Available online at www.sciencedirect.com

ScienceDirect

The Journal of Finance and Data Science 5 (2019) 48–59

http://www.keaipublishing.com/en/journals/jfds/

Testing market response to auditor change filings: A comparison of machine learning classifiers

Richard Holowczak [a,1], David Louton [b,*,2], Hakan Saraoglu [b,2]

[a] *Computer Information Systems, Baruch College, City University of New York, Box B11-220, 1 Bernard Baruch Way, New York, NY 10010, USA*

[b] *Department of Finance, Bryant University, 1150 Douglas Pike, Smithfield, RI 02917-1284, USA*

Received 29 January 2018; revised 1 August 2018; accepted 15 August 2018
Available online 23 August 2018

Abstract

The use of textual information contained in company filings with the Securities Exchange Commission (SEC), including annual reports on Form 10-K, quarterly reports on Form 10-Q, and current reports on Form 8-K, has gained the increased attention of finance and accounting researchers. In this paper we use a set of machine learning methods to predict the market response to changes in a firm's auditor as reported in public filings. We vectorize the text of 8-K filings to test whether the resulting feature matrix can explain the sign of the market response to the filing. Specifically, using classification algorithms and a sample consisting of the Item 4.01 text of 8-K documents, which provides information on changes in auditors of companies that are registered with the SEC, we predict the sign of the cumulative abnormal return (CAR) around 8-K filing dates. We report the correct classification performance and time efficiency of the classification algorithms. Our results show some improvement over the naïve classification method.

1. Introduction

The use and interpretation of unstructured data such as text is becoming an increasingly important area of focus in accounting and finance research. Advances in machine learning techniques and the availability of large textual data sets are opening new avenues of exploration, especially when combined with traditional structured data sets such as market data. In this paper we employ a collection of machine learning techniques to predict the market response to changes in a firm's auditor as reported in public filings.

[*] Corresponding author. Fax: +401 232 6319.

E-mail addresses: richard.holowczak@baruch.cuny.edu (R. Holowczak), dlouton@bryant.edu (D. Louton), saraoglu@bryant.edu (H. Saraoglu).

Peer review under responsibility of China Science Publishing & Media Ltd.

[1] Fax: +646 312 1541.

[2] Fax: +401 232 6319.

https://doi.org/10.1016/j.jfds.2018.08.001

Market response to auditor changes of publicly traded companies has been widely studied in the accounting and finance literature. Whisenant, et al.,[34] find that reportable events disclosed in Form 8-K filings of auditor changes are considered by investors to have information content leading to a statistically significant market response. Griffin and Lont,[14] show that the main drivers of investor response to the disclosure of auditor resignations and dismissals relate more to economic fundamentals than to the auditor change attributes. They find that investors react most negatively to resignations and much less for dismissal announcements. The market response increases for companies that have prior securities litigation and higher risk of bankruptcy. Hennes, Leone, and Miller [16] study the conditions that relate to financial restatements that eventually lead to the dismissal of external auditors, and assess the market response to the dismissal announcements. They report that auditors are more likely to be dismissed after more severe restatements. They also show that the market reaction to the dismissal is significantly more positive following more severe restatements.[2] Aldhizer et al,[2] examine whether auditor realignment disclosures filed in Form 8-K (Item 4.01) have information content to investors using a post Sarbanes-Oxley (SOX) sample. Their results show that internal control, material weakness and non-reliance on management representation disclosures carry negative information content, while audit scope limitation, earnings restatement and client–auditor disagreement disclosures do not have information content. Chang et al,[6] find that stock market reaction is relatively more positive to clients switching from a Big 4 auditor to a smaller third-tier auditor after major regulatory changes.

A common thread in the above-mentioned studies is that their authors rely on processing textual data that is contained in companies' SEC filings to frame their statistical analyses. In other words, understanding the context of an auditor change announcement requires reading the content of Item 4.01 in 8-K filings and transforming the information in the document's text into a set of variables that will be used in the analysis. For example, Audit Analytics, which provides detailed research on over 150,000 active audits and more than 10,000 accounting firms, processes text documents using extensive human involvement to read thousands of dense documents. We examine whether, and to what extent, machine learning can be helpful in interpreting a large collection of 8-K filings with minimal human interaction. Specifically, we use text classification to predict whether the text of a particular 8-K Item 4.01 filing will be associated with a positive or negative market response.

Machine learning methods that use text to explain a real-valued variable or classify observations into categories have been applied to such tasks as predicting author's age from a text,[17,26] predicting opening weekend revenue for a movie using the text of film critics' reviews,[20] and predicting online review helpfulness.[25] Algorithms have also been developed for classifying textual data into binary or multiple categories with applications in e-mail spam filtering,[3,9,28,31] e-mail classification into categories,[7,8,21,29] and classification of recommendations and reviews.[12,23] Joachims presents a detailed description of using the Support Vector Machines (SVM) in text classification.[19] Aggarwal and Zhai provide a survey of text classification algorithms.[1]

Textual analysis has also been increasingly applied to studies in finance and economics. While sentiment analysis of financial texts for prediction purposes has been widely studied,[5,24] use of text data to classify observations or explain real-valued variables in finance is still in its early stages. Sun et al[32] investigate the potential use of textual information from user-generated microblogs to predict the stock market. Foster et al[10] use text to predict real estate prices. Shimon Kogan, et al[22] apply well-known regression techniques to a large corpus of freely available financial reports and construct regression models to explain volatility for the period following a report. Trusov et al[33] use multiple text representations in a regression framework to predict financial risks. Frankel et al[11] examine the usefulness of support vector regressions (SVRs) in assessing the content of unstructured, qualitative disclosures by relating Management's Discussion and Analysis accruals to actual accruals. Antweiler and Frank[4] use computational linguistics methods to measure the bullishness of the messages on internet message boards. They find significant predictive content from message posting to trading volume, from message posting to volatility and from the degree of message bullishness to trading volume. Guo et al[15] review the literature and describe different methods used in textual analysis, especially machine learning. Gentzkow et al[13] also provide an overview of methods for analyzing text and a survey of current applications in economics and related social sciences.

