Exploring the Forecasting Potential of Company Annual Reports

Xin Ying Qiu
Management Sciences Department, Tippie College of Business, The University of Iowa. E-mail: xinying.qiu@gmail.com

Padmini Srinivasan
Management Sciences Department, Tippie College of Business, School of Library and Information Science, The University of Iowa

Nick Street
Management Sciences Department, Tippie College of Business, The University of Iowa.

Abstract

Previous research indicates that the narration disclosure in company annual reports can be used to assist in assessing the company's short-term financial prospects. However, not much effort has been made to systematically and automatically assess the predictive potential of such reports using text classification, information retrieval, and machine learning techniques. In this study, we built SVM-based predictive models with different feature selection methods from ten years of annual reports of 30 companies. We used feature selection methods to reduce the term space and studied the class-related vocabulary. Evaluation of predictive accuracy is performed with cross validation and t-test significance tests. We compare different models' performance and analyze misclassification rates by year and by industry. We identify the strengths and weaknesses of each model. Our results support the feasibility of automatically predicting next-year company financial performance from the current year's report. We suggest text features can be further studied to understand their roles as indicators of company's future performance. This research paves the way for large-scale automatic analysis of the relationship between annual reports and short-term performance, as well as the identification of interesting signals within annual reports.

Introduction

Company annual reports (10K filings) are freely available to the public and contain required disclosures, quantitative summaries of the company's financial performance as well as textual discussions. These reports are of great importance in helping investors, corporate managers, and financial analysts with their decision-making. Studies have shown that the narration sections of 10K filings provide information that is as useful as the financial ratios to financial analysts while predicting the company's future prospects (Roger & Grant,1997, Schipper, 1991). The SEC (Securities and Exchange Commission) also requires the reporting of the firm's strategies and managerial priorities, and its view of the past year's performance and future prospects. The major mandatory disclosures in annual reports include reasons for price and sales changes, reasons for revenue and cost changes, planned expenditures, known trends, and future liquidity positions.

Annual reports have been studied as a marketing and communication tool that the corporation uses to convey an image or messages to its stakeholders (Herreman & Ryans, 1995). 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 (Smith & Taffler, 2000), management discussion and analysis (MD&A) (Bryan, 1997), president's letter (Abrahamson & Amir, 1996) as well as the general writing style and readability (Subramanian et al., 1993). 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. The main contributions of these studies 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. (1993) 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 & Taffler (2000) identified thematic keywords from chairman's statements and generated discriminant functions to predict company failure. Bryan (1997) 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 & Segars (1997) studied president's letters in annual reports and suggested that poor performing firms tended 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 statement can assist the prediction of company performance and even failure and bankruptcy. Predictive models for financial performance have been studied mainly using the financial data and various machine-learning techniques such as classification (Turetken, 2004). The goal is generally to identify prominent financial ratios with good predictive power, or better performing classification algorithms such as neural networks. Surprisingly, little research has been done on utilizing the textual content of annual reports to build predictive models, despite findings that the reports have the potential to serve as indicators of company future prospects. The most related work in this direction is that of Kloptchenko et al. (2002) and Visa et al. (2000). In Kloptchenko et al. (2002), company quarterly reports and corresponding financial ratios were clustered with prototype matching clustering and SOM clustering respectively. Although the two clusters did not coincide, the authors found that changes in textual reports tended to occur ahead of the changes in financial performance.

The wealth of information in these reports, especially the narration (textual) portions acknowledged as important for human analysts, remains untapped for machine learning applications. Set in this background literature, we see a well-justified opportunity to explore methods for predicting company financial performance from annual reports. In this research, we explore the feasibility of using the textual content of annual reports for a given year to predict the company's financial performance in the next year. We measure financial performance as return on equity (ROE) ratio. We applied the bag-of-words vector representation to represent each annual report, and performed Support Vector Machine (SVM) based classification with cross-validation to evaluate the predictive accuracy. We also experimented with different feature selection methods to reduce the term space and examine the vocabulary subset capable of predicting a certain performance class. The goal of our study is to establish a baseline for building predictive models from the textual content of annual reports and to analyze different models' strengths and weaknesses in this application domain. More specifically, we address the following research problems:

- Determine the feasibility of building predictive models from annual reports measured by classification accuracy
- Evaluate different predictive models' strengths and weaknesses in predicting the specific class of future financial performance
- Examine the potential of detecting interesting textual (vocabulary) features from annual reports that may serve as signals of future financial performance
- Detect patterns that may exist in different industries and different years

The key contribution of our research is the application of text classification methods in the financial domain for knowledge discovery. We attempt to build and study different classification models to better capture the predictive signals in company annual reports, if they exist. Our analysis also explores the trade-off between different models and the challenges faced when building such applications.

