under-perform (25%)	23.39%	52.58%	24.04%	364
Small Firm	Out-perform (25%)	14.89%	69.20%	15.90%	371
	Average (50%)	19.36%	66.15%	14.49%	737
	Under-perform (25%)	20.57%	64.27%	15.18%	364
Profit Firm	Out-perform (25%)	16.53%	68.41%	15.06%	584
	Average (50%)	19.56%	64.36%	16.09%	1158
	Under-perform (25%)	27.83%	49.18%	23.00%	364
Loss Firm	Out-perform (25%)	16.55%	71.91%	11.56%	200
	Average (50%)	12.38%	77.57%	10.05%	396
	Under-perform (25%)	9.05%	80.71%	10.23%	194

Table 4.22: **Classification Contingency Table of SVM-score for Predicting $SAR^{f_i}_{(t+1,t+2)}$ with 25-50-25% Class Definition of Sub Sample Firms**

Year	Glamour Firm	Value Firm	Big Firm	Small Firm	Profit Firm	Loss Firm
1999	111.61%	-0.23%	5.49%	83.25%	27.86%	-74.63%
2000	-45.52%	-24.07%	-23.86%	-42.44%	-38.74%	-14.25%
2001	18.61%	-4.08%	14.51%	-1.14%	6.57%	13.33%
2002	15.12%	12.42%	7.99%	6.92%	5.09%	8.08%
Average	24.96%	-3.99%	1.03%	11.65%	0.199%	-16.87%

Table 4.23: **Portfolio Return of Sub Sample Firms Based on Prediction with SVM-score Model and 25-50-25% Class Definition**

CHAPTER V
ANALYSIS

5.1 Model Performance

Our research goal is to assess the potential of building predictive models using the textual content of annual reports. To achieve this goal, we test three models (i.e. SVM-score, SVM-prob and SVM-multi) using different evaluation benchmarks: predictive accuracy, cost of errors, comparison with majority vote baseline and analysts forecasts, portfolio return, and robustness with different class definitions. There are interesting observations to make about these three models.

The SVM-Prob model seems to stand out as the overall best model when evaluated with predictive accuracy and cost of errors, against the baseline. As discussed in Section 4.1, with the cross-validation experiments (i.e. design one), SVM-Prob in most cases of measure/year experiments has more balanced trade-off between true predictive rates and error rates for predicting out-performing and under-performing firms. SVM-Score risks the highest error rates to achieve the highest true predictive rates, in predicting out-performing and under-performing. SVM-Multi is the most conservative that sacrifices true predictive rates for lowest error rates. On the other hand, with the implementable design (i.e. Design Two), SVM-Prob seems to be the most conservative among the three by predicting only "average" firms and under-performing firms.

With design two, models are trained on data from a time period different from that of the test data. There may be some factors such as temporal change in economic status that is reflected in the test data but not in the training data. We expect that design two experiments are harder to achieve good predictive results than design one. It is interesting to see that SVM-Prob, while still the best by predictive accuracy

and cost of errors, does not provide predictions for out-performing firms. Interestingly, while SVM-Score is ranked the third or the worst among all three models for both designs evaluated with predictive accuracy and cost of errors, it is the only one suitable for building mock portfolio by predicting all three classes with design two. As discussed in Section 2.5.3, the three models (SVM-score, SVM-prob, and SVM-multi) differ in their approaches to give multi-class predictions. SVM-score uses the highest score from the three binary classifiers to give a final class label to a document. SVM-prob transfers a binary classifier's scores to probabilities of belonging to the positive class, and then assigns a class label based on the highest probability. SVM-multi formulates the multi-class categorization problem as a constrained optimization problem with a quadratic objective function that yields a direct method for multi-class prediction. All three models are designed to achieve as many correct predictions as possible. It appears that SVM-score is able to predict all three classes with a relatively simpler design that bases its final decision on the crude scores, even though the overall accuracy is not high. There are questions to answer as to why SVM-prob is not able to predict out-performing, and why out-performing instead of perhaps under-performing. We leave these questions for future exploration.

We also observe that year 1999 is the easiest year to predict with design one and 2001 is the easiest with design two. The tentative explanation is that 1999 is the year when the market reached its peak before the financial turbulence happened in 2000. We speculate that with the apparent thriving economic development, the annual reports may be rich in information that facilitates a more accurate prediction with a design that utilizes the data from the same time period for model training and testing. Predicting 2001 with design one means that we use the reports of 1999 and financial data in 2000 to build the predictive model, apply the model to reports of 2000, and predict 2001 financial performance. Possibly, since both 1999 and 2000 are years of considerable financial development or turbulence, the reports of 1999 and

2000 reflect abundant and contrasting information that makes the implementable design easier to predict even if the model is built with data of an unseen time period.

We find that $\Delta ROE^{f_i}_{(t,t+1)}$ is the hardest to predict in both design one and design two. The possible reason that it does not possess good predictive power could be because the definition of ROE itself has many variations and implications for evaluating firms performance. Exploring other forms to formulate ROE could be of interest to our future study.

5.2 Post Hoc Analysis of Different Classes of Firms as Predicted by Model

As presented in Section 4.2.4, we design a mock portfolio experiment to test the models' cost of errors as well as the trading strategy based on model predictions. We see from Table 4.21 that the portfolios composed of firms based on model predictions are able to provide positive average returns over years, with 2 types of class definitions. We examined the numerical details of all the firms in this portfolio experiment to examine the characteristics of the three classes of firms as predicted by the model. Tables 5.1 and 5.2 demonstrate that firms predicted as belonging to three classes show distinctive difference in terms of their financial information. Using Table 5.1 with 25-50-25% class definition as an example, we find that relative to the predicted out-performing firms, the predicted under-performing firms are smaller (as measured with assets), have lower market-to-book ratios, sales, price momentum, and SAR, and are less profitable (as measured with EPS). "Average" firms have the largest assets and sales, and are the most profitable. Similar observations can be made of these firms when experiments with 10-80-10% class definition, as shown in Table 5.2. These clear distinctions among the three groups of firms predicted by textual model provide additional evidence about the information content of disclosures. Since our model is built with textual features, we believe the text in the annual reports is

enough to identify distinctive groups of under- and out-performing firms.

	Firm/Year From $SAR^{f_i}_{(t+1,t+2)}$ Design Two Experiment (With 25-50-25% Class Definition)		
	Predicted Under-Perform Mean (Median)	**Predicted Average-Perform Mean (Median)**	**Predicted Out-Perform Mean (Median)**
Assets ($ million)	832.96 (144.21)	4358.87 (634.51)	2441.23 (336.97)
Sales ($ million)	774.45 (81.46)	3701.34 (645.4)	2103.29 (281.58)
EPS	-0.3111 (-0.2475)	0.8122 (0.82)	0.6049 (0.6661)
Market-to-Book	1.6141 (1.0134)	1.7113 (0.9988)	3.0409 (1.9556)
Leverage	0.4150 (0.3617)	0.5136 (0.5330)	0.4203 (0.3929)
Earnings Surprise	-0.1419 (-0.175)	-0.2563 (-0.1)	-0.1145 (-0.0056)
Price Momentum	-0.3163 (-0.3787)	-0.0238 (-0.0697)	0.6633 (0.3153)
Size Adjusted Return (annual)	-0.0522 (-0.1777)	0.0087 (-0.0926)	0.0234 (-0.1251)

Table 5.1: **All Firms' Characteristics Based on SVM-score Model Prediction for $SAR^{f_i}_{(t+1,t+2)}$ with 25-50-25% Class Definition**

5.3 Cross-sectional Regression

As discussed in Section 4.2.4.1, we identify certain factors such as total assets, market-to-book ratio, and EPS value that influence firms' information environment. We divide firms into sub-samples such as big vs small, glamour vs value, and profit vs loss firms. The portfolio returns based on model predictions show distinctive differences among firms in the sub samples. These difference suggest that the predictive models identify greater value-relevant information in the disclosures made by glamour firms and by small firms.

