Exploring the Forecasting Potential of Company Annual Reports

Xin Ying Qiu* xin-qiu@uiowa.edu
Padmini Srinivasan!* padmini-srinivasan@uiowa.edu
W. Nick Street* nick-street@uiowa.edu

Management Sciences Department, Tippie College of Business*
School of Library & Information Science!
University of Iowa
Iowa City, IA 52242

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 annual 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 prediction accuracy is performed with cross validation and t-test significance test. 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 interesting signals within annual reports.

Keywords: text classification, feature Selection, text mining application, predictive model, knowledge discovery

1 Introduction

Company annual reports (10K filings) are freely available to the public and contain required disclosure on the quantitative summary of the company's financial performance as well as textual discussion. These reports are of great importance in helping investors, corporate managers, and financial analysts with judgment and decision making. Studies have shown that the narration sections of 10K filings provide information as useful as the financial ratios to the financial analysts in predicting the company's future prospects [11] [9]. The usefulness of the textual contents in annual reports is also well justified by the SEC requirements on reporting the firm's strategies and managerial priorities, and view of the past year's performance and future prospects. The major mandatory disclosure includes reasons for price and sales changes, reasons for revenue and cost changes, planned expenditures, known trends, and future liquidity position.

Annual reports have been studied as a marketing and communication tool that the corporation uses to convey an image or messages to its stakeholders [3]. More recent studies on the relationship between the reports and the firm's performance have focused on special sections of the reports, such as the chairman's statement [13], management discussion and analysis (MD&A) [2], president's letter [1] and the general writing style and readability [14]. 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. [14] 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. [13] identified thematic keywords from chairman's statements and generated discriminant functions to predict company failure. Bryan [2] 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. [7] 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 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 [15]. 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 utilized 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. [6] and Visa et al. [16]. In Kloptchenko and others study in 2002 [5], the company quarterly reports and corresponding financial ratios were clustered separately 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 machine learning 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 financial performance in the next year. We measure financial performance as return on equity (ROE) ratio. We applied the traditional bag-of-words vector representation to represent each annual report, and performed Support Vector Machine (SVM) based classification with cross-validation to evaluate the prediction accuracy. We also experimented with different feature selection methods to reduce the term space and examine the vocabulary special to 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 analyze different models' strengths and weaknesses. More specifically, we address the following research problems:

- Determine the feasibility of building predictive models from annual reports, measured with 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 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 our experience and analysis from applying machine-learning-based text classification in the financial domain for knowledge discovery. We attempt to build and study different models to better capture the predictive signals from company annual reports, if they exist. Our analysis also presents the tradeoff between different models and the challenges faced in building such applications.

The rest of the paper is organized as follows: In section 2, we will present our methodologies in data collection, prediction problem modeling, choice of methods to build models, and evaluation methods. In section 3, we will present the experiment results in terms of prediction accuracy. We will compare different models to address the research problems presented above. In section 4, we will make a few general observations and project on future study.

2 Methodology

2.1 Data Collection and Class Definition

In this study, we first had a domain expert (Ph.D candidate in Accounting) help us select a total of 30 companies from 3 industries (pharmacology, IT, banking) which have at least 10 years of consecutive filings with the SEC and whose performance might fluctuate over the time frame. We retrieved automatically from EdgarScan[1] all the 10K filings of these companies for years 1990 to 2003. Our domain expert 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. 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.

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

- If the ROE ratio in year $t+1$ is great 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 signal to predict the next year's performance. Thus, we paired each of the 300 annual reports (30 companies $\times$ 10 years) with the class of the next year's performance. These 300 report-class pairs form our pool of instances, and we use these to build our predictive model as a 3-class text classification problem.

2.2 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 [10] to remove stop words, perform stemming, and construct vector-space representation 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 $\times$ inverse document frequency" (TF$\times$IDF) weights. TF$\times$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 documents are those that occur very often in this document but not much in the other documents. This document vector representation produces a large scale feature space of around 50,000 terms, and the document vectors are sparse. SMART provides multiple weighting options. In the later classification evaluation with cross-validation, we experimented with two TF$\times$IDF motivated weighting schemes, "ltc" and "atc" that are explained in [12], and they were significantly similar based on paired 2-tailed t-test. Therefore, for the rest of the discussion, we refer only to text vector with "atc" weight.