The rest of the paper is organized as follows: In section 2, we will present our methodology related to data collection, prediction problem modeling, selection of modeling approach, and evaluation methods. In section 3, we will present our experimental results in terms of predictive accuracy. We will compare different models to address the research goals presented above. In section 4, we will present our general observations and outline our plans for further study.

Methodology

Data Collection and Class Definition

In this study, we first had several domain experts (two professors and a Ph.D. candidate in Accounting) help us identify our data. With their help we selected a total of 30 companies from 3 industries (pharmacology, IT, and banking). Each company has at least 10 years of consecutive filings with the SEC and has had performance that has fluctuated over the time frame. We retrieved automatically from EdgarScan[1] all the 10K filings of these companies for years 1990 to 2003. Our domain experts also helped us collect the financial measurements for each firm/year from the COMPUSTAT database. We calculated the Return On Equity (ROE) ratio for each firm/year. ROE ratio measures the earnings power of owners' equity. It shows how much income was earned for every dollar invested by owners. The increase and decrease in ROE ratio indicates the change of the firm's earnings power. With the guidance from our domain expert's analysis, the ROE ratios were partitioned into 3 classes. We use t to refer to the year corresponding to the annual report year and t + 1 to refer to the following year.

- If the ROE ratio in year t + 1 is within 5% of the ROE ratio in year t, then the company's year t + 1 performance is classified as belonging to a “neutral" class.
- If the ROE ratio in year t+1 is greater than the ROE ratio in year t by more than 5%, then the company's year t + 1 performance is classified as belonging to a “positive" class.
- If the ROE ratio in year t + 1 is less than the ROE ratio in year t by more than 5%, then the company's year t + 1 performance is classified as belonging to a “negative" class.

Our hypothesis is that the reports carry enough signals to predict the next year's performance. Thus we pair the document with the class of the next year's performance. When we retrieved all the 10K filings for these 30 companies, we obtained 316 documents. Among these, 37 documents could not be labeled because of lack of ROE data, resulting in 279 labeled documents. These form our pool of instances to build our predictive models for a 3-class text classification problem.

Experiment Design

Before applying classification to the documents, we first preprocessed the documents by removing HTML tags, tables, and numbers. We used the SMART system (Salton, 1989) to remove stop words, perform stemming, and construct vector-space representations of the documents. SMART is a document indexing and information retrieval system available free for research purposes. Each report is represented as a vector of its distinctive terms and their “term frequency inverse document frequency" (TF*IDF) weights. TF*IDF is the most successful and widely used weighting scheme to estimate 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. This document vector representation produces a large feature space of around 50,000 terms, while the document vectors are quite sparse. SMART provides multiple weighting options. In our cross-validation experiments (explained later), we experimented with two TF*IDF motivated weighting schemes, “ltc" and “atc", that are explained in Singhal et al. (1996), and they were significantly similar based on paired 2-tailed t-test. Therefore, for the rest of the discussion, we refer only to text representation vectors with “atc" weight.

The main classifier we used throughout this study is the SVM-Light[2] implementation of Support Vector Machines with linear kernel function and default parameter settings. SVMs have been recognized as being able to efficiently handle high-dimensional

[1] http://edgarscan.pwcglobal.com/servlets/edgarscan

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

problems with many thousands of support vectors. Previous research has shown that SVMs can perform text categorization better than some other classifiers such as naive Bayes, Rocchio, and k-NN (Joachims, 1998). In our current study, we applied SVMs based text classification in the financial domain. Rigorous comparisons of alternative machine learning methods and different kernel functions in SVMs will follow this study.

The classification task is defined as assigning each annual report to exactly one of the three classes: predicting better performance in the next year (positive class), predicting same performance (neutral class), and predicting worse performance (negative class). Since standard SVMs are designed for 2-class problems, we implement this multi-class classification with three 2-class (binary) classifiers. In the training step, three individual SVM classifiers were trained to predict each of the three classes (positive, negative, and neutral) against the other two. In the testing step, each of the three classifiers gives a decision score for each testing document. We use the highest score to assign the document to exactly one class. We randomly split the 279 documents into 2/3 for training and 1/3 for testing and performed one classification. We repeated the process 10 times where we referred to each repetition of the process as one fold. The accuracy of each fold is recorded and average accuracies computed.