	Firm/Year From $SAR^{f_i}_{(t+1,t+2)}$ Design Two Experiment (With 10-80-10% Class Definition)		
	Predicted Under-Perform Mean (Median)	**Predicted Average-Perform Mean (Median)**	**Predicted Out-Perform Mean (Median)**
Assets ($ million)	703.40 (123.47)	3669.63 (451.43)	1825.39 (317.61)
Sales ($ million)	538.60 (77.84)	3131.42 (408.53)	1656.08 (225.55)
EPS	-0.3991 (-0.31)	0.6283 (0.655)	0.5919 (0.58)
Market-to-Book	1.3923 (0.7626)	1.8552 (1.0935)	3.3298 (2.2887)
Leverage	0.4406 (0.3964)	0.4881 (0.4929)	0.3908 (0.3414)
Earnings Surprise	0.2367 (-0.1933)	-0.2602 (-0.09)	-0.0004 (0)
Price Momentum	-0.4063 (-0.4899)	0.0083 (-0.0679)	0.8578 (0.4950)
Size Adjusted Return (annual)	-0.0309 (-0.1438)	0.0004 (-0.1069)	0.0267 (-0.1464)

Table 5.2: **All Firms' Characteristics Based on SVM-score Model Prediction for $SAR^{f_i}_{(t+1,t+2)}$ with 10-80-10% Class Definition**

To further explore the relationship between the incremental information content of the annual reports and the size-adjusted cumulative return, we employ cross-sectional regression to control some factors that affect returns. We regress SAR on model prediction, firm size, market-to-book ratio, price momentum and earnings surprise. The regression formula examines the effect of the narrative content on size-adjusted return and its interaction effect with other numeric estimates. We are interested in whether the information in the narrative disclosure is absorbed by or is incremental to the information in the quantitative disclosure. The formal definition

is:

$$\begin{aligned} SAR =& \alpha + \beta_1 Dummy + \beta_2 Size + \beta_3 MTB + \beta_4 PM + \beta_5 Surprise + \\ & \beta_6 Size \times Dummy + \beta_7 MTB \times Dummy + \beta_8 PM \times Dummy + \\ & \beta_9 Surprise \times Dummy + error \end{aligned}$$

where

SAR = Size adjusted buy-and hold return for the year

Dummy = 1 if the firm is classified as out-perform and 0 for predicted under-performing firms

Size = The size of the firm, measured as the natural logarithm of total assets

MTB = The natural log of market to book ratio (a valuation proxy), using the closing market price as of the start of the holding period

PM = Price momentum, measured as the SAR for the six months preceding the start of the holding period

Earnings Surprise = Actual EPS - Forecast EPS, where the forecast is the latest available consensus analyst forecast

With this formulation, a positive β_1 indicates that the narrative disclosure provides incremental value-relevant information. A significant non-zero interaction term implies that the textual content alters the investors' confidence in the numeric estimates. We obtain the results as in Table 5.3. The regression model 1 examines the main effects of model prediction and other controlling factors. We observe that the coefficient for the model prediction is significantly positive, indicating value-relevant information incremental to that provided by known factors. Consistent with previous studies, we find that larger firms earn smaller returns [80], and that higher market-to-book ratio predicts lower returns [25]. However, our data do not show the expected positive relation between price momentum and returns [44].

The regression model 2 is a complete model that includes the interaction of

model prediction with the other numerical factors. We find that the main effect of the model prediction is no longer significant. However, the interaction of the prediction with price momentum indicates that the narrative disclosure reinforces the market's confidence in price momentum disclosure and the implication of price momentum for returns.

	Regression Model 1		**Regression Model 2**	
Item	**Estimate**	**t-value**	**Estimate**	**t-value**
Intercept	0.12910	1.64	0.15807	1.26
Dummy for model prediction	0.08588	1.85*	-0.00311	-0.02
Log(Total Assets)	-0.02561	-1.99*	-0.02586	-1.27
Log(Market-to-book)	-0.00928	-1.41	-0.03026	-1.75*
Earnings Surprise	-0.00226	-0.55	0.00256	0.45
Price Momentum	-0.05776	-2.39**	-0.01257	-0.48
Dummy × log(Total Assets)			0.00391	0.15
Dummy × log(Market-to-book)			0.03287	1.75
Dummy × earnings surprise			-0.01004	-1.22
Dummy × price momentum			-0.29054	-4.20***
N		1,108		1,108
Adjusted R-square		0.009		0.023
F-Value		3.01**		3.93***

*Note: * is significant at 1%; ** at 0.1%; *** at 0.01%*

Table 5.3: **Cross-sectional Regression Results**

5.4 Textual Feature Analysis

With the results and analysis presented above and in Chapter 4, we confirm that the narrative disclosure provides additional incremental information about firms'

financial performance. We are interested in further exploring the textual features in the annual reports to study what kind of textual features contribute to predicting financial performance. We hope to be able to explain how the language in annual reports predict future performance. Our document representation method is the "bag of words" approach with document frequency thresholding to filter out less informative words. There are many options to further analyze the syntactic and semantic features of these words, such as restricting to noun phrases, examining the verbal tones, or using terms clustered as meta-features. As an exploratory step, we experiment with using a set of predefined verbal tones to examine the potential of relating the semantics of the documents with future financial performance. We leave the systematic analysis of the textual features for future work.

Diction is a software package that examines a text for its verbal tones [35]. We select the 31 dictionaries that are defined within Diction to categorize the terms in the annual reports into 31 semantic groups. These 31 lists of words describe the following semantic types: accomplishment, aggression, ambivalence, blame, centrality, cognition, collectives, communication, concreteness, cooperation, denial, diversity, exclusion, familiarity, hardship, human interest, inspiration, leveling, liberation, motion, numerical, passivity, past concern, praise, present concern, rapport, satisfaction, self reference, spatial terms, temporal terms, and tenacity. There are a total of 9293 terms in these 31 dictionaries. 750 of the terms appear in more than one dictionary, resulting in a total of 8543 unique terms. The definition of each dictionary is provided in Appendix E along with sample words.

We test the usage of the dictionary terms with SVM-Prob model predicting $\Delta EPS^{f_i}_{(t,t+1)}$ using design one (i.e. the cross-validation design) for year 2002. We experiment with two alternatives. First we replace each dictionary word in the annual report with its dictionary name. Thus each document is represented as a vector of terms that include both the words not in the dictionaries and the dictionary names.

In other words, we replace part of the annual report words with their higher-level semantic types. We call this model SVM-Prob-Combine-Diction. In the second approach, we keep only the words that appear in the 31 dictionaries and replace them with their dictionary names. Therefore, each annual report is represented as a vector of terms that are generalized to a higher-level semantic types. We name this model SVM-Prob-Diction-Only. We compare the models augmented with dictionary words with the original model (SVM-Prob) built with documents without dictionary replacement.

Table 5.4 shows the results. We find that replacing words with their semantic meanings does not change significantly the original model's predictive performance, even though the augmented model still performs better than the baseline. Using only words from the dictionaries and abstracting the words to their semantic meanings deteriorates the original model's performance compared to that of the majority-vote baseline. More experiments that examine all the years for all three measure and both design one and two could be helpful in making a more general observation on our current observation. In addition, exploring other ways to construct the dictionaries may be helpful in studying the relation between textual reports and firms performance at the semantic level. For example, a dictionary or thesaurus of accounting terms may be of more relevance and value in terms of providing insights on the use of domain-specific language.

	SVM-Prob-Combine-Diction	**SVM-Prob-Diction-Only**	**SVM-Prob**	**Majority-Vote Baseline**
Average accuracy	0.5533	0.4962	0.5502	0.5
P-value with SVM-Prob	0.17	***0.0052***	–	–
P-value with Baseline	***0.0123***	0.2986	***0.0144***	–

Table 5.4: **Comparing the Models Built with Dictionary Words: Predicting $\Delta EPS^{f_i}_{(t,t+1)}$ for 2002**

CHAPTER VI
CONCLUSIONS AND FUTURE WORK

Text classification methods have been applied to the business domain for discovering relationships or extracting opinions from resources such as news releases and web pages. Company annual reports, one of the most important sources of disclosures, have remained largely untapped by the machine learning and text mining community. Previous studies have shown that the narrative discussions in annual reports are important for assessing firm performance. However, researchers have mainly exploited numerical data to build models to forecast firm performance. Set in this background, our research goal is to study the potential of building predictive models from annual reports. We apply text classification method to annual reports and predict company future financial performance. We shape our research problem along five key dimensions. These involve measures of firm performance; performance aspects to forecast; model evaluation methods; document representations; and experiment designs. We discuss options and our selections along each dimension. Part of our intent in these discussions has been to demonstrate the complexity and challenges involved in problem definition when dealing with research that derives from real-world goals. These discussions also outline strategies for future work as for example, with a different experiment design, with different measures of financial performance or with an alternative set of company reports. We believe our research contributes to the future study in these dimensions as well as to other research that considers real-world applications in the business and finance domains.

We explore three financial performance measures, two of which ($\Delta EPS^{f_i}_{(t,t+1)}$, $\Delta ROE^{f_i}_{(t,t+1)}$) use accounting measures, and one is a market response measure ($SAR^{f_i}_{(t+1,t+2)}$). We observe that with design one and SVM-prob, our best predictive model, SAR has the best predictability in terms of accuracy, followed by EPS and then ROE. It appears

that all three SVM models (SVM-prob, SVM-score and SVM-multi) have difficulty in giving better-than-baseline predictions for ROE, compared with the other two measures. With design two, $\Delta ROE^{f_i}_{(t,t+1)}$ again is the hardest to predict with the lowest average accuracy across years. $\Delta EPS^{f_i}_{(t,t+1)}$ is the easiest to predict with the highest average predictive accuracy over years, although the differences in average accuracy are very small compared with the other two financial measures.

Based on our observations of the data distribution, we decide to formulate our research problem as a three-class classification problem. The choice is based on several reasons. As an exploratory study, we would like to pursue a coarse-grained category prediction as a first step instead of a real value prediction. Given the near normal data distribution, an odd number of classes would be appropriate. Three would be the minimum. In addition, as a real-world application from the operation perspective, we are most interested in the two ends that are the out and under-performing firms. Further, the larger the number of classes to predict, the smaller the training set will be for the binary classifiers and thus the harder the classification problem. Since the firms are ordered in each year for each measure, in future research, this prediction problem could also be formulated as a ranking problem where we are interested in finding the top and bottom ranking firms. We test our model's robustness with a 10-80-10% class definition for the portfolio return evaluation. The results are comparable with that of 25-50-25% class definition.