We contribute to the finance literature by: (1) applying a vectorization algorithm to the problem of feature extraction from financial texts; and (2) exploring the relative performance of different machine learning algorithms in classifying financial texts based on the features extracted from them. Our paper is organized as follows. Section 2 describes our sample. Section 3 explains our methodology and the algorithms we use in the process, and provides the results of our study. Section 4 provides a sensitivity analysis and section 5 presents a summary of our findings and directions for future research.

2. Data

We collect 8-K and 8-K/A (amended) filings that contain Item 4.01 information from the SEC's EDGAR database for the period February 2001 to December 2016 using a custom Linux shell script. As 8-K documents are typically loosely formatted using HTML markup tags, we employ a custom Perl script to locate and extract the Central Index Key (CIK), Company Name, File Date and the full text of Item 4.01. The current mandated format of 8-K documents (see https://www.sec.gov/files/form8-k.pdf) that employ separate item numbers as headings (e.g., Item 1.01, Item 1.02, etc.) appears to have been taken up in mid-2003. As a result, we are able to locate and extract many more Item 4.01 sections using post 2003 data. Additional filters drop entries containing only placeholder information or boilerplate text (e.g., "Item 4.01 Changes in Registrant's Certifying Accountant. Not Applicable.").

Some 8-K documents include other items that may be material and thus impact financial markets, but these cases would tend to obscure the market response to Item 4.01 filings. We therefore drop any observations that include items in addition to Item 4.01 and 9.01 Exhibits. Item 9.01 is often used to amplify and explain material disclosed under Item 4.01. Finally, we drop any Item 4.01 filings occurring within 30 days of the previous Item 4.01 filing from the same firm. After applying these filters, our sample includes 12,435 documents.

In order to measure market response to Item 4.01 8-K filings, we merge our sample with stock returns data from the Center for Research in Security Prices (CRSP). A significant number of 8-K filings are generated by wholly owned subsidiaries, private firms that meet the SEC regulatory thresholds in terms of assets or number of shareholders and by other entities that are not exchange traded. These filings are not useful in the context of our study. In addition, we impose a 180-day market model estimation period preceding the event window for each filing. This has the effect of excluding firms that recently began trading under their current ticker symbol. After applying these important filters we are left with a final sample consisting of 3509 usable observations.

3. Methodology

The first step in our analysis is to assess market response to the announcement of auditor changes in companies. To do so, we use the single index model to calculate cumulative abnormal returns (CARs) around the public disclosure date of the change, with a window length of 9 days. Specifically, we calculate the CARs during a period of 4 days before the event day, the event day itself, and 4 days afterward. The event window length that we select is in line with the length employed in previous market response studies where it varies between 2 days and 11 days.[3] As the main purpose of our study is to test the performance of different machine learning classifiers applied to SEC filings, we start out by assigning the CAR of each observation i, based on its sign, to a positive market response or negative market response category denoted by y.

$$y_i = \begin{cases} 1, & \textit{if } CAR_i > 0 \\ 0, & \textit{otherwise} \end{cases} \tag{1}$$

The resulting data set S_N contains 3509 pairs of market response category and Item 4.01 text documents that correspond to the filings around which the CARs are calculated:

$$S_N = (\text{text\_document}_i, y_i) \; \textit{for } i = 1, \ldots, N_S,$$

where N_s is the number of observations in our sample. Table 1 presents the summary statistics of the CARs and market response categories based on the signs of CARs, Table 2 presents a breakdown of the market response variables for training and testing subsets, and Table 7 shows two representative observations from the dataset.

In the next step, we divide the sample randomly into a training subsample and a test subsample. We initially divide the sample in half as our base case where training and test samples contain 50 percent of the observations each. We explore the impact of different training-testing subsample proportions in section 4.

Using the training sample, we tokenize the Item 4.01 text into words and remove punctuation and stop words such as *the*, *a*, *an*, *and*, *but*, *if*, *of*, etc. from the text. Then, we transform the tokenized text of the complete set of documents into a matrix of numerical values using the term frequency–inverse document frequency (tf-idf) transform (see[30]). Elements of the matrix are the tf–idf weights where each weight represents the importance of a word to a specific document in the corpus. The tf-idf weight of a

[3] Companies have four business days to file a Form 8-K for the events specified in the items in Sections 1–6 and 9 above. https://www.sec.gov/fast-answers/answersform8khtm.html. The existing literature uses a similar event window. Furthermore, larger windows will introduce confounding factors, which can potentially bias the results.

Table 1
Descriptive Statistics of the Cumulative Abnormal Returns (CARs) and CAR Categories.

	CARs	CAR Categories
N	3509	3509
Mean	0.064 (2.8687, 0.0041)	0.489
Standard deviation	0.132	0.500
Skewness	6.521	0.044
Kurtosis	116.147	−1.999
Minimum	−0.787	0
25th percentile	−0.040	0
50th percentile	−0.002	0
75th percentile	0.043	1
Maximum	2.966	1

We present the descriptive statistics of the cumulative abnormal returns (CARs) and CAR categories that are represented by the sign of the CAR. We assign the CAR of each observation *i*, based on its sign, to a positive market response or negative market response category denoted by *y*.

$$y_i = \begin{cases} 1, & if\ CAR_i > 0 \\ 0, & otherwise \end{cases}$$

The t-value and *p*-value, respectively, corresponding to the mean of CARs are shown in parentheses.

word *i* in document *j* is obtained by multiplying term frequency (TF), which is the number of times word *i* appears in document *j*, by inverse document frequency (IDF), which represents the frequency of the word, or token, in the set of documents, as follows:

$$\mathrm{TF} - \mathrm{IDF}_{i,j} = \mathrm{TF}_{i,j} \times \mathrm{IDF}_i, \tag{2}$$

where

$$\mathrm{IDF}_i = \log\left(\frac{1+N_T}{1+\mathrm{DF}_i}\right) + 1, \tag{3}$$

N_T is the number of documents in the training sample and DF_i is the number documents in which the word, or token, *i* appears.