Feature Selection

Our choice of document representation with bag-of-words vectors uses word stems. This is a common approach in text classification. Recent research by Moschitti & Baili (2004) 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 predictive accuracy of SVMs. Therefore, in our current study, we focus on studying basic word stem features and their impact on the prediction problem. We leave consideration of the more complex or fine-grained linguistic information to future study.

Previous research on SVMs (Joachims,1998) suggests that they eliminate the need for feature selection to achieve high classification accuracy. The argument is that SVMs use functions that could separate the data space with the widest margin, and thus do not depend on the number of features. However, during our interactions with accounting experts, it became clear that users in the financial arena are unlikely to value a system that cannot, in some sense, explain its logic. It is therefore an important goal to be able to explain, to the extent possible, the logic behind our classifiers. Thus our interest is to proactively explore feature selection in our application to understand how a report's textual content indicates changes in a firm's future financial performance. We would like to see if we could construct an appropriate “vocabulary" for each class. Moreover, our current term space, even after our preprocessing step, is still large with 50,000 terms, most of which have extremely low frequency and little meaning. Thus we explore feature selection methods.

Yang & Pedesen (1997) systematically evaluated five feature selection techniques by applying them to text categorization problems on a large-scale corpus. One of their conclusions is that document frequency method and χ^2 method, as defined below, eliminated 90% of the unique terms without loss of categorization accuracy. We tested these two methods with our prediction problem and also suggested a novel statistical method utilizing the z-test.

- **Document frequency thresholding (DF)**
  Document frequency is the number of unique documents in which a term occurs. We computed each term's document frequency in the training data set, and applied a heuristic threshold to eliminate terms that appeared in less than three documents. The assumption is that terms that rarely appear in the corpus

carry little class-specific information and do not affect the global prediction performance (Yang & Pedesen, 1997). In our implementation, the DF threshold removes on average 75% of the total terms.

- **χ^2 Statistic (CHI)**

  For a term feature, the χ^2 statistic tests the null hypothesis that the observed term frequency in a training document is not different from its statistically expected frequency. Otherwise, if the term frequency is significantly different from expectation, it implies this term is important in defining the class of the document. We implemented the χ^2 measurement following Yang & Pedesen (1997) so that each term has three χ^2 scores for the three classes. We picked the maximum score[3] and tested with one degree of freedom at the 5% significance level to decide if we should assign this term to a class, or eliminate it from the vocabulary. Therefore, the constructed class vocabularies contain mutually exclusive sets of terms. In the cross-validation experiments, the χ^2 method reduced the vocabulary by 7% up to 55%. In all folds, the negative class vocabulary is the largest.
- **Z-test statistic (Z-test)**

  The Z-test statistic measures the independence between the mean term frequencies in the two classes. Given a term t and a class label c, we computed average term frequency per document ($\mu_{(t,c)}$) when the term appears in the class documents and when it does not ($\mu_{(t,c_0)}$). Then z-test scores are measured as:

  $$Z(t,\ c) = \frac{\mu_{(t,c)} - \mu_{(t,c_0)}}{\sqrt{\frac{\sigma^2_{(t,c)}}{n_c} + \frac{\sigma^2_{(t,c_0)}}{n_{c_0}}}}$$

  Each term has one z-test score for each of the three classes. The scores at the 5% significance level determine the labeling of the term. Thus each term may be eliminated or assigned to as many as three classes. The class vocabularies constructed in this way have overlapping terms. The method reduces the size of total term space by 20% to 40%.

The above feature selection methods are implemented before applying the SVM classifiers. In the case of the DF threshold method, training documents and testing documents vector representations are reconstructed with the same reduced vocabulary selected from the global[4] term space. With the CHI and Z-test methods, the training documents' vector representations include only the terms from its class vocabulary, while the global vocabulary is used to represent the testing documents.

<u>Evaluation</u>

Since our prediction problem is modeled as a 3-class classification question, we evaluated both the accuracy for predicting all 3 classes at one time, and the accuracy for predicting each class independently. Overall, we have six different predictive models all using the SVM classifier approach:

- No feature selection (SVM)
- DF threshold (DF-SVM)
- χ^2 Statistic (CHI-SVM)
- Z-test Statistic (Z-SVM)

[3] We used the maximum χ^2 following Yang and Pedesen (1997).