We find that of our three models, SVM-Prob performs the best with both experiment design one and two, when evaluated with predictive accuracy and cost of errors. More specifically, with design one, SVM-Prob manages a more balanced tradeoff between achieving higher accuracy and lower cost of error. With design two, SVM-Prob is the most conservative by giving a large portion of predictions to the "average class" and not predicting the "out-performing" class at all. However, this prediction pattern renders SVM-Prob useless for building portfolios or suggesting trading strategies.

SVM-score, an obvious risk-taker among the three and ranked the worst in terms of accuracy and cost, turns out to be the best candidate in providing predictions for portfolio construction. Further exploration is needed to study why SVM-Prob is unable to predict out-performing firms with design two, and why avoiding out-performing but not under-performing as both are minority class compared with the "average" class. A reasonable approach to start with is by examining the plots of the probabilities for the three binary classes generated by the SVM-Prob model.

Overall our results confirm that models can be successfully built using the textual content of annual reports to predict future performance. We find that with design one, our model is able to perform significantly better than the majority vote baseline in 12 out of 17 (71%) sets of financial measure/year predictions. The design one experiment is designed to use contemporaneous data to build and test the model. This design assumes perfect information about the future economic environment. The design two is an implementable simulation with historical data that may give us a more accurate estimation of a model's real-world performance. As expected, our model's performance deteriorate relative to design one. However, to further examine SVM-prob model's performance with design two, we employ Kappa statistic, a measure to evaluate inter-rater reliability, to test the agreement between the model prediction and the true class labels of the test data points. The Kappa statistics show significant agreement between model predictions on the test data's class labels and the test data's true class labels.

We use analysts EPS forecasts and stock recommendations as another evaluation benchmark for our model. We find that analysts' forecasts on EPS are significantly and consistently better than both the majority vote baseline and our model's prediction. On the other hand, analysts' predictive accuracies based on their stock recommendation are worse than baseline. These findings are consistent with previous studies that speculate on analysts having important information sources for accurate

EPS forecast and biasing in stock recommendation because of management incentives. Future study could be done to see if predictions from models built with the narrative disclosure from annual reports may be effectively combined with the analysts' forecasts. This in turn will shed light on the nature of the analysts' information sources.

We represent the annual reports by selecting all the words from the documents. We apply document frequency threshold to filter out less informative words. This "bag of words" approach has be shown in other studies to be effective in text classification. We could explore more complex term definitions that incorporate phrases or syntactic features. However, it could be more important to study what are the content-rich features that capture the semantics most related to predictions. We experiment using 31 dictionaries to complement our document representation. The scale of our experiments is small and the findings are not encouraging at current stage. It is of great interest and value to explore alternate ways to study the semantic attributes and gain insights into what characteristics of the narrative disclosure lead to certain predictions.

We confirm with our study that the narrative discussion of the annual reports contain information that add value to the numerical estimates of financial performance. Post hoc analysis on the firms of different classes as predicted by the textual model shows distinctive financial differences among firms. Sub-sample analysis also indicates that firms of different cross-sectional features provide information disclosure differently and the textual differences could be captured by our model. Predictive models can be built with the narrative disclosure to detect and utilize the incremental information for financial forecast. One limitation is that this study employs only annual reports as the source of narrative information. To improve our model's predictive performance, other information sources can be added that may capture the temporal change in economic condition. These sources include news releases from the Federal

Reserve, sector-specific forecasts by trade associations, quarterly reports, company press releases, and business news, among many others. Further, our models could by refined by incorporating other parameters such as economic forecasts, and industry and product-life cycle. Alternatively, predictions from our text-based models could be combined with numeric models for possibly better predictive accuracy.

To close, we present research that applies text classification techniques to accounting reports. The scope of future study along this direction is considerable. We see a rich set of follow-up research questions with the promise of further insight in this research area.

APPENDIX A
EXAMPLES OF DATA DISTRIBUTION

Figure A.1: **Histogram for ROE Measure ($\Delta ROE^{f_i}_{(t,t+1)}$) for 2002**

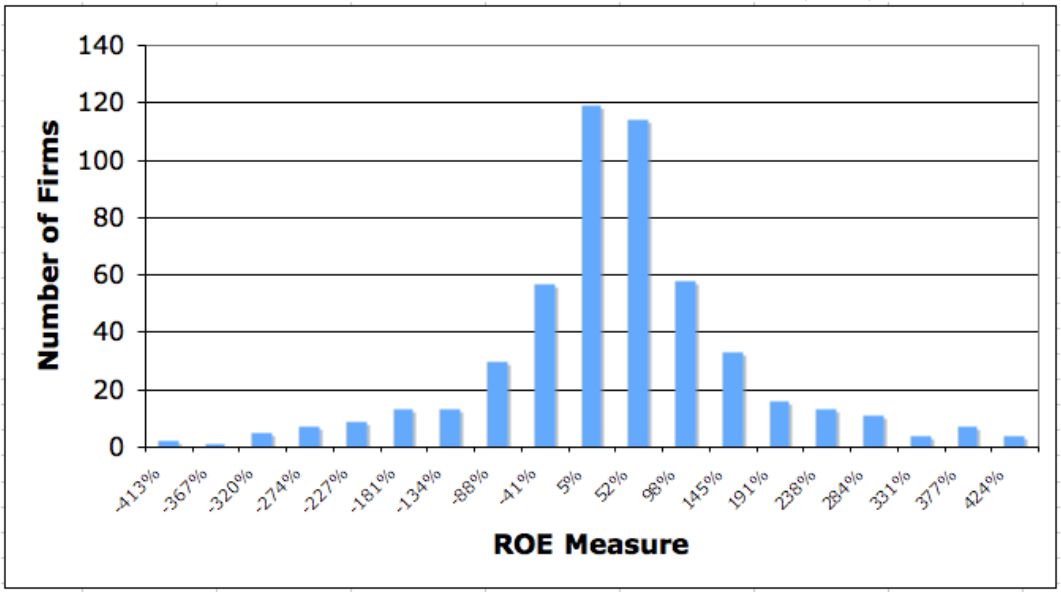


Figure A.2: **Histogram for ROE Measure ($\Delta ROE^{f_i}_{(t,t+1)}$) for 2003**