Dimensions of the resulting matrix **X**, which we will call feature matrix henceforth, are $[N_T, p]$ where p is the number of tokens included in the vocabulary derived from the document corpus consisting of all the 8K − Item 4.01 texts included in our training sample. We include not only individual words, but also word sequences (tuples) of dimension 2 through 6 because their discriminatory power is greater than that of individual words alone. This results in $p = 1{,}056{,}128$. Putting the feature matrix **X** and the vector of market response categories *y* together, we obtain a training sample of features−category pairs denoted by T_{N_T}:

$$T_{N_T} = \left(\left(x_{i1}, \ldots, x_{ip}\right), y_i\right) for\ \ i = 1, \ldots, N_T, \tag{4}$$

where x_{ij} is the tf-idf weight of feature *j* for observation *i* and *p* is the number of features.

In the next step, we use the tf-idf matrix and the market response categories from the training sample to fit the following classification models that are commonly used in machine learning research: (1) Ridge Classifier, (2) k-Nearest Neighbor (kNN), (3) Random Forest, (4) Linear Support Vector Classifier with L2 Penalty, and (5) Multinomial Naïve Bayes. Then, we test the models in our test sample and report classification accuracy for each model. Table 7 illustrates this process by showing the raw text for two representative observations from the dataset, along with the raw and categorical forms of the cumulative abnormal returns (CARs) associated with the event window and the predicted categorical CAR produced by each of the classifiers included in our study.

Ridge regression satisfies the following criterion:

Table 2
Categories of Cumulative Abnormal Returns (CARs) by Sign.

Category	Training Sample		Test Sample		Total Sample	
	N	Proportion	N	Proportion	N	Proportion
0	908	51.77%	885	50.43%	1793	51.10%
1	846	48.23%	870	49.57%	1716	48.90%
Total	1754	100.00%	1755	100.00%	3509	100.00%

We present the number of observations for each CAR category and their proportions in the training and test samples.

$$\min\left[\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\sum_{j=1}^{p}\beta_j^2\right], \tag{5}$$

where $y=0$ or 1 (from equation (1)) and x represents an element of the tf-idf matrix, $\lambda \geq 0$ is a tuning parameter. β_0, β_0, …, β_p are coefficients to be estimated by minimizing the regression sum of squares. The minimization problem in the above equation trades of two different criteria: (1) obtaining coefficient estimates that fit the data well, represented by the first term in the brackets, and (2) shrinking the magnitude of the coefficient estimates toward zero through the second term, which represents a shrinkage penalty. This tradeoff, which is a key feature of the ridge regression classifier, produces an accurate predicted classification in the case of both of the example observations shown in Table 7.

The role of the parameter λ is to strike a balance between the two terms in equation (5). When λ is zero, the penalty term has no impact, while as λ increases and the shrinkage penalty grows, the ridge regression will produce coefficient estimates that are closer to zero. We estimate model performance under λ values ranging from 0 to 10. These results are presented in Table 6.

The k-nearest neighbors (kNN) algorithm is a simple non-parametric method that can be used for classification and regression problems in pattern recognition. kNN classifies the observations in our test sample by the majority vote of k nearest neighbors selected from our training sample of documents. Given a test observation (X_i^*, y_i^*) with features X_i^* and category label y_i^*, the algorithm first searches and selects its k nearest neighbors from the training sample based on the Euclidean distance between the observations, yielding a set of nearest neighbors, $S_{k,NN} = ((X_1^{NN}, y_1^{NN}), \ldots, (X_k^{NN}, y_k^{NN}))$. Then, it predicts the category label $\widehat{y}_i^*$ of the test observation through the majority vote of k nearest neighbors as:

$$\widehat{y}_i^* = \underset{c}{\operatorname{argmax}} \sum_{\left(X_i^{NN}, y_i^{NN}\right) \in S_{k,NN}} \delta(c), \tag{6}$$

where $c \in C$, C is the category set, and y_i^{NN} is the category label of the ith nearest neighbor, and

$$\delta(c) = \begin{cases} 1, & if\ c = y_i^{NN} \\ 0, & otherwise \end{cases}. \tag{7}$$

As can be seen in Table 7, the k nearest neighbors classifier only predicts the correct classification for one of the two examples.

We estimate model performance under k values ranging from 5 to 105. These results are presented in Table 6 and discussed in the Sensitivity Analysis section.

Random forests, which combine decision trees that are used in statistical learning to improve their prediction performance, are particularly suitable for classifying text documents as they can deal with a large number of features. We classify the documents in our test sample by using the following random forest algorithm. We build a number of decision trees from a sample drawn with replacement from our training sample of documents. When we construct a decision tree, each time a split in the tree is considered, we select a random sample of m features from the full set of p predictors in the feature vector, as split candidates. The split then uses one of the m features while a new sample of m features is used at each split. The algorithm combines features by averaging their probabilistic prediction. Using random subsets extracted from the full features set, results in a slight increase in the bias of the random forest as compared to the bias of a single decision tree. However, averaging decreases the variance and results in a better model. Our random forest classifier accurately classifies both of the example observations shown in Table 7.