[4] Global implies from all 279 instances.

- DF and χ^2 (DF-CHI-SVM)
- DF and z-test (DF-Z-SVM)

Each model's performance is evaluated with the average predictive accuracy of 10 repetitions of random-split of data. In each repetition, the data is randomly split into 2/3 for training and 1/3 for testing. Each model's average accuracy is compared with the majority-vote baseline and pair-wise with each other using t-test significance tests. The random split of data for each of the 10 repetitions is the same across the models to assure comparability. As a final step to study the features, we applied DF-CHI and DF-Z feature selection models to the complete document set to perform a qualitative analysis of the class-specific features.

Results and Analysis

Overall Prediction Accuracy

We can observe from Table 1 the overall classification accuracies of different models, and their differences against the baseline as measured with p-value. We can say that the SVM model without feature selection and DF-SVM perform significantly better than baseline. This suggests that it is possible to build automatic predictive models with accuracy better than majority vote. However, adding feature selection to the SVM model did not result in better accuracy. The only marginally successful feature selection model is DF-SVM, which achieved the same performance as SVM-only model. However, DF-SVM reduces the original term space by 80% as illustrated in Tables 7. Considering both accuracy and feature set size, we believe DF-SVM is better than SVM-only model, mainly for its ability to generate much smaller vocabulary without degrading the predictive accuracy.

Table 1: T-test comparing performance in predicting all 3 class: Numbers in paretheses represent average accuracies. Significant p-values are denoted with underline.

P-value	**SVM (0.593)**	**DF-SVM (0.591)**	**CHI-SVM (0.535)**	**Z-SVM (0.574)**	**DF-CHI-SVM (0.534)**	**DF-Z-SVM (0.565)**
Baseline (0.556)	0.02	0.02	0.07	0.14	0.01	0.47
SVM (0.593)		0.89	0.003	0.02	0.002	0.003

We also look into the confusion matrices of three models: DF-SVM, DF-CHI-SVM, and DF-Z-SVM, to understand the misclassification errors. These will give us insights into how the tool works and how to make further improvements. As illustrated in Tables 2, 3, and 4, DF-SVM and DF-Z-SVM generated class distribution more similar to the true class distribution than DF-CHI-SVM. DF-SVM has a positive-neutral-negative distribution of 21%, 71% and 7%; DF-Z-SVM has 20%, 67% and 14%; DF-CHI-SVM has 5%, 88%, and 7%; while the true class distribution is 26%, 56%, and 18%. We conclude that even though DF-Z-SVM did not perform as well as DF-SVM in terms of overall accuracy, its predicted class distribution is more similar to the true class distribution than DF-CHI-SVM and that of DF-SVM. From this perspective, DF-Z-SVM is more promising than DF-CHI-SVM.

Now we consider the types of errors made. There are two types of errors that are particularly important: predicting the negative class as positive class, and predicting the positive class as negative. The former represents loss with high cost, while the latter is loss of opportunity. The first error rates are 5.5% for DF-SVM, 4.7% for DF-Z-SVM, and 1.5% for DF-CHI SVM. The second error rates are 3.6% for DF-SVM, 6.6% for DF-Z-SVM and 2.6% for DF-CHI-SVM. For both errors, DF-CHI-SVM has the lowest error rates. We can conclude that while DF-SVM and DF-CHI-SVM approximated the true class distribution better, DF-CHI-SVM avoided high-cost errors by predicting a much larger majority of neutral class.

Table 2: DF-SVM Average Normalized Confusion Matrix

	Predicted Class			
True Class	**+1**	**0**	**-1**	**Total**
+1	9.75%	12.85%	3.59%	26.19%
0	5.4%	47.86%	2.36%	55.62%
-1	5.46%	11.28%	1.45%	18.19%
Total	20.82%	71.99%	7.4%	100%

Table 3: DF-Z Average Normalized Confusion Matrix

	Predicted Class			
True Class	**+1**	**0**	**-1**	**Total**
+1	9.3%	10.29%	6.6%	26.19%
0	5.5%	45.19%	4.93%	55.62%
-1	4.71%	11.47%	2.01%	18.19%
Total	19.51%	66.95%	13.54%	100%