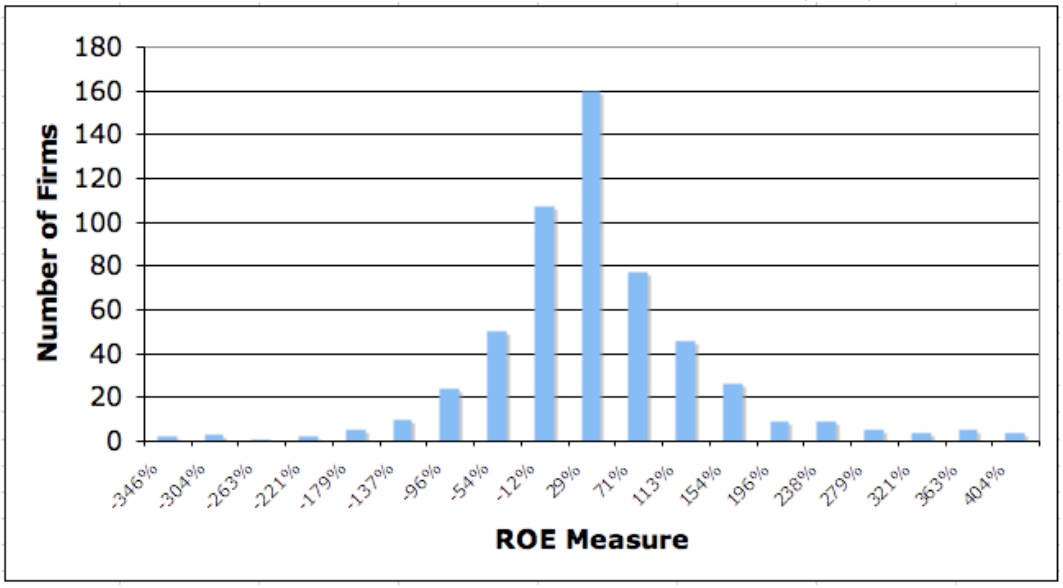


Figure A.3: **Histogram for SAR Measure ($SAR^{f_i}_{(t+1,t+2)}$) for 1997**

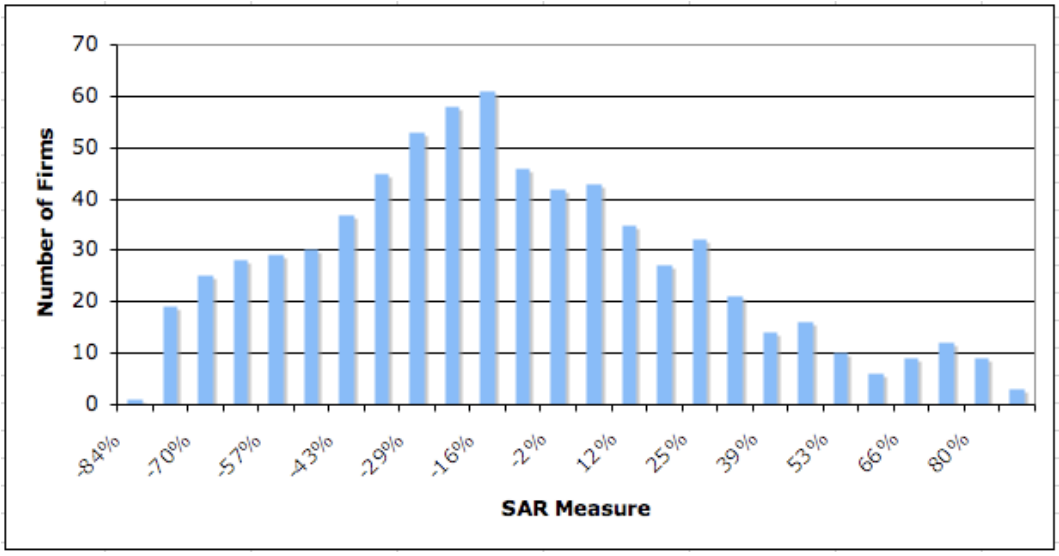
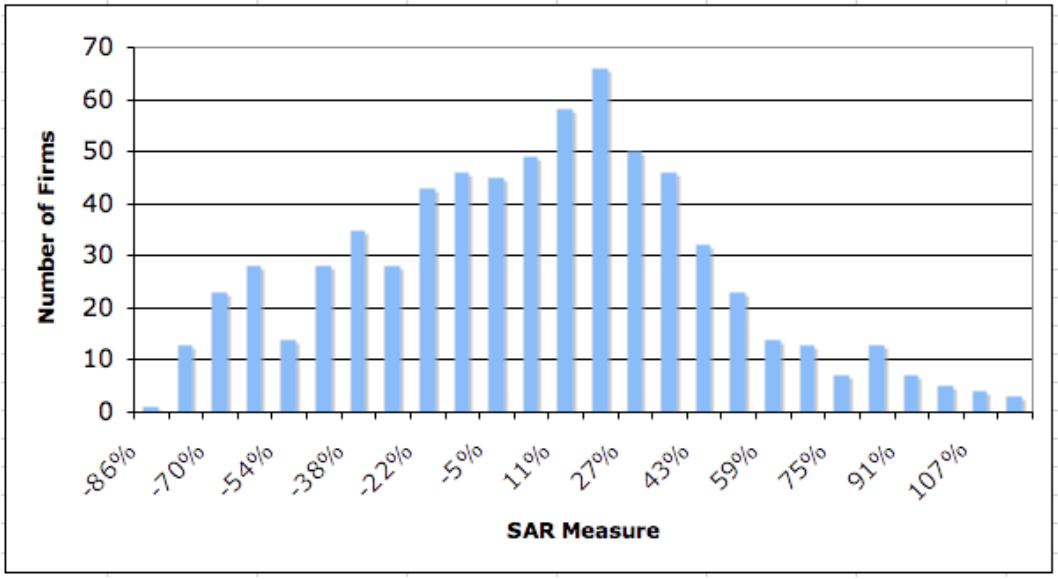


Figure A.4: **Histogram for SAR Measure ($SAR^{f_i}_{(t+1,t+2)}$) for 2001**



APPENDIX B
CLASS DEFINITION OF DATA FOR SUB-SAMPLE EXPERIMENTS

Year (Total Sample Size, Size After Cutoff)	Before 2% Cutoff Each Tail				After 2% Cutoff Each Tail					
	Bottom 2%		Top 2%		Bottom 25%		Middle 50%		Top 25%	
	SAR Measure Threshold	Samples Lost	SAR Measure Threshold	Samples Lost	SAR Measure Threshold	Sample Size	SAR Measure Threshold	Sample Size	SAR Measure Threshold	Sample Size
1998 (531, 510)	<-70.02%	10	>118.22%	11	<-39.65%	127	>=-39.65%, <2.72%	255	>=2.72%	128
1999 (488, 469)	<-124.65%	9	>506.3%	10	<-66.36%	117	>=-66.36%, <-1.34%	234	>=-1.34%	118
2000 (490, 471)	<-65.93%	9	>120.19%	10	<-21.79%	117	>=-21.79%, <41.72%	235	>=41.72%	119
2001 (487, 468)	<-76.15%	9	>179.07%	10	<-17.25%	116	>=-17.25%, <30.75%	234	>=30.75%	118
2002 (418, 401)	<-59.01%	8	>50.25%	9	<-19.21%	100	>=-19.21%, <14%	200	>=14%	101

Figure B.1: **Class Definition with $SAR^{f_i}_{(t+1,t+2)}$ for Profit Firms**

Year (Total Sample Size, Size After Cutoff)	Before 2% Cutoff Each Tail				After 2% Cutoff Each Tail					
	Bottom 2%		Top 2%		Bottom 25%		Middle 50%		Top 25%	
	SAR Measure Threshold	Samples Lost	SAR Measure Threshold	Samples Lost	SAR Measure Threshold	Sample Size	SAR Measure Threshold	Sample Size	SAR Measure Threshold	Sample Size
1998 (136, 131)	<-67.76%	2	>132.21%	3	<-40.15%	32	>=-40.15%, <13.82%	66	>=13.82%	33
1999 (156, 149)	<-123.54%	3	>955.46%	4	<-29.13%	37	>=-29.13%, <245.9%	74	>=-245.9%	38
2000 (144, 139)	<-79.16%	2	>89.46%	3	<-50.45%	34	>=-50.45%, <8.38%	70	>=8.38%	35
2001 (149, 144)	<-97.1%	2	>218.06%	3	<-57.91%	35	>=-57.91%, <14.48%	72	>=14.48%	37
2002 (236, 227)	<-80.17%	4	>86.76%	5	<-50.3%	56	>=-19.21%, <14%	114	>=-6.64%	57

Figure B.2: **Class Definition with $SAR^{f_i}_{(t+1,t+2)}$ for Loss Firms**

Year (Total Sample Size, Size After Cutoff)	Before 2% Cutoff Each Tail				After 2% Cutoff Each Tail					
	Bottom 2%		Top 2%		Bottom 25%		Middle 50%		Top 25%	
	SAR Measure Threshold	Samples Lost	SAR Measure Threshold	Samples Lost	SAR Measure Threshold	Sample Size	SAR Measure Threshold	Sample Size	SAR Measure Threshold	Sample Size
1998 (309, 296)	<-69.89%	6	>67.15%	7	<-39.77%	73	>=-39.77%, <0.093%	148	>=0.093%	75
1999 (303, 290)	<-106.86%	6	>245.9%	7	<-57.29%	72	>=-57.29%, <-9.05%	145	>=-9.05%	73
2000 (309, 296)	<-62.84%	6	>121.16%	7	<-3.94%	73	>=-3.94%, <43.83%	148	>=43.83%	75
2001 (298, 287)	<-68.06%	5	>105.93%	6	<-8.24%	71	>=-8.24%, <28.79%	144	>=28.79%	72
2002 (317, 304)	<-55.61%	6	>43.56%	7	<-20.19%	75	>=-20.19%, <11.74%	152	>=11.74%	77

Figure B.3: **Class Definition with $SAR^{f_i}_{(t+1,t+2)}$ for Large Firms**

Year (Total Sample Size, Size After Cutoff)	Before 2% Cutoff Each Tail				After 2% Cutoff Each Tail					
	Bottom 2%		Top 2%		Bottom 25%		Middle 50%		Top 25%	
	Return Threshold	Samples Lost	Return Threshold	Samples Lost	Return Threshold	Sample Size	Return Threshold	Sample Size	Return Threshold	Sample Size
1998 (308, 295)	<-66.58%	6	>162.21%	7	<-38.88%	73	>=-38.88%, <9.17%	148	>=9.17%	74
1999 (303, 290)	<-131.02%	6	>881.37%	7	<-68.1%	72	>=-68.1%, <124.2%	145	>=124.2%	73
2000 (309, 296)	<-76.26%	6	>104.82%	7	<-42.49%	73	>=-42.49%, <18.87%	148	>=18.87%	75
2001 (298, 287)	<-93.48%	5	>255.56%	6	<-39.17%	71	>=-39.17%, <30.88%	144	>=30.88%	72
2002 (317, 304)	<-73.41%	6	>77.17%	7	<-42%	75	>=-42%, <6.97%	152	>=6.97%	77

Figure B.4: **Class Definition with $SAR^{f_i}_{(t+1,t+2)}$ for Small Firms**

Year (Total Sample Size, Size After Cutoff)	Before 2% Cutoff Each Tail				After 2% Cutoff Each Tail					
	Bottom 2%		Top 2%		Bottom 25%		Middle 50%		Top 25%	
	SAR Measure Threshold	Samples Lost	SAR Measure Threshold	Samples Lost	SAR Measure Threshold	Sample Size	SAR Measure Threshold	Sample Size	SAR Measure Threshold	Sample Size
1998 (309, 296)	<-68.96%	6	>130.79%	7	<-36.67%	73	>=-36.67%, <11.65%	148	>=11.65%	75
1999 (303, 290)	<-122.44%	6	>773.2%	7	<-50.13%	72	>=-50.13%, <93.45%	145	>=93.45%	73
2000 (309, 296)	<-76.26%	6	>107.6%	7	<-41.95%	73	>=-41.95%, <25.91%	148	>=25.91%	75
2001 (298, 287)	<-90.24%	5	>131.25%	6	<-29.8%	71	>=-18.37%, <33.92%	144	>=23.03%	72
2002 (317, 304)	<-67.61%	6	>77.17%	7	<-35.69%	75	>=-35.69%, <11.74%	152	>=11.74%	77