The support vector classifier (SVC) classifies a test observation depending on which side of a hyperplane it lies. The hyperplane is chosen to separate most of the training observations into two classes, but may misclassify a few observations as follows[18]:

$$\max_{\beta_0,\beta_1,\ldots,\beta_p,\varepsilon_0,\varepsilon_1,\ldots,\varepsilon_p} M \tag{8}$$

$$\text{subject to } \sum_{j=1}^{p}\beta_j^2 = 1,$$

$$y_i\left(\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+\ldots+\beta_p x_{ip}\right) \geq M(1-\varepsilon_i),$$

$$y_i = [0,1]$$

$$\varepsilon_i \geq 0,$$

$$\sum_{j=1}^{n}\varepsilon_i \leq C,$$

where C is a tuning parameter that is non-negative and controls the trade-off between bias and variance in classification. M is the width of the margin that surrounds the separating hyperplane and $\varepsilon_1, \ldots, \varepsilon_n$ are slack variables that allow individual observations to be on the wrong side of the hyperplane or the margin. After the solution to the maximization problem is obtained, an observation X^* from the test sample can be classified based on the sign of $\widehat{y}(X^*) = \beta_0 + \beta_1 x_1^* + \beta_2 x_2^* + \ldots + \beta_p x_p^*$. The SVC classifier accurately predicts the classification of both of the example observations shown in Table 7.

Naive Bayes classifiers apply the Bayes' theorem to the classification problems by using the assumption that the elements of the feature vector are independent of each other. Bayes' theorem is represented by the following equation:

$$P(y|x_1, \ldots, x_n) = \frac{P(y)P(x_1, \ldots, x_n|y)}{P(x_1, \ldots, x_n)}, \tag{9}$$

where y denotes the class variable y and $x_1, \ldots, x_n$ represent the features in the sample. The independence assumption implies that:

$$P(x_i|y, x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n) = P(x_i|y), for\ i = 1, \ldots, n. \tag{10}$$

Based on the previous equation, we can write equation (9) as:

$$P(y|x_1, \ldots, x_n) = \frac{P(y)\prod_{i=1}^{n} P(x_i|y)}{P(x_1, \ldots, x_n)}. \tag{11}$$

As $P(x_1, \ldots, x_n)$ is constant given the sample, $P(y|x_1, \ldots, x_n)$ is proportional to $P(y)\prod_{i=1}^{n} P(x_i|y)$, and we can classify the observations using the following equation:

$$\widehat{y} = \underset{y}{\operatorname{argmax}} P(y)\prod_{i=1}^{n} P(x_i|y), \tag{12}$$

and estimating $P(y)$, which is the relative frequency of y in the training sample, and $P(x_i|y)$, which is the probability of feature i appearing in a sample belonging to y. Multinomial Naïve Bayes, which is implemented with multinomially distributed features, can be used with features that are represented by tf-idf scores of the text sample. The multinomial naïve Bayes classifier accurately predicts the correct classification of both of the example observations shown in Table 7.

The following algorithm, which we implement using the scikit-learn[27] module in the Python programming language, summarizes our methodology:

1. Divide the text sample into two subsamples: (1) 4.01_train (Train, T) and (2) 4.01_test (Query, Q).
2. Obtain the token matrix $\mathbf{X}_T$, which can be represented as a SciPy sparse matrix, from subsample 4.01_train using the tf-idf transform. The shape of $\mathbf{X}_T$ will be $[N_T, p]$ where N_T is the number of documents in 4.01_train and p is the number of words and tuples of dimension 2 through 6 obtained from the transformation of 4.01_train.
3. Obtain the classification vector y_T for the training sample where $y_{T,i} = 1$ if $\mathrm{CAR}_{T,i} > 0$, $y_{T,i} = 0$ otherwise, for $i = 1$ to N_T.
4. Use $\mathbf{X}_T$ and y_T to fit a classification model. Use (1) Ridge Classifier, (2) k-Nearest Neighbor (kNN), (3) Random Forest, (4) Linear Support Vector Classifier with L2 Penalty, and (5) Multinomial Naïve Bayes.
5. Apply the tf-idf transform to obtain the token matrix $\mathbf{X}_Q$ from subsample 4.01_test using the same words (vocabulary) obtained from the transformation of 4.01_train. The shape of $\mathbf{X}_Q$ will be $[N_Q, p]$ where N_Q is the number of documents in 4.01_test and p is the number of words.
6. Use the transformed matrix $\mathbf{X}_Q$ and the fit classifier model obtained from the training sample to predict the classification $\widehat{y}_Q$ of each observation in the test sample.
7. Determine the classification accuracy of the model using $\widehat{y}_Q$ and y_Q.
8. Repeat 4–7 using a different classifier model.
9. Report process time and classification accuracy results from each classifier model used.

The results are shown in Table 3 and Table 4. For the base case, using a training sample comprised of 50 percent of the observations, the Linear Support Vector Classifier with L2 penalty performs the best among the five models with a classification accuracy of 54.5%.[4] It should be noted, however, that the accuracy rate that is achieved by the Linear Support Vector Classifier with L2 penalty represents only a minor improvement over the base rate that can be achieved using a naïve classification method, which would equal the proportion of the more frequent class in the sample. In this case, that rate, which is shown in Table 3, is around 50%.

[4] We also ran the Linear Support Vector Classifier using the L1 penalty, which is less sensitive to the presence of outliers. This did not result in a noteworthy difference in classification accuracy.

Table 3
Confusion Matrices.