Table 4: DF-CHI Average Normalized Confusion Matrix

	Predicted Class			
True Class	**+1**	**0**	**-1**	**Total**
+1	1.7%	21.94%	2.56%	26.19%
0	1.79%	50.7%	3.23%	55.62%
-1	1.45%	15.72%	1.02%	18.19%
Total	4.84%	88.36%	6.8%	100%

Class-Specific Accuracy

Next, we would like to compare the models with respect to their ability to predict for specific performance classes. Table 5 shows that in predicting the positive class (i.e., better next-year financial performance), all feature selection models help produce a much smaller vocabulary of specific interest to the positive class documents, at no cost of prediction accuracy. The accuracies among all feature selection models are very close to each other. DF-Z-SVM has the highest accuracy by very a small margin.

Table 6 shows that when predicting the negative class (i.e., worse next-year financial performance), only DF-SVM maintains the same accuracy as the pure SVM model. All other feature selection methods affected SVM negatively.

Table 5: T-test comparing performance in predicting positive class: Numbers in paretheses represent average accuracies. Significant p-values are underlined.

P-value	**DF-SVM (0.7319)**	**CHI-SVM (0.7290)**	**Z-SVM (0.7292)**	**DF-CHI-SVM (0.7218)**	**DF-Z-SVM (0.7373)**
SVM (0.7329)	0.87	0.81	0.57	0.54	0.44
DF-SVM (0.7319)		0.86	0.68	0.60	0.45

Table 6: T-test comparing performance in predicting negative class: Numbers in paretheses represent average accuracies. Significant p-values are underlined.

P-value	DF-SVM (0.8138)	CHI-SVM (0.7480)	Z-SVM (0.7962)	DF-CHI-SVM (0.7399)	DF-Z-SVM (0.7873)
SVM (0.8138)	1	<0.001	0.03	<0.001	0.005
DF-SVM (0.8138)		<0.001	0.03	<0.001	0.005

Textual Feature Analysis

Positive and Negative Class Vocabularies

Table 7 shows a unique observation about the positive class vocabulary. The positive class has vocabulary size ranging from a few hundred to nearly 15,000 generated by different feature selection methods. However, referring to Table 5, all methods performed the same as the pure SVM with around 45,000 words. We may conclude that positive class of companies is easier to identify regardless the size of the feature set.

Table 7 also shows that CHI's positive and neutral class vocabularies are only about 7% the size of the negative class vocabulary. A look at the negative class vocabulary from one fold of CHI shows that 88% of the terms are 3-letter terms most of which have little meaning. In other words, we found many more meaningless words in the negative class vocabulary than the positive or neutral class vocabulary. So far none of the feature selection methods has been successful in identifying a subset of terms special to the negative class without loss of predictive accuracy. This may coincide with earlier research findings (Subramanian et al., 1993) that poor performing firms' reports are hard to read and tend to use significantly more jargon and modifiers.

Table 7: Average vocabulary size by model from cross-validation training model

Vocabulary Size	Positive Class	Neutral Class	Negative Class	Total
CHI	858	996	11598	13454
Z-test	14938	22368	23754	[a]
DF-CHI	603	996	883	2482
DF-Z	7231	8809	7978	[a]
DF	--	--	--	10548
Original Total	--	--	--	44607

Interesting Features

We now take a look at some sample words from the three class vocabularies generated by DF-CHI as shown in Table 8. Since we used linear kernel function to build SVM models, the signs of the feature weights in the model can be used to explain the term's contribution to the classification of a document. Looking into the weights of the features in DF-CHI model yields some interesting observations. For example, “discret", “stockhold", “intellig", “profit", “divers", “extraordin", “innovat" and “succeed" have positive weights in the positive class predictive model but negative weights in the negative class predictive model. This implies that these terms contribute to classifying a positive class document but not to a negative class document. Similarly, “stress", “cumulat", “lessee", “unknow", “doubt" have positive weights in the negative class predictive model

[a] Z-test models have overlapping class vocabularies. In the feature selection step, each class vocabulary is recorded but not the total vocabulary.

and negative weights in the positive class model. Interestingly, “delay", “uncertain", “web" and “internet" have positive weights in the positive class model, but negative weights in the negative class model, while “satisfact", “portfolio" and “award" have negative weights in the positive class model but positive weights in the negative class model.