Figure B.5: **Class Definition with $SAR^{f_i}_{(t+1,t+2)}$ for Glamor Firms**

Year (Total Sample Size, Size After Cutoff)	Before 2% Cutoff Each Tail				After 2% Cutoff Each Tail					
	Bottom 2%		Top 2%		Bottom 25%		Middle 50%		Top 25%	
	SAR Measure Threshold	Samples Lost	SAR Measure Threshold	Samples Lost	SAR Measure Threshold	Sample Size	SAR Measure Threshold	Sample Size	SAR Measure Threshold	Sample Size
1998 (308, 295)	<-68.87%	6	>95.89%	7	<-41.14%	73	>=-41.14%, <-3.03%	148	>=-3.03%	74
1999 (303, 290)	<-127.65%	6	>431.22%	7	<-72.29%	72	>=-72.29%, <-6.26%	145	>=-6.26%	73
2000 (308, 295)	<-62.97%	6	>120.19%	7	<-8.91%	73	>=-8.91%, <42.69%	148	>=42.69%	74
2001 (298, 287)	<-92.34%	5	>255.56%	6	<-18.37%	71	>=-18.37%, <33.92%	144	>=33.92%	72
2002 (317, 304)	<-65.48%	6	>43.8%	7	<-28.55%	75	>=-28.55%, <9.11%	152	>=9.11%	77

Figure B.6: **Class Definition with $SAR^{f_i}_{(t+1,t+2)}$ for Value Firms**

APPENDIX C
MODELS' CLASSIFICATION CONTINGENCY TABLES WITH DESIGN ONE

Design One	SVM-Score Prediction on $\Delta ROE^{f_i}_{(t,t+1)}$			
25-50-25% Class Definition	Out-Perform	Average	Under-Perform	True Distribution
True Out-Perform	23.59%	64.62%	11.79%	25.10%
True Average	11.01%	78.36%	10.63%	49.98%
True Under-Perform	12.13%	65.83%	22.06%	24.92%
Prediction Distribution	14.45%	71.79%	13.76%	100.00%

Table C.1: **Classification Contingency Table of SVM-Score Predicting $\boldsymbol{\Delta ROE^{f_i}_{(t,t+1)}}$ with Design One**

Design One	SVM-Prob Prediction on $\Delta ROE^{f_i}_{(t,t+1)}$			
25-50-25% Class Definition	Out-Perform	Average	Under-Perform	True Distribution
True Out-Perform	14.66%	79.74%	5.62%	25.10%
True Average	5.41%	88.43%	6.17%	49.98%
True Under-Perform	4.41%	79.50%	16.08%	24.92%
Prediction Distribution	7.53%	83.99%	8.48%	100.00%

Table C.2: **Classification Contingency Table of SVM-Prob Predicting $\boldsymbol{\Delta ROE^{f_i}_{(t,t+1)}}$ with Design One**

Design One	SVM-Multi Prediction on $\Delta ROE^{f_i}_{(t,t+1)}$			
25-50-25% Class Definition	Out-Perform	Average	Under-Perform	True Distribution
True Out-Perform	17.84%	77.03%	5.14%	25.10%
True Average	8.68%	85.57%	5.75%	49.98%
True Under-Perform	11.56%	77.93%	10.51%	24.92%
Prediction Distribution	11.69%	81.55%	6.76%	100.00%

Table C.3: **Classification Contingency Table of SVM-Multi Predicting $\boldsymbol{\Delta ROE^{f_i}_{(t,t+1)}}$ with Design One**

Design One	SVM-Score Prediction on $SAR^{f_i}_{(t+1,t+2)}$			
25-50-25% Class Definition	Out-Perform	Average	Under-Perform	True Distribution
True Out-Perform	19.99%	68.64%	11.36%	25.07%
True Average	8.10%	78.02%	13.88%	50.00%
True Under-Perform	5.77%	59.93%	34.30%	24.93%
Prediction Distribution	10.51%	71.16%	18.34%	100.00%

Table C.4: **Classification Contingency Table of SVM-Score for Predicting $\boldsymbol{SAR^{f_i}_{(t+1,t+2)}}$ with Design One**

Design One	SVM-Prob Prediction on $SAR^{f_i}_{(t+1,t+2)}$			
25-50-25% Class Definition	Out-Perform	Average	Under-Perform	True Distribution
True Out-Perform	18.64%	75.44%	5.93%	25.07%
True Average	5.92%	87.11%	6.97%	50.00%
True Under-Perform	3.92%	70.64%	25.44%	24.93%
Prediction Distribution	8.60%	80.11%	11.30%	100.00%

Table C.5: **Classification Contingency Table of SVM-Prob for Predicting $\boldsymbol{SAR^{f_i}_{(t+1,t+2)}}$ with Design One**

Design One	SVM-Multi Prediction on $SAR^{f_i}_{(t+1,t+2)}$			
25-50-25% Class Definition	Out-Perform	Average	Under-Perform	True Distribution
True Out-Perform	12.91%	80.18%	6.92%	25.07%
True Average	4.82%	87.98%	7.20%	50.00%
True Under-Perform	4.71%	72.42%	22.86%	24.93%
Prediction Distribution	6.82%	82.15%	11.04%	100.00%

Table C.6: **Classification Contingency Table of SVM-Multi for Predicting $\boldsymbol{SAR^{f_i}_{(t+1,t+2)}}$ with Design One**

APPENDIX D
MODELS' CLASSIFICATION CONTINGENCY TABLES WITH DESIGN TWO

Design Two	SVM-Score Prediction on $\Delta EPS^{f_i}_{(t,t+1)}$			
25-50-25% Class Definition	Out-Perform	Average	Under-Perform	True Distribution
True Out-Perform	23.04%	50.58%	26.40%	25.12%
True Average	14.18%	74.92%	10.90%	50.03%
True Under-Perform	22.69%	55.39%	21.91%	24.81%
Prediction Distribution	18.52%	63.95%	17.53%	100.00%

Table D.1: **Classification Contingency Table of SVM-Score Predicting $\boldsymbol{\Delta EPS^{f_i}_{(t,t+1)}}$ with Design Two**

Design Two	SVM-Prob Prediction on $\Delta EPS^{f_i}_{(t,t+1)}$			
25-50-25% Class Definition	Out-Perform	Average	Under-Perform	True Distribution
True Out-Perform	0%	68.18%	31.82%	25.15%
True Average	0%	87.82%	12.18%	50.03%
True Under-Perform	0%	71.05%	28.95%	24.81%
Prediction Distribution	0%	78.72%	21.28%	100.00%

Table D.2: **Classification Contingency Table of SVM-Prob for Predicting $\boldsymbol{\Delta EPS^{f_i}_{(t,t+1)}}$ with Design Two**

Design Two	SVM-Multi Prediction on $\Delta EPS^{f_i}_{(t,t+1)}$			
25-50-25% Class Definition	Out-Perform	Average	Under-Perform	True Distribution
True Out-Perform	12.63%	73.71%	13.64%	25.15%
True Average	5.99%	88.03%	5.98%	50.03%
True Under-Perform	14.08%	73.80%	12.11%	24.81%
Prediction Distribution	9.67%	80.90%	9.43%	100.00%

Table D.3: **Classification Contingency Table of SVM-Multi for Predicting $\boldsymbol{\Delta EPS^{f_i}_{(t,t+1)}}$ with Design Two**

Design Two	SVM-Score Prediction on $\Delta ROE^{f_i}_{(t,t+1)}$			
25-50-25% Class Definition	Out-Perform	Average	Under-Perform	True Distribution
True Out-Perform	17.73%	53.58%	28.69%	25.11%
True Average	13.55%	76.90%	9.55%	49.96%
True Under-Perform	24.46%	55.03%	20.51%	24.93%
Prediction Distribution	17.32%	65.59%	17.09%	100.00%

Table D.4: **Classification Contingency Table of SVM-Score Predicting $\boldsymbol{\Delta ROE^{f_i}_{(t,t+1)}}$ with Design Two**

Design Two	SVM-Prob Prediction on $\Delta ROE^{f_i}_{(t,t+1)}$			
25-50-25% Class Definition	Out-Perform	Average	Under-Perform	Total
True Out-Perform	0%	69.12%	30.88%	25.11%
True Average	0%	88.70% 11.30%	49.96%	
True Under-Perform	0%	76.32%	23.68%	24.93%
Total	0%	80.70%	19.30%	100.00%

Table D.5: **Classification Contingency Table of SVM-Prob Predicting $\boldsymbol{\Delta ROE^{f_i}_{(t,t+1)}}$ with Design Two**

Design Two	SVM-Multi Prediction on $\Delta ROE^{f_i}_{(t,t+1)}$			
25-50-25% Class Definition	Out-Perform	Average	Under-Perform	Total
True Out-Perform	13.61%	66.83%	19.58%	25.11%
True Average	9.65%	83.10%	7.26%	49.96%
True Under-Perform	16.70%	68.49%	14.83%	24.93%
Total	12.40%	75.37%	12.24%	100.00%