	Prediction = 0	Prediction = 1	Total
Ridge Classifier			
Actual = 0	534 (60.34%) (53.94%)	351 (39.66%) (45.88%)	885
Actual = 1	456 (52.41%) (46.06%)	414 (47.59%) (54.12%)	870
Total	990	765	1755
k-Nearest Neighbor (kNN)			
Actual = 0	867 (97.97%) (50.35%)	18 (2.03%) (54.55%)	885
Actual = 1	855 (98.28%) (49.65%)	15 (1.72%) (45.45%)	870
Total	1722	33	1755
Random Forest			
Actual = 0	503 (56.84%) (54.03%)	382 (43.16%) (46.36%)	885
Actual = 1	428 (49.20%) (45.97%)	442 (50.80%) (53.64%)	870
Total	931	824	1755
Linear Support Vector Classifier with L2 Penalty			
Actual = 0	533 (60.23%) (54.39%)	352 (39.77%) (45.42%)	885
Actual = 1	447 (51.38%) (45.61%)	423 (48.62%) (54.58%)	870
Total	980	775	1755
Multinomial Naïve Bayes			
Actual = 0	496 (56.05%) (53.91%)	389 (43.95%) (46.59%)	885
Actual = 1	424 (48.74%) (46.09%)	446 (51.26%) (53.41%)	870
Total	920	835	1755

For each machine learning algorithm that we use in our study, we present confusion matrices of the classification results corresponding to the test sample. The results in this table correspond to the base case for each classification algorithm, which uses 50% of the sample for training and 50% for testing.

Table 4
Performance of Classifiers.

	Training Time (sec)	Test Time (sec)	Accuracy
Ridge Classifier	0.0063	0.7625	55.58%
k-Nearest Neighbor (kNN)	0.6891	0.0109	51.26%
Random Forest	1.1766	17.1199	54.11%
Linear Support Vector Classifier with L2 Penalty	0.0031	0.5469	55.58%
Multinomial Naïve Bayes	0.0281	0.0672	56.55%

We present performance of each classifier in terms of (1) training time, (2) test time, and (3) classification accuracy. The performance results in this table correspond to the base case for each classification algorithm, which uses 50% of the sample for training and 50% for testing.

4. Sensitivity analysis

To test the sensitivity of classification performance to a change in parameter values, we first run our analysis with different cut-off points for dividing the sample into training and test subsamples. Then, for each machine learning algorithm in our analysis, we vary the key input parameters and measure the potential improvement in classification accuracy.

As the value of the cut-off point for obtaining the training and test subsamples is an important parameter, we begin our sensitivity analysis by checking the impact of varying subsample sizes on performance. First, we use a cut-off point where 10 percent of the sample is selected as the training subsample and 90 percent as the test subsample.

Then, we increment the cut-off point by 10 percentage points up to a 90 percent-10 percent division of the sample to training and test subsamples, respectively. For each subsample size, we randomly form 10 different test and training subsample pairs and report the group means of the classification accuracy percentage. Table 5 shows that the Linear Support Vector Classifier with L2 penalty achieves a 61.14 percent classification accuracy when the training proportion of the dataset is 90 percent. This is more than a 10-percentage point improvement over the base rate accomplished by the naïve classification method.

We also conduct sensitivity analysis of our results with different values of the classification algorithm parameters. In doing so, we keep the training sample size at 90 percent of the overall sample. These results are shown in Table 6. For Ridge Regression, we set the value of the tuning parameter λ to 1 for the base case. Larger values of λ correspond to higher shrinkage in coefficient estimates and lower variance. We vary λ in the interval [0, 10] with a step size of 1 and test the performance of the ridge classifier for each value of the λ parameter.

Table 5
Classification Accuracy for Different Training Sample Proportions.

	Training Sample Proportions								
	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
Ridge Classifier	51.33%	52.07%	53.57%	54.46%	55.58%	56.05%	56.71%	58.05%	58.43%
k-Nearest Neighbor (kNN)	50.72%	50.92%	51.09%	51.84%	51.26%	51.75%	51.79%	49.84%	50.31%
Random Forest	51.15%	52.27%	52.93%	53.15%	54.11%	54.94%	55.27%	56.45%	56.24%
Linear Support Vector Classifier with L2 Penalty	51.37%	52.22%	53.50%	54.70%	55.58%	56.15%	56.72%	57.72%	58.21%
Multinomial Naive Bayes	51.73%	52.49%	54.60%	55.93%	56.55%	58.00%	58.47%	59.74%	61.00%

We present the sensitivity of classification accuracy to the size of training sample. First, we use a cut-off point where 10 percent of the sample is selected as the training subsample and 90 percent as the test subsample. Then, we increment the cut-off point by 10 percentage points up to a 90 percent-10 percent division of the sample to training and test subsamples, respectively. For each subsample size, we randomly form 10 different test and training subsample pairs and report the group means of the classification accuracy percentage.

Table 6
Sensitivity Analysis Results.

Ridge		kNN		Random Forest		LSVC		Multinomial NB	
λ	Score	k	Score	Maximum m	Score	C	Score	α	Score
0	58.06%	5	52.62%	5	57.66%	0.1	57.69%	0	60.28%
1	57.55%	10	52.19%	10	59.09%	1	59.26%	0.1	59.74%
2	57.49%	15	51.97%	15	57.75%	10	59.40%	0.2	59.60%
3	56.64%	20	51.79%	20	57.61%	100	59.63%	0.3	59.63%
4	56.64%	25	51.71%	25	56.52%	1000	59.00%	0.4	59.97%
5	56.18%	30	51.77%	30	57.92%			0.5	59.52%
6	55.95%	35	52.14%	35	56.30%			0.6	59.52%
7	55.81%	40	52.22%	40	55.90%			0.7	59.06%
8	55.61%	45	52.05%	45	56.41%			0.8	59.29%
9	55.53%	50	52.17%	50	56.84%			0.9	59.12%
10	55.56%	55	52.48%	55	57.04%			1.0	59.06%
		60	52.08%	60	56.84%				
		65	52.42%	65	57.81%				
		70	51.62%	70	57.52%				
		75	51.42%	75	56.72%				
		80	51.82%	80	56.50%				
		85	52.31%	85	56.58%				
		90	51.79%	90	56.58%				
		95	52.25%	95	57.15%				
		100	52.11%						
		105	51.77%						

We conduct sensitivity analysis by varying the key parameter values for each classification algorithm. We present the classification accuracy corresponding to different values of the key parameters. We keep the training sample size at 90 percent of the overall sample for conducting the sensitivity analysis.