Table 8: Sample words from DF-CHI vocabulary

DF-CHI Class +1	admissibl, approach, award, career, certif, chairperson, charact, cordial, cultur, dear, disclos, doubt, effic, feasibl, flexibl, harm, hostil, incent, industrial, infeasibl, intangibl, magic, modest, monitor, necessit, neighbor, opposit, penalt, permissibl, perpetual, portfolio, postemploy, predetermin, preestabl, promis, punctual, purpos, reevalu, restrain, satisfact, shortcom, stress, survey, truth, unannounc, uncommit, underpaid, unguaranteed, unknow, unmatur, vary, wealth, wrongdo
DF-CHI Class 0	adapt, attitud, bargain, behavior, catalog, categor, compatibl, competit, consensus, default, deterior, disagree, disapprov, dissatisfact, diversif, dynam, financiac, focus, foreclosur, foreseen, guarantee, imbalanc, inconsist, indetermin, insuffic, intellectual, interrupt, invalid, know, moderat, obsolet, overdu, payoff, preclud, prepay, prerefund, prospectus, protocol, prudent, questionnair, realloc, redesign, redetermin, reexamin, refocus, reinvest, reissu, reliabl, reorgan, reputat, research, retrain, satisf, scientif, securit, signal, specul, sustain, teamwork, techniqu, threat, thrift, trademark, trust, unaffect, undevelop, unfair, unforeseen, unidentif, unnecess, unplan, unsatisf, valid, violat, wrong
DF-CHI Class -1	burgeon, chanc, collaps, complex, conspicu, contend, cope, corrupt, crucial, curtain, delay, detain, devast, downtim, eminent, extraordin, fatal, forecast, forego, indefeasibl, mileston, mission, overdraft, owe, payback, pendent, philharmon, profit, promin, redirect, reinforc, relocat, resuppl, rewrit, setback, shortfal, succeed, superior, troubleshoot, turnov, unauthor, unbudget, uncertain, undergon, unforeseeabl

Analysis by Industry and by Year

We selected DF-Z-SVM model to further analyze performance by industry and by year. The choice is made because of its better tradeoff with different measures: it produces smaller class-specific vocabulary; it generates better predicted class distribution relative to the true class distribution; and it performs well in predicting both positive and neutral class. We take the 10-repetition experiments of DF-Z-SVM and calculate the average accuracy for each industry and for each year separately. The results are given in Table 9 and Figure 1.

Table 9 shows that IT and Banking have similar average accuracy, while Pharmacology is clearly different from the other two. Pharmacology is one major subdivision with unique characteristics in the “manufacturing" industry where the IT subdivision also belongs. Results in Table 9 suggest that relatively speaking, there exist predictable patterns in the Banking and IT industries that could be captured with machine learning models with fair accuracy, but Pharmacology industry's future performance may be more difficult to predict.

Figure 1 shows the predictive accuracy and standard deviation by year. We did not observe a pattern and the large standard deviation also indicates the lack of useful information in this analysis. Each company has 10 consecutive years of data ranging from 1990 to 2003. While each industry has on average about 100 documents for one fold of training and testing the models, each year has only on average about 20 documents for training and testing of one fold. We believe that the poor predictive accuracy by year results from the limited data we had for each year. In our future research, we will use more company data for each year to fully assess if there are predictable patterns by year.

Table 9: DF-Z Average accuracy by industry

Industry	Avg. Accuracy	Standard Deviation
IT	0.58521	0.0811
Pharmacy	0.52501	0.0776
Banking	0.58036	0.0833