The main classifier we used throughout this study is SVM-Light[2] implementation of Support Vector Machines with default parameter settings and linear kernel function. SVMs have been recognized as being able to efficiently handle high-dimensional problems with many thousands of support vectors. Previous research has shown that SVMs can perform text categorization better than conventional classifiers such as naive Bayes, Rocchio, and k-NN [4]. 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 performed 10-fold cross-validation where in each fold all 300 documents are randomly split into 2/3 for training and 1/3 for testing. The accuracy of each fold is recorded.

2.2.1 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 [8] 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. Therefore, in our current study, we focus on studying the basic word stem features and their relationship to the prediction, without considering the more complex or fine-grained linguistic information.

Previous research on SVMs [4] suggest 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. 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 can 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.

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

Yang et al. [17] 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 which appeared in less than three documents. The assumption is that terms that rarely appear in the corpus carry little category-specific information and do not affect the global prediction performance [17]. 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 the formula given in [17] 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 10-fold 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 labelling of the term. Thus each term may be eliminated or assigned to as many as three classes. The class vocabluaries 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.

2.3 Evaluation

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

Each model's performance is evaluated with 10-fold cross-validation and compared with the majority-vote baseline and pairwise with each other using t-test significance tests. The random split of data for each of the 10 folds 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.

3 Results and Analysis

3.1 Overall Prediction Accuracy

We can observe from Table 1 the overall classification

[3] We used the maximum χ^2 following Yang and others' methodology in [17].

[4] Global implies from all 300 instances.

Table 1: T-test comparing performance in predicting all 3 classes: Numbers in parentheses represent average accuracies. Significant p-values are denoted in bold

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**

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

	SVM Result			
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%

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

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 neutral class: Numbers in parentheses represent average accuracies. Significant p-values are denoted in bold

P-value	DF-SVM (0.7119)	CHI-SVM (0.5690)	Z-SVM (0.6978)	DF-CHI-SVM (0.5566)	DF-Z-SVM (0.7110)
SVM (0.7053)	0.21	**<0.001**	0.29	**<0.001**	0.60
DF-SVM (0.7119)		**<0.001**	0.12	**<0.001**	0.90

Table 7: T-test comparing performance in predicting negative class: Numbers in parentheses represent average accuracies. Significant p-values are denoted in bold

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**

Table 8: 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	*
DF	-	-	-	10548
Original Total	-	-	-	44607

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

Table 9: Vocabulary size by models from all documents

Vocabulary size	Positive Class	Neutral Class	Negative Class	Total
DF-Z	9978	11562	11282	13850
DF-CHI	664	1541	1103	3308
DF	-	-	-	13850
Original Total	-	-	-	174128

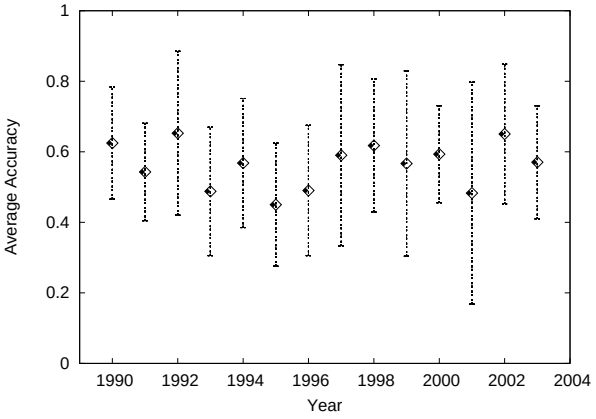
Table 10: Sample words from DF-CHI vocabularies

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

Table 11: DF-Z Average Accuracy by Industry

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

Figure 1: DF-Z Average Accuracy by Year



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 method 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. DF-SVM also reduces 80% of the orginial term space as illustrated in Tables 8 and 9. 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 prediction accuracy.

We also look into the confusion matrices of three models: DF-SVM, DF-CHI-SVM, and DF-Z-SVM in order to understand the misclassification errors. 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 which has significantly better accuracy than baseline. From this perspective, DF-Z-SVM is more promising than DF-CHI-SVM.

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. We can observe that while DF-SVM and DF-CHI-SVM approximated the true class distrubtion better, DF-CHI-SVM avoided high-cost errors by predicting a much larger majority of neutral class.

3.2 Class-Specific Accuracy

Next, we would like to see the differences of each model in predicting different 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 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 in predicting the neutral class (i.e., the same next-year financial performance), all feature selection methods except for Chi-square achieve accuracies similar to the SVM-only approach. The Chi-square models perform significantly worse than SVM-only.

Table 7 shows that in 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.