We measure the sensitivity of the performance of the kNN algorithm to the number of nearest neighbors by varying k between 5 and 105, with a step size of 5. Although there is a local maximum at around $k = 5$, classifier performance does not vary significantly over this range.

We test the performance sensitivity of the linear support vector classifier by varying the tuning parameter C that controls the trade-off between bias and variance in classification. Larger values of C correspond to a smaller-margin hyperplane and smaller C values will result in a larger-margin hyperplane. We use values for C that are in the set [0.1, 1, 10, 100, 1000] where a C value of 1 corresponds to the base case of our analysis. The local maximum classification score for this range occurs at $C = 100$. However, in our application of the linear support vector classifier, performance is not particularly sensitive to the choice of C parameter.

Table 7
Examples of 8K - Item 4.01 filings and Corresponding Classifier Predictions.

CAR	Categorical CAR	Predicted Classification				
		Ridge Classifier	k-Nearest Neighbors	Random Forest	Linear Support Vector	Multinomial Naïve Bayes
Example 1 − FairPoint − 3/20/2008						
10.86%	1	1	0	1	1	1

Item 4.01 text

ITEM 4.01 CHANGES IN Certifying Accountant Previous Independent Public Accounting Firm KPMG LLP was previously the principal accountants for FairPoint Communications, Inc. and subsidiaries ("FairPoint" or the "Company"). On March 18, 2008, KPMG LLP was dismissed and Ernst & Young LLP was engaged as principal accountants. The decision to change accountants was approved by the audit committee of the board of directors of FairPoint. During the two fiscal years ended December 31, 2007, and in the subsequent interim period through March 18, 2008, there were no: (1) disagreements with KPMG LLP on any matter of accounting principles or practices, financial statement disclosure, or auditing scope or procedures, which disagreements if not resolved to their satisfaction would have caused them to make reference in connection with their opinion to the subject matter of the disagreement, or (2) reportable events, except that KPMG LLP advised FairPoint of the following material weakness: management oversight and review procedures designed to monitor the effectiveness of control activities in the northern New England division were ineffective. The audit reports of KPMG LLP on the consolidated financial statements of FairPoint Communications, Inc. and subsidiaries as of December 31, 2007 and 2006 and for the three years ended December 31, 2007 did not contain any adverse opinion or disclaimer of opinion, nor were they qualified or modified as to uncertainty, audit scope or accounting principles, except as follows: KPMG LLP's report on the consolidated financial statements of FairPoint as of and for the three years ended December 31, 2007, contained a separate paragraph stating that "As discussed in note 2 to the consolidated financial statements, the Company adopted the provisions of Financial Accounting Standards Board ("FASB″) Interpretation No. 48, Accounting for Uncertainty in Income Taxes an interpretation of FASB Statement No. 109, effective January 1, 2007 and the provisions of Statement of Financial Accounting Standards (SFAS) No. 123 (revised 2004), Share-Based Payment, effective January 1, 2006." The audit report of KPMG LLP on the effectiveness of internal control over financial reporting as of December 31, 2007 did not contain any adverse opinion or disclaimer of opinion, nor were they qualified or modified as to uncertainty, audit scope, or accounting principles, except that KPMG LLP's report indicates that FairPoint did not maintain effective internal control over financial reporting as of December 31, 2007 because of the effect of a material weakness on the achievement of the objectives of the control criteria and contains an explanatory paragraph that states management oversight and review procedures designed to monitor the effectiveness of control activities in the northern New England division were ineffective. A letter from KPMG LLP is attached as Exhibit 16.1 to this Current Report on Form 8-K. New Independent Registered Public Accounting Firm On March 18, 2008, FairPoint approved the engagement of Ernst & Young LLP ("E&Y″) as its new independent registered public accounting firm to audit FairPoint's financial statements for the year ending December 31, 2008 and to review the financial statements to be included in 2** FairPoint's quarterly reports on Form 10-Q for each of the financial quarters of 2008. The decision to engage E&Y as FairPoint's independent registered public accounting firm was approved by the Audit Committee. E&Y audited Spinco's financial statements for the years ended December 31, 2005, 2006 and 2007. Following the Merger, Spinco will be treated as the acquiror in the Merger for accounting purposes. Except for E&Y's role as the independent registered public accounting firm for Spinco and except that, in the role as the independent registered public accounting firm of Spinco, E&Y has audited the financial statements that will become the historical financial statements of FairPoint, prior to the engagement of E&Y, neither FairPoint nor anyone on behalf of FairPoint consulted with E&Y during FairPoint's two most recent fiscal years and through the subsequent interim period regarding either: 1. the application of accounting principles to any specified transaction, either completed or proposed, or the type of audit opinion that might be rendered on FairPoint's financial statements (as described in Item 304 (a) (2) (i) of Regulation S-K); or 2. any matter that was either a subject of disagreement or event (as defined in Item 304 (a) (1) (iv) of Regulation S-K and the related instruction to Item 304), or a reportable event (as described in Item 304 (a) (1) (v) of Regulation S-K). FairPoint has participated in discussions with E&Y, in its capacity as Spinco's auditors, in connection with certain discussions regarding the potential impact of the Merger on FairPoint's 2008 financial statements.