Table D.6: **Classification Contingency Table of SVM-Multi Predicting $\boldsymbol{\Delta ROE^{f_i}_{(t,t+1)}}$ with Design Two**

APPENDIX E
DICTIONARY DESCRIPTION AND SAMPLE WORDS

Dictionary Name	Description	Sample Words
Accomplishment	Words expressing task-completion and organized human behavior, including capitalistic terms, modes of expansion, general functionality and programmatic language.	establish, influence, leader, increase, strengthen, proceed, produce, grow, working, succeed, agenda
Aggression	Words embracing human competition and forceful action, physical energy, social domination, goal-directedness, personal triumph, excess human energy, disassembly, and resistance.	blast, collide, conquest, violation, challenging, overcome, veto, prevent, reduce
Ambivalence	Words expressing hesitation or uncertainty, implying a speakers inability or unwillingness to commit to the verbalization being made, including hedges, statements of inexactness, confusion, restrained possibility, and mystery.	perhaps, allegedly, almost, vague, puzzling, could, would, dilemma, suppose, seems
Blame	Terms designating social inappropriateness, downright evil, unfortunate circumstances, unplanned vicissitudes, outright denigrations.	naive, stupid, malicious, bankrupt, detrimental, illegitimate, repugnant
Centrality	Terms denoting institutional regularities and/or substantive agreement on core values, indigenous terms, designations of legitimacy, systematicity, typicality, congruence, predictability, and universality.	basic, innate, decorum, paradigm, standardized, conformity, reliable, expected, landmarks

Table E.1: **31 Dictionaries Description and Sample Words (Adapted from Diction 5.0 Manual)**

Dictionary Name	Description	Sample Words
Cognition	Words referring to cerebral processes, both functional and imaginative, modes of discovery, domains of study, mental challenges, institutional learning practices, and intellection.	learn, consider, compare, biology, economics, examine, teaching, invent, speculate, strategies, analyze, software, estimate
Collectives	Singular nouns connoting plurality that function to decrease specificity, including social groupings, task groups, and geographical entities.	team, humanity, staff, world, congress, republic
Communication	Terms referring to social interaction, both face-to-face and mediated, modes of intercourse, moods of intercourse, social actors, a variety of social purposes.	interview, read, speak, videotape, e-mail, broadcast, declare, demand, reporter, advocates, respond, persuade
Concreteness	Terms possessing no thematic unity other than tangibility and materiality, including sociological units, occupational groups, political alignments, physical structures, forms of diversion, terms of accountancy, modes of transportation, body parts, articles of clothing, household animals and foodstuffs, and general elements of nature.	peasants, manufacturer, congressman, courthouse, store, television, mortgage, finances, airplane, stomach, shirts, cat, wine, silk, sand
Cooperation	Terms designating behavioral interactions among people that often result in a group product, including designations of formal work relations, and informal associations to more intimate interactions, neutral interactions, job-related tasks, personal involvement, and self-denial.	unions, partner, friendship, mediate, network, teamwork, contribute, public-spirited
Denial	Words expressing standard negative contractions, including negative functions words, and terms designating null sets.	shouldn't, don't, not, nor, nothing, nobody

Table E.2: **31 Dictionaries Description and Sample Words (Adapted from Diction 5.0 Manual)**

Dictionary Name	Description	Sample Words
Diversity	Words describing individuals or groups of individuals differing from the norm. Such distinctiveness may be comparatively neutral, but it can also be positive and negative. Functionally, heterogeneity may be an asset or a liability as can its characterizations.	inconsistent, contrasting, exceptional, unique, illegitimate, dispersed, deviancy, rare, distinctive
Exclusion	Words describing the sources and effects of social isolation. Such seclusion can be phrased passively as well as positively and negatively. It can result from voluntary forces and involuntary forces and from both personality factors and political factors.	displaced, self-sufficient, outlaws, secede, loneliness, right-wingers, discard
Familiarity	Common prepositions, demonstrative pronouns and interrogative pronouns, and a variety of particles, conjunctions and connectives.	across, over, this, that, who, what, a, for, so
Hardship	Words about natural disasters, hostile actions, censurable human behavior, unsavory political outcomes, normal human fears, and incapacities.	bankruptcy, pollution, exploitation, unemployment, error, weakness
Human interest	Standard personal pronouns, family members and relations, and generic terms.	he, his, them, cousin, wife, uncle, friend, human, person
Inspiration	Abstract virtues deserving of universal respect, including nouns isolating desirable moral qualities as well as attractive personal qualities, and social and political ideals.	faith, honesty, virtue, courage, dedication, wisdom, success, mercy

Table E.3: **31 Dictionaries Description and Sample Words (Adapted from Diction 5.0 Manual)**

Dictionary Name	Description	Sample Words
Leveling	Words used to ignore individual differences and to build a sense of completeness and assurance, including totalizing terms, adverbs of permanence, and resolute adjectives.	everyone, each, fully, always, completely, inevitably, consistently, absolute
Liberation	Terms describing the maximizing of individual choice and the rejection of social conventions. Liberation is motivated by both personality factors and political forces and may produce dramatic outcomes. Liberatory terms also admit to rival characterizations.	autonomous, options, radical, freedom, deliverance, disentangle, loophole, uninhibited
Motion	Terms connoting human movement, physical processes, journeys, speed, and modes of transit.	job, leap, circulate, momentum, travels, zip, ride, fly
Numerical terms	Any sum, date, or product specifying the facts in a given case, including common numbers in lexical format, terms indicating numerical operations and quantitative topics.	one, hundred, percentage, subtract, multiply, mathematics
Passivity	Words ranging from neutrality to inactivity, including terms of compliance, docility, cessation, tokens of inertness, disinterest, and tranquility.	allow, submit, refrain, backward, inhibit, unconcerned, quietly
Past concern	The past-tense forms of the verbs contained in the Present Concern dictionary.	
Praise	Affirmations of some person, group, or abstract entity, including terms isolating important social qualities, physical qualities, intellectual qualities, entrepreneurial qualities, and moral qualities.	dear, beautiful, bright, vigilant, reasonable, successful, renowned, good

Table E.4: **31 Dictionaries Description and Sample Words (Adapted from Diction 5.0 Manual)**

Dictionary Name	Description	Sample Words
Present concern	Present-tense verbs about general physical activity, social operations, and task-performance.	taste, take, govern, meet, make, print
Rapport	Words describing attitudinal similarities among groups of people, including terms of affinity, assent, deference, and identity	congenial, warrants, willing, permission, equivalent, consensus
Satisfaction	Terms associated with positive affective states, with moments of undiminished joy and pleasurable diversion, or with moments of triumph, including words of nurturance.	passionate, happiness, thanks, welcome, excited, pride, encourage, secure
Self-reference	All first-person references.	I, I'd, I'll, I'm, me, mine, my
Spatial terms	Terms referring to geographical entities, physical distances, and modes of measurement, including general geographical terms as well as specific ones, politically defined locations, points on the compass, as well as terms of scale, quality, and change.	abroad, Poland, county, east, coastal, kilometer, vacant, disoriented, migrated, frontier
Temporal terms	Terms that fix a person, idea, or event within a specific time-interval, thereby signaling a concern for concrete and practical matters, including literal time, metaphorical designations, calendrical terms, elliptical terms, and judgmental terms.	century, instant, seniority, nowadays, year-round, postpone, transitional, premature, obsolete
Tenacity	All uses of the verb to be, three definitive verb forms and their variants, as well as all associated contractions.	is, will, shall, has, must do, he'll, they've

Table E.5: **31 Dictionaries Description and Sample Words (Adapted from Diction 5.0 Manual)**

REFERENCES

[1] E. Abrahamson and E. Amir. The information content of the president's letter to shareholders. *Journal of Business Finance and Accounting*, 23(8):1157–82, 1996.

[2] AIMR. *Association for Investment Management and Research (AIMR) Corporate Disclosure Survey: A Report to AIMR.* Fleishman-Hillard Research, 2000.

[3] E. L. Allwein, R. E. Schapire, and Y. Singer. Reducing multiclass to binary: A unifying approach for margin classifiers. *Journal of Machine Learning Research*, 1:113–141, 2000.

[4] C. Apte, F. Damerau, and S. Weiss. Automated learning of decision rules for text categorization. *ACM Transactions on Information Systems*, 12(3):233–251, 1994.

[5] B. Barber, R. Lehavy, M. McNichols, and B. Trueman. Prophets and losses: Reassessing the returns to analysts stock recommendations. *Financial Analysts Journal*, March/April, 2003.

[6] R. Bekkerman, R. El-Yaniv, N. Tishby, and Y. Winter. On feature distributional clustering for text categorization. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and Development in Information Retrieval*, pages 146–153, 2001.

[7] V. L. Bernard. The Feltham-Ohlson framework: Implications for empiricists. *Contemporary Accounting Research*, 11(2):733–747, 1995.

[8] A. Bernstein, S. Clearwater, S. Hill, C. Perlich, and F. Provost. Discovering Knowledge from Relational Data Extracted from Business News. In *Proceedings of Workshop on Multi-Relational Data Mining (MRDM 2002)*, 2002.