To summarize, we find DF-SVM and DF-Z-SVM are better in predicting positive and neutral with the benefit of smaller vocabularies.

3.3 Textual Feature Analysis

3.3.1 Vocabulary Size and Prediction Accuracy

To examine the potential of detecting interesting textual features from annual reports, we would like to first look at the effects of reducing vocabulary size on the prediction accuracy. Table 8 shows the average vocabulary size for each class from cross validation by each feature selection method. Both Z-test and DF-Z produce vocabularies with overlapping terms for the three classes. Their vocabularies are larger than those of CHI and DF-CHI respectively. Z-test and DF-Z degraded SVM's performance in predicting negative class but not the positive or neutral class. CHI and DF-CHI produced vocabularies with mutually exclusive terms for the three classes. DF-CHI generates the smallest vocabulary size for all three classes. When applied to SVM classification, CHI and DF-CHI did not change SVM's performance in predicting positive class, but affected SVM negatively on all other classifications.

3.3.2 Positive and Negative Class Vocabularies

Table 8 shows a unique observation about the positive class. The positive class has vocabulary size ranging from a few hundred to nearly 15,000 generated by different feature selection methods. However, all methods performed the same as the pure SVM with around 45,000 words. We may conclude that positive class of companies are easier to identify regardless the size of the feature set.

Table 8 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. We applied DF, DF-Z and DF-CHI feature selections to the complete data set. Table 9 shows the vocabulary size and distribution. In DF-CHI, 27% of the negative class words are of three or fewer letters in length, compared with 10% and 9% in the positive class and neutral class respectively. In DF-Z, 46% of the negative class words are of three or fewer letters in length, compared with 4% and 17% in the positive class and neutral class respectively. Most of the 3-letter words 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 no methods have been successful in identifying a subset of terms special to the negative class without loss of prediction accuracy. This may coincide with earlier research findings [14] that poor performing firms' reports are hard to read and tend to use significantly more jargon and modifiers.

3.3.3 Interesting Features

We now take a look at some sample words from the three class vocabularies generated by DF-CHI as shown in Table 10. The vocabularies are exclusive from each other. Given the non-directional nature of the χ^2 statistic, the inclusion of a word in a class-vocabulary implies that either the presence or the absence of the word in the documents could be indicative of the class.

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. We look into the weights of the features in DF-CHI model and made 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 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.

3.4 Analysis by Industry and by Year

We selected DF-Z-SVM model to further analyze patterns by industry and by year. The choice is made because of its better tradeoff with different measures: it produces smaller and 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-fold experiments of DF-Z-SVM and calculate the average accuracy for each industry and for each year separately. The results are given in Table 11 and Figure 1.

Table 11 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 11 suggest that there exist predictable patterns in the Banking and IT industries which 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 prediction accuracy and standard deviation by year. We did not observe a pattern and the large standard deviation also indicates the lack of useful information. 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 prediction 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.

4 Conclusion & Discussion

This study confirms the feasibility of using text classification on annual reports to predict future short-term financial performance. We performed cross validation and t-test 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 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 prediction 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.

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, stock price changes may be used as dependent prediction variables. In future research, we will also test predictive models built with a given year's annual reports using the next year's reports.

5 Acknowledgment

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

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

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

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

[5] Anotonina Kloptchenko, Tomas Eklund, Barbro Back, Jonas Karlsson, Hannu Vanharanta, and Ari Visa. Combining data and text mining techniques for analyzing financial reports. *Proceedings of Eighth Americas Conference on Information Systems*, 2002.

[6] Anotonina Kloptchenko, Camilla Magnusson, Barbro Back, Ari Visa, and Hannu VAnharanta. Mining textual contents of quarterly reports. *Turku Center for Computer Science Technical Reports*, 2002.

[7] 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.

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

[9] R.K. Rogers and J. Grant. An empirical investigation of the relevance of the financial reporting process to financial analysts. *Unpulished*, 1997.

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

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

[12] Amit Singhal, Chris Buckley, and Mandar Mitra. Pivoted document length normalization. *Proceedings of the 1996 ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 21–29, 1996.

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

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

[15] Ozgur Turetken. Predicting financial performance of publicly traded Turkish firms: A comparative study. *Unpublished*, 2004.

[16] Ari Visa, Jarmo Toivonen, Piia Ruokonen, Hannu Vanharanta, and Barbro Back. Knowledge discovery from text documents based on paragraph maps. *Proceedings of the 33rd Hawaii International Conference on System Sciences*, 2000.

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