(*continued on next page*)

Table 7 (*continued*)

CAR	Categorical CAR	Predicted Classification				
		Ridge Classifier	k-Nearest Neighbors	Random Forest	Linear Support Vector	Multinomial Naïve Bayes
Example 2 − Varsity Group − 6/26/2007						
−4.59%	0	0	0	0	0	0

Item 4.01 text

Item 4.01. Changes in Registrant's Certifying Accountant On June 21, 2007, Varsity Group Inc., (the "Company") dismissed PricewaterhouseCoopers LLP (the "former auditor"), as its independent registered public accounting firm. Effective June 21, 2007, the Company engaged McGladrey & Pullen, LLP as its new independent registered public accounting firm. The Company's board of directors has approved the dismissal of the former auditor, and the appointment of McGladrey & Pullen, LLP as its new independent registered public accounting firm. The reports of the former auditor on the Company's financial statements for the years ended December 31, 2006 and December 31, 2005 contained neither an adverse opinion, or a disclaimer of opinion, and were not qualified or modified as to uncertainty, audit scope or accounting principle. During the years ended December 31, 2006 and December 31, 2005 and through June 21, 2007, there were no disagreements with the former auditor on any matter of accounting principles or practices, financial statement disclosure or auditing scope or procedure, which disagreements, if not resolved to the former auditors satisfaction, would have caused them to make reference thereto in their reports on the Company's consolidated financial statements for such years. During the years ended December 31, 2006 and December 31, 2005 and through June 21, 2007, there were no reportable events, as defined in Item 304(a)(1)(v) of Regulation S-K, except where noted below: As of December 31, 2005, March 31, 2006, June 30, 2006, September 30, 2006, December 31, 2006 and March 31, 2007, the Company did not maintain effective controls over the accuracy of the calculation of earnings per share. Effective controls were not in place over the calculation of diluted shares outstanding for purposes of calculating diluted earnings per share. This control deficiency resulted in a computational error of the number of shares to be assumed as repurchased in the application of the treasury stock method that was not prevented or detected. Additionally, this control deficiency could result in a misstatement of earnings per share that would have resulted in a material misstatement to annual or interim financial statements that would not have been prevented or detected. Accordingly, management has determined that this control deficiency constitutes a material weakness. The Company has previously disclosed this control deficiency with the SEC. The Company has authorized the former auditors to respond fully to the inquiries of McGladrey & Pullen, LLP relating to the material weakness. The Company provided the former auditor with a copy of this Current Report on form 8-K and requested that they furnish us with a letter addressed to the SEC stating whether or not they agree with the above statements. A copy of this letter is filed as an exhibit to this Form 8-K. During the years ended December 31, 2006 and December 31, 2005 and through June 21, 2007, the Company did not consult McGladrey & Pullen, LLP regarding either: 1) the application of accounting principles to any specified transaction, either completed or proposed, or the type of audit opinion that might be rendered on the Company's financial statements, and neither a written report was provided to the Company nor oral advice was provided that McGladrey & Pullen, LLP concluded was an important factor considered by the Company in reaching a decision as to the accounting, auditing or financial reporting issue; or 2) any matter that was either the subject of disagreement, as that term is defined in Item 304(a)(1)(iv) of Regulation S-K and the related instructions to Item 304 of Regulation S-K, or a reportable event, as that term is defined in Item 304(a)(1)(v) of Regulation SK.

The two examples included in this table were chosen because they contain relatively direct language and were associated with non-trivial cumulative average residual returns (CARs) during the event window. Apart from that, they are typical of the 8K − Item 4.01 filings included in our sample. The classifiers were trained on a randomly chosen sub-sample consisting of 90 percent of the observations in our dataset.

For the random forest algorithm, we test for a potential change in performance by varying m, the maximum number of features used at each split, from 5 to 95, with a step size of 5. Although we do not find significant variation in classifier performance over this range of m values, there is a modest improvement at around $m = 30$.

In Multinomial Naïve Bayes, the probability of a feature appearing in a sample belonging to a class is estimated by maximum likelihood. In the base case, we use smoothed maximum likelihood and assign a value of 0.01 to the smoothing parameter a, which corresponds to Lidstone smoothing. To perform sensitivity analysis for the performance of the Multinomial Naïve Bayes algorithm, we vary the smoothing parameter in the interval of [0, 1], with a step size of 0.1, where a value of 0 implies no smoothing and a value of 1 is associated with Laplace smoothing. Performance of this classifier does not vary significantly over the range of a values investigated.

The results of our sensitivity analysis are summarized in Table 6. We find that the ridge regression classifier is sensitive to a change in the value of parameter λ. Increasing λ from its base case value of 1 reduces the performance of the ridge classifier in our sample, while a value of 0 results in an improvement. Classification performance of kNN, Random Forest, SVC, and Multinomial NB are not as sensitive to changes in their respective tuning parameters.

5. Conclusion and directions for future research

In this study, we implement an algorithm that vectorizes the text of 8-K filings to test whether the resulting matrix can explain the direction of the market response to the filing using various machine learning classifiers. Our sample consists of the Item 4.01 text of 8-K documents, which provide information on changes in auditors of companies that are registered with the Securities Exchange Commission (SEC). We report the classification performance and time efficiency of classification algorithms. Our results show an improvement over the base rate that can be accomplished by using a naïve method of random classification and would equal the proportion of the more frequent class in the sample. It should be noted that using CARs as the basis for categorizing a given Item 4.01 text into labels of 0 or 1 assumes that the market response directly reflects the content of that text. A sensitivity analysis of our results provides further insight into the potential classification performance of the models we use in our analysis.

Our work in progress to improve and expand this study includes: (1) using a different proxy for categorizing the Item 4.01 texts, for example a classification score assigned to the texts in the training sample by human judgement, and (2) going beyond a simple bag-of-words approach by incorporating context-specific phrases in the analysis.

Further research could explore the application of the algorithms that we propose in this study to different types of financial texts including other items in 8-K filings.