[9] L. Brown, G. Richardson, and S. J. Schwager. An information interpretation of financial analyst superiority in forecasting earnings. *Journal of Accounting Research*, 25:49–67, 1997.

[10] S. H. Bryan. Incremental information content of required disclosures contained in management discussion and analysis. *The Accounting Review*, 72(2):285–301, 1997.

[11] W. B. Cavnar and J. M. Trenkle. N-gram-based text categorization. In *Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval*, 1994.

[12] N. Chawla, K. Bowyer, L. Hall, and W. Kegelmeyer. SMOTE: Synthetic minority over-samplling technique. *Journal of Artificial Intelligence Research*, 16:321–357, 2002.

[13] W. W. Cohen and H. Hirsch. Joins that generalize: text classification using WHIRL. In *Proceedings of KDD-98, 4th International Conference on Knowledge Discovery and Data Mining*, pages 169–173, 1998.

[14] W. W. Cohen and Y. Singer. Context-sensitive learning methods for text categorization. *ACM Transactions on Information Systems*, 17(2):141–173, 1999.

[15] K. Crammer and Y. Singer. On the algorithm implementation of multi-class svms. *JMLR*, 2001.

[16] M. Craven. Learning to extract relations from medline. In *AAAI-99 Workshop on Machine Learning for Information Extraction*, 1999.

[17] M. Craven, D. Freitag, A. McCallum, T. Mitchell, K. Nigam, and C. Y. Quek. Learning to extract symbolic knowledge from the world wide web. In *Proceedings of the 14th International Conference on Machine Learning*, 1997.

[18] K. Dave, S. Lawrence, and D. Pennock. Mining the peanut gallery: opinion extraction and semantic classification of product review. In *Proceedings of WWW2003*, 2003.

[19] A. Davis, J. Piger, and L. Sedor. Beyond the numbers: an analysis of optimistic and pessimistic language in earnings press releases. Working paper, Washington University at St. Louis, Federal Reserve Bank of St. Louis and University of Notre Dame.

[20] S. Deerwester, S. Dumais, G. Furnas, T. Landauer, and R. Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407, 1990.

[21] T. Dietterich and G. Bakiri. Solving multiclass learning problems via error-correcting output codes. *Journal of Artificial Intelligence Research*, 2:263–286, 1995.

[22] S. Dumais. Latent semantic indexing (LSI): Trec-3 report. In *TREC Proceedings*, 1994.

[23] S. Dumais. Improving the retrieval information from external sources. *Behavior Research Methods, Instruments and Computers*, 23(2):229–236, 1998.

[24] S. Dumais, J. Platt, D. Heckerman, and M. Sahami. Inductive learning algorithms and representations for text categorization. In *Proceedings of CIKM-98, 7th ACM International Conference on Information and Knowledge Management*, pages 148–155, 1998.

[25] E. Fama and K. French. The cross section of expected stock returns. *Journal of Finance*, 47:427–465, 1992.

[26] G. A. Feltham and J. A. Ohlson. Valuation and clean surplus accounting for operating and financial activities. *Contemporary Accounting Research*, 11(2):689–731, 1995.

[27] T. D. Fields, T. Z. Lys, and L. Vincent. Empirical research on accounting choice. *Journal of Accounting and Economics*, 31:255–307, 2001.

[28] P. Foltz. Using latent semantic indexing for information filtering. In *Proceedings of the ACM Conference on Office Information Systems*, pages 40–47, 1990.

[29] J. Furnkranz, T. Mitchell, and E. Riloff. A case study in using linguistic phrases for text categorization on the WWW. In *Proceedings of the 1998 AAAI/ICML Workshop*, pages 5–12, 1998.

[30] L. Galavotti, F. Sebastiani, and M. Simi. Feature selection and negative evidence in automated text categorization. In *Proceedings of ECDL-00, 4th European Conference on Research and Advanced Technology for Digital Libraries*, 2000.

[31] K. Gee. Using latent semantic indexing to filter spam. In *Proceedings of the 2003 ACM Symposium on Applied Computing*, pages 460–464, 2003.

[32] F. Ginter, J. Boberg, J. Jarvinen, and T. Salakoski. New techniques for disambiguation in natural language and their application to biological text. *Journal of Machine Learning Research*, 5:605–621, 2004.

[33] N. Glance, M. Hurst, K. Nigam, M. Siegler, R. Stockton, and T. Tomokiyo. Deriving marketing intelligence from online discussion. In *Proceedings of KDD'05*, 2005.

[34] M. Gordon, R. K. Lindsay, and W. Fan. Literature-based discovery on the world wide web. *ACM Transaction on Internet Technologies*, 2(4):261–275, 2002.

[35] R. P. Hart. Redeveloping diction: Theoretical considerations. *Progress in Communication Sciences: Theory, Method, And Practice in Computer Content Analysis*, 16:43–60, 2001.

[36] T. Hastie and R. Tibshirani. Classification by pairwise coupling. *The Annals of Statistics*, 26(2):451–471, 1998.

[37] P. J. Hayes and S. P. Weinstein. Construe/tis: a system for content-based indexing of a database of news stories. In *Second Annual Conference on Innovative Applications of Artificial Intelligence*, 1990.

[38] E. Henry. Market reaction to verbal components of earnings press releases: event study using a predictive algorithm. *Journal of Emerging Technologies in Accounting*, 3:1–19, 2006.

[39] I. Herreman and J. Ryans Jr. The case for better measurement and reporting of marketing performance. *Business Horizons*, 38(5):51–60, 1995.

[40] W. Hersh, C. Buckley, T. J. Leone, and D. Hickam. Ohsumed: an interactive retrieval evaluation and new large test collection for research. In *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 192–201, 1994.

[41] D. J. Ittner, D. D. Lewis, and D. D. Ahn. Text categorization of low quality images. *Symposium on Document Analysis and Information Retrieval*, pages 301–315, 1995.

[42] N. Japkowicz and S. Stephen. The class imbalance problem: A systematic study. *Intelligent Data Analysis*, 6, 2002.

[43] N. Jegadeesh, J. Kim, S. D. Krische, and C. M. C. Lee. Analyzing the analysts: When do recommendations add value? *Journal of Finance*, 59(3):1083–1124, 2004.

[44] N. Jegadeesh and S. Titman. Returns to buying winners and selling losers: Implications for stock market efficiency. *Journal of Finance*, 48:65–91, 1993.

[45] T. Joachims. A probabilistic analysis of the Rocchio algorithm with TFIDF for text categorization. In *Proceedings of ICML-97, 14th International Conference on Machine Learning*, pages 143–151, 1997.

[46] T. Joachims. A probabilistic analysis of the Rocchio algorithm with TFIDF for text categorization. In *Proceedings of the Fourteenth International Conference on Machine Learning*, pages 143–151, 1997.

[47] T. Joachims. Text categorization with support vector machines: learning with many relevant features. In *Proceedings of the European Conference on Machine Learning*, pages 137–142, 1998.

[48] G. H. John, R. Kohavi, and K. Pfleger. Irrelevant features and the subset selection problem. In *Proceedings of 11th International Conference on Machine Learning*, pages 121–129, 1994.

[49] D. V. Khmelev and W. J. Teahan. A repetition based measure for verification of text collections and for text categorization. In *SIGIR'03*, 2003.

[50] S. Kiritchenko. *Hierarchical Text Categorization and Its Application to Bioinformatics*. PhD thesis, University of Ottawa, 2005.

[51] A. Kloptchenko, T. Eklund, B. Back, J. Karlsson, H. Vanharanta, and A. Visa. Combining data and text mining techniques for analyzing financial reports. In *Proceedings of Eighth Americas Conference on Information Systems*, 2002.

[52] A. Kloptchenko, C. Magnusson, B. Back, A. Visa, and H. VAnharanta. Mining textual contents of quarterly reports. Turku Center for Computer Science Technical Reports, 2002.

[53] R. Kohavi and G. H. John. Wrappers for feature subset selection. *Artificial Intelligence*, 97(1-2):273–324, 1997.

[54] G. Kohut and A. Segars. The president's letter to stockholders: an examination of corporate communication strategy. *Journal of Business Communication*, 29(1):7–21, 1992.

[55] R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins. Trawling the web for emerging cyber-communities. In *Proceedings of the Eighth International Conference on World Wide Web (WWW-8)*, pages 1481–1493, 1999.

[56] L. S. Larkey. Automatic essay grading using text categorization techniques. In *Proceedings of ICML-95, 12th International Conference on Machine Learning*, pages 90–95, 1998.

[57] D. D. Lewis. An evaluation of phrasal and clustered representations on a text categorization task. In *Proceedings of SIGIR-92, 15th ACM International Conference on Research and Development in Information Retrieval*, pages 37–50, 1992.

[58] D. D. Lewis. Applying support vector machines to the trec-2001 batch filtering and routing tasks. In *Proceedings of Text Retrieval Conference*, 2001.

[59] D. D. Lewis and W. Gale. A sequential algorithm for training text classifiers. In *Proceedings of SIGIR-94, 17th ACM International Conference on Research and Development in Information Retrieval*, pages 3–12, 1994.

[60] D. D. Lewis and M. Ringuette. A comparison of two learning algorithms for text categorization. In *Proceedings of the Third Annual Symposium on Document Analysis and Information Retrieval (SDAIR'94)*, 1994.