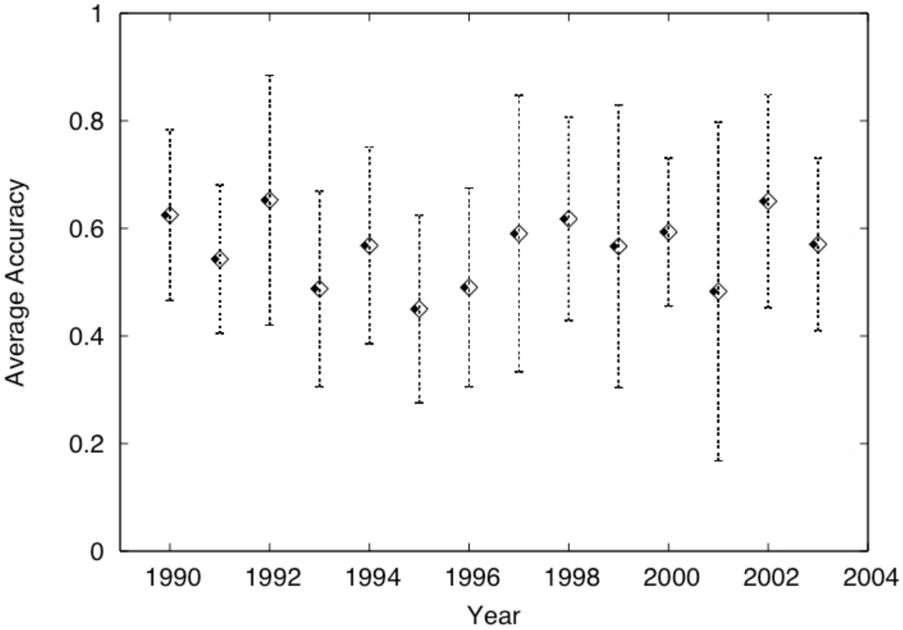

Figure 1: DF-Z Average Accuracy by year

Conclusion & Discussion

The major conclusion from this study is that it confirms the feasibility of using text classification on annual reports to predict future short-term financial performance. We performed cross validation and t-tests to rigorously assess the performance of different models. To explore ways of understanding the forecasting relations, we experimented with two existing feature selection methods and one novel application of the z-test statistical method. We evaluated the tradeoff of each feature selection method and further looked into some of the interesting textual features from the annual reports. We observed that DF thresholding is an effective and simple method to greatly reduce the term space without affecting predictive accuracy. We find our Z-test feature selection method to be promising in future research. We detected the existence of patterns by industry and will further explore the patterns by year in our future work.

The significance of our study lies in two aspects: 1) Annual reports are a vast and abundant data source that remains untapped by text mining and machine learning researchers. This is an important observation given the current interest in mining text collections from different domains. 2) The development and refinement of the techniques to relate annual reports with future financial performance may result in an implementable predictive system. This kind of tool could be of value to analysts as an additional source of indication, to stockholders as a consulting tool, and to companies as a check on their own forecast and disclosure.

We would like to extend our current study in several ways. First, the three industries' annual reports were pooled together to form the training and testing sets. We traced the prediction results back by industry and by year. Alternatively, we can build predictive models by industry and by year separately and evaluate the performances of industry model and the year model. Second, other measurements besides ROE such as earning per share or stock price changes may be used as dependent prediction variables.

Acknowledgments
The authors thank our accounting domain experts Jeffrey Burks, Professor Ramji Balakrishnan, Professor Morton Pincus, and Professor Guojin Gong for helping us selecting companies, collecting financial data, and providing insights and advice.

References

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

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

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

Joachims, T. (1998). Text categorization with support vector machines: learning with many relevant features. *Proceeding of the European Conference on Machine Learning*, 137-142.

Kloptchenko, A. et al. (2002). Combining data and text mining techniques for analyzing financial reports. *Proceedings of Eighth Americas Conference on Information Systems*.

Kloptchenko, A. et al. (2002). Mining textual contents of quarterly reports. *Turku Center for Computer Science Technical Reports.*

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

Moschitti, A. & Basili, R. (2004). Complex linguistic features for text classification: A comprehensive study. *Proceedings of the 26th European Conference on Information Retrieval (ECIR)*, 181-196.

Rogers, R. & Grant, J. (1997). An empirical investigation of the relevance of the financial reporting process to financial analysts. *Unpulished*.

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

Schipper, K. (1991). Analysts' forecasts. *Accounting Horizons*, 5(4):105-21.

Singhal, A., Buckley, C., & Mitra, M. (1996). Pivoted document length normalization. *Proceedings of the 1996 ACM SIGIR Conference on Research and Development in Information Retrieval*, 21-29.

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

Subramanian, R., Insley, R. G., & Blackwell, R. D. (1993). Performance and readability: A comparison of annual reports of pro table and unprofitable corporations. *Journal of Business Communication*, 30:50-61.

Turetken, O. (2004). Predicting financial performance of publicly traded Turkish firms: A comparative study. *Unpublished*.

Visa, A. et al. (2000). Knowledge discovery from text documents based on paragraph maps. *Proceedings of the 33rd Hawaii International Conference on System Sciences*.

Yang, Y. & Pedesen, J. O. (1997). A comparative study in feature selection in text categorization. *Proceedings of the Fourteenth International Conference on Machine Learning*, 412-420.