References

1. Aggarwal CC, Zhai CX. A survey of text classification algorithms. In: *Mining Text Data*. vol. 9781461432234. USA: Springer; 2012:163–222.
2. Aldhizer III GR, Martin DR, Cotter JF. Do markets react to required and voluntary disclosures associated with auditor realignments? *Adv Account*. June 2009;25(1):1–12.
3. Androutsopoulos I, Palioras G, Karkaletsis V, Sakkis G, Spyropoulos C, Stamatopoulos P. Learning to filter spam e-mail: a comparison of a naïve Bayesian and memory-based approach. In: *Proc. 4th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD)*. 2000:1–13.
4. Antweiler W, Frank MZ. Is all that talk just noise? The information content of internet stock message boards. *J Finance*. 2004;59(3):1259–1294.
5. Bodnaruk A, Loughran, McDonald B. Using 10-K text to gauge financial constraints. *J Financ Quant Anal*. August 2015;50(4):623–646.
6. Chang H, Cheng CSA, Reichelt KJ. Market reaction to auditor switching from big 4 to third-tier small accounting firms. *Audit: J Pract Theor*. November 2010;29(2):83–114.
7. Cohen WW. Learning trees and rules with set-valued features. In: *Proceedings of the Thirteenth National Conference on Artificial Intelligence*. vol. 1. AAAI Press; 1996:709–716.
8. de Carvalho VR, Cohen W. On the collective classification of email 'speech acts'. In: *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. New York, NY, USA: ACM; 2005:345–352.
9. Drucker H, Wu D, Vapnik V. Support vector machines for spam categorization. *IEEE Trans Neural Network*. 1999;10(5):1048–1054.
10. Foster DP, Liberman M, Stine AR. *Featurizing Text: Converting Text into Predictors Regression Analysis*. Working Paper. Department of Statistics, University of Pennsylvania; 2013:1–37. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.591.6715.
11. Frankel R, Jennings J, Lee J. Using unstructured and qualitative disclosures to explain accruals. *J Account Econ*. November–December 2016;62(2-3):209–227.
12. Ganu G, Elhadad N, Marian A. Beyond the stars: improving rating predictions using review text content. In: *Proceedings of the 12th International Workshop on the Web and Databases*. vols. 1–6. 2009.
13. Gentzkow M, Kelly BT, Taddy M. *Text as Data. NBER Working Paper No. 23276*; March 2017. http://www.nber.org/papers/w23276.
14. Griffin PA, Lont DH. Do investors care about auditor dismissals and resignations? What drives the response? *Audit: J Pract Theor*. November 2010;29(2):189–214.
15. Guo L, Shi F, Tu J. Textual analysis and machine leaning: crack unstructured data in finance and accounting. *J Finance Data Sci*. September 2016;2(3):153–170.
16. Hennes KM, Leone AJ, Miller BP. Determinants and market consequences of auditor dismissals after accounting restatements. *Account Rev*. May 2014;89(3):1051–1082.
17. Hong J, Mattmann C, Ramirez P. *Ensemble Maximum Entropy Classification and Linear Regression for Author Age Prediction*. CoRR; October 2016. https://arxiv.org/abs/1610.00852. Accessed September 16, 2017.
18. James G, Witten D, Hastie T, Tibshirani R. *An Introduction to Statistical Learning with Applications in R*. New York: Springer; 2013.
19. Joachims T. *Learning to Classify Text Using Support Vector Machines – Methods, Theory, and Algorithms*. Kluwer/Springer; 2002.
20. Joshi M, Das D, Gimpel K, Smith NA. *Movie Reviews and Revenues: An Experiment in Text Regression*. PA 15213, USA: Language Technologies Institute, Carnegie Mellon University Pittsburgh; 2010.
21. Kiritchenko S, Matwin S. Email classification with co-training. Email classification with co-training. In: Stewart Darlene A, Howard Johnson J, eds. *Proceedings of the 2001 Conference of the Centre for Advanced Studies on Collaborative Research*. vol. 8. IBM Press; 2001.
22. Kogan S, Levin D, Routledge BR, Sagi JS. Predicting risk from financial reports with regression. In: *NAACL '09 Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. 2009:272–280.
23. Koren Y, Bell R, Volinsky C. Matrix factorization techniques for recommender systems. *Computer*. August 2009;42(8):30–37.
24. Loughran T, McDonald B. When is a liability not a liability? textual analysis, dictionaries, and 10-Ks. *J Finance*. February 2011;66(1):35–65.
25. Ngo-Ye TL, Sinha AP. The influence of reviewer engagement characteristics on online review helpfulness: a text regression model. *Decis Support Syst*. May 2014;61:47.

26. Nguyen D, Smith NA, Rose CP. Author age prediction from text using linear regression. In: *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*. June 2011:115−123.
27. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in Python. *J Mach Learn Res*. 2011;12:2825−2830.
28. Provost J. *Naïve-bayes Vs. Rule Learning in Classification of Email*. Technical Report AI-TR-99-284. University of Texas at Austin, Artificial Intelligence Lab; 1999.
29. Rennie JDM. ifile: an application of machine learning to e-mail filtering. In: *Proc. Of the KDD-2000 Text Mining Workshop*. 2000:95−98.
30. Salton G, Buckley C. Term-weighting approaches in automatic text retrieval. *Inf Process Manag*. December 1988;24(5):513−523.
31. Sahami M, Dumais S, Heckerman D, Horvitz E. A Bayesian approach to filtering junk e-mail. In: *Learning for Text Categorization, Papers from the AAAI Workshop*. 1998:55−62. AAAI Technical Report WS-98-05.
32. Sun A, Lachanski M, Fabozzi F. Trade the tweet: social media text mining and sparse matrix factorization for stock market prediction. *Int Rev Financ Anal*. 2016;48:272−281.
33. Trusov R, Natekin A, Kalaidin P, Ovcharenko S, Knoll A, Fazylova A. Multi-representation approach to text regression of financial risks. In: *Artificial Intelligence and Natural Language and Information Extraction, Social Media and Web Search FRUCT Conference*. 2015:110−117.
34. Whisenant S, Sankaraguruswamy S, Raghunandan K. Market reactions to disclosure of reportable events. *Audit: J Pract Theor*. March 1, 2003;22(1):181−194. March 2003.