[61] F. Li. Annual report readability, current earnings and earnings persistence. Working paper, University of Michigan, Ann Arbor, 2006.

[62] F. Li. Do stock market investors understand the risk sentiments of corporate annual reports? Working paper, University of Michigan, Ann Arbor, 2006.

[63] Y. Li and A. Jain. Classification of text documents. *The Computer Journal*, 41(8):537–546, 1998.

[64] H. T. Lin, C. J. Lin, and C. H. Weng. A note on platt's probabilistic outputs for support vector machines. Technical report. Department of Computer Science and Information Engineering. National Taiwan University, 2003.

[65] R. J. Lundholm. A tutorial on the Ohlson and Feltham/Ohlson models: answers to some frequently asked questions. *Contemporary Accounting Research*, 11(2):749–761, 1995.

[66] D. Mladenic. Feature subset selection in text learning. In *Proceedings of ECML-98, 10th European Conference on Machine Learning*, pages 95–100, 1998.

[67] S. Morinaga, K. Yamanishi, K. Tateishi, and T. Fukushima. Mining product reputations on the web. In *Proceedings of SIGKDD'02*, 2002.

[68] A. Moschitti and R. Basili. Complex linguistic features for text classification: A comprehensive study. In *Proceedings of the 26th European Conference on Information Retrieval (ECIR)*, pages 181–196, 2004.

[69] I. Muslea. *Active Learning with Multiple Views*. PhD thesis, University of Southern California, 2002.

[70] H. Ng, W. Goh, and K. Low. Feature selection, perceptron learning, and a usability case study for text categorization. In *Proceedings of SIGIR-97, 20th ACM International Conference on Research and Development in Information Retrieval*, pages 67–73, 1997.

[71] K. Nigam, A. K. McCallum, S. Thrun, and T. M. Mitchell. Text classification from labeled and unlabeled documents using em. *Machine Learning*, 39(2/3):103–134, 2000.

[72] J. A. Ohlson. Earnings, book values, and dividends in equity valuation. *Contemporary Accounting Research*, 11(2):661–687, 1995.

[73] F. Peng and D. Schuurmans. Combining naive bayes and n-gram language models for text categorization. In *Proceedings of The 25th European Conference on Information Retrieval Research (ECIR03)*, 2003.

[74] C. Perez-Iratxeta, P. Bork, and M. A. Andrade. Association of genes to genetically inherited diseases using data mining. *Nature Genetics*, 31(3):316–319, 2002.

[75] J. Platt. Probabilistic outputs for support vector machines and comparison to regularized likelihood methods. *Advances in Large Margin Classifiers*, pages 61–74, 1999.

[76] M. Porter. An algorithm for suffix stripping. In S. Jones and Willett, editors, *Readings in Information Retrieval*, pages 313–316. 1997.

[77] F. Provost and T. Fawcett. Robust classification systems for imprecise environments. *Machine Learning*, 42:203–231, 2001.

[78] X. Y. Qiu and P. Srinivasan. Go for gene documents. In *CIKM TMBIO 2006*, 2006.

[79] X. Y. Qiu, P. Srinivasan, and N. Street. Exploring the forecasting potential of company annual reports. In *Proceedings of the American Society of Information Science and Technology 2006 Annual Meeting*, 2006.

[80] Reinganum. Misspecification of the capital asset pricing: Empirical anomalies based on earnings' yield and market values. *Journal of Financial Economics*, 9:19–46, 1981.

[81] S. B. Rice, G. Nenadic, and B. J. Stapley. Mining protein function from text using term-based support vector machines. *BMC Bioinformatics*, 6(Suppl 1)(S22), 2005.

[82] J. J. J. Rocchio. *Document retrieval systems – optimization and evaluation.* PhD thesis, Harvard University, 1966. Report ISR-10, to the National Science Foundation.

[83] J. J. J. Rocchio. Relevance feedback in information retrieval. In *The Smart System – experiments in automatic document processing*, pages 313–323. 1971.

[84] K. Rogers and J. Grant. Content analysis of information cited in reports of sell-side financial analysts. *Journal of Financial Statement Analysis*, 3:17–30, 1997.

[85] M. E. Ruiz and P. Srinivasan. Hierarchical neural networks for text categorization. In *Proceedings of SIGIR-99, 22nd ACM International Conference on Research and Development in Information Retrieval*, pages 281–282, 1999.

[86] M. E. Ruiz and P. Srinivasan. Hierarchical text categorization using neural networks. *Information Retrieval*, 5(1):87–118, 2002.

[87] E. Saad, D. Prokhorov, and D. Wunsch. Comparative study of stock trend prediction using time delay, recurrent and probabilistic neural networks. *IEEE Transactions on Neural Networks*, 9(6):1456–1470, 1998.

[88] G. Salton. *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer*. Addison-Wesley, 1989.

[89] G. Salton and C. Buckley. Improving retrieval performance by relevance feedback. *Journal of the American Society for Information Science*, 41(4):288–297, 1990.

[90] G. Salton and M. McGill. *Introduction to Modern Information Retrieval*. McGraw Hill Publishing Company, 1983.

[91] R. Schapire and Y. Singer. Boosting and rocchio applied to text filtering. In *Proceedings of SIGIR-98, 21st ACM International Conference on Research and Development in Information Retrieval*, pages 215–223, 1998.

[92] R. E. Schapire and Y. Singer. Boostexter: A boosting-based systems for text categorization. *Machine Learning*, 39(2/3):135–168, 2000.

[93] K. Schipper. Analysts' forecasts. *Accounting Horizons*, 5(4):105–21, 1991.

[94] H. Schutze, D. Hull, and J. Pedersen. A comparison of classifiers and document representations for the routing problem. In *Proceedings of SIGIR-95, 18th ACM International Conference on Research and Development in Information Retrieval*, pages 229–237, 1995.

[95] M. Smith and R. J. Taffler. The chairman's statement: A content analysis of discretionary narrative disclosures. *Accounting Auditing & Accountability Journal*, 13(5):624–646, 2000.

[96] P. Srinivasan. Text Mining: Generating Hypotheses from MEDLINE. *Journal of the American Society for Information Science and Technology (JASIST)*, 55(5):396–413, 2004.

[97] R. Subramanian, R. G. Insley, and R. D. Blackwell. Performance and readability: A comparison of annual reports of profitable and unprofitable corporations. *Journal of Business Communication*, 30:50–61, 1993.

[98] H. Taira and M. Haruno. Feature selection in svm text categorization. In *Proceedings of AAAI-99, 16th Conference of the American Association for Artificial Intelligence*, pages 480–486, 1999.

[99] C.-M. Tan, Y.-F. Wang, and C.-D. Lee. The use of bigrams to enhance text categorization. *Information Processing and Management: an International Journal*, 38(4):529–546, 2002.

[100] O. Turetken. Predicting financial performance of publicly traded Turkish firms: A comparative study. Unpublished, Fox School of Business and Management, Temple University, Philadelphia, PA, 2004.

[101] K. Tzeras and S. Hartman. Automatic indexing based on bayesian inference networks. In *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'93)*, pages 22–34, 1993.

[102] V. N. Vapnik. *The Nature of Statistical Learning*. Springer, 1995.

[103] A. Visa, J. Toivonen, P. Ruokonen, H. Vanharanta, and B. Back. Knowledge discovery from text documents based on paragraph maps. In *Proceedings of the 33rd Hawaii International Conference on System Sciences*, 2000.

[104] A. S. Weigend, E. D. Wiener, and J. O. Pedersen. Exploiting hierarchy in text categorization. *Information Retrieval*, 1(3):193–216, 1999.

[105] J. Weston and C. Watkins. Support vector machines for multi-class pattern recognition. In *Proceedings of the Seventh European Symposium on Artificial Neural Networks*, 1999.

[106] E. Wiener, J. O. Pedersen, and A. S. Weigend. A neural network approach to topic spotting. In *Proceedings of SDAIR-95, 4th Annual Symposium on Document Analysis and Information Retrieval*, pages 317–332, 1995.

[107] D. M. Wilkinson and B. A. Huberman. A method for finding communities of related genes. In *Proceedings of the National Academy of Sciences (PNAS)*, volume 101, pages 5241–5248, 2004.

[108] Y. Yang and C. G. Chute. A linear least squares fit mapping method for information retrieval from natural language texts. In *Proceedings of the 14th International Conference on Computational Linguistics*, pages 447–453, 1992.

[109] Y. Yang and C. G. Chute. An example-based mapping method for text categorization and retrieval. *ACM Transaction on Information Systems (TOIS)*, pages 253–277, 1994.

[110] Y. Yang and X. Liu. A re-examination of text categorization methods. In *Proceedings of SIGIR-99, 22nd ACM International Conference on Research and Development in Information Retrieval*, pages 42–49, 1999.

[111] Y. Yang and J. O. Pedesen. A comparative study in feature selection in text categorization. In *Proceedings of the Fourteenth International Conference on Machine Learning*, pages 412–420, 1997.

[112] W. Zhang, Q. Cao, and M. Schniederjans. Neural network earnings per share forecasting models: A comparative analysis of alternative methods. *Decision Sciences*, 35(2):205–237, 2004.