Predicting Corporate 8-K Content Using Machine Learning Techniques

Min Ji Lee
Graduate School of Business
Stanford University
Stanford, California 94305
E-mail: minjilee@stanford.edu

Hyungjun Lee
Department of Statistics
Stanford University
Stanford, California 94305
E-mail: hyungjun@stanford.edu

***Abstract*—This project tries to predict the contents of corporate 8-K filings using Naive Bayes Assumption and Support Vector Machine. We manually classify the contents of 300 corporate 8-K filings as financial, operational, legal, administrative or HR, train the Naive Bayes Classifier and multi-class SVM based on this classification, and test their performances. Both techniques yield error rates as high as one-third due to the limited sample size. When we test binary classification (financial and non-financial), accuracy improves as more data points are fed into the classification process.**

I. Introduction

Companies are required to file form 8-K to the Securities and Exchange Commission (SEC) to notify its investors of any material events. Examples of such material events include changes in management, departure of directors, bankruptcy, and layoffs. In presence of a material event, it is likely that the market will react or has reacted to the event. Thus, we predict that there will be an association between the content of the 8-K filing and stock returns. In this project, we study the contents of 8-K filings using methods of machine learning. However, the scope of this project does not extend to analyzing their relationship with stock reactions. This project is primarily inspired by Li (2010) [1], which investigates the information content of the 'Management Discussion and Analysis' (MD&A) section in various firms' annual reports, where management discusses various aspects of the company. As with Li (2010) [1], the purpose of this project is to examine the qualitative information that can potentially be useful to investors in addition to hard data.

Specifically, we try to predict the topics covered by "Item 8.01 Other Events" in the corporate 8-K filings. Unlike the others sections and items of 8-K filings, which all have their respective topics defined in the filing requirements, Item 8.01 can cover a range of topics not covered by the other sections. Thus, predicting the topics of the 8-K sections filed under Item 8.01 could add valuable insight into understanding the content of those filings.

The project comprises three parts. The first part is to download and extract 8-K filings and to split the filings' text into words using Perl. Then, we screen 7,363 filings that contain Item 8.01 from the total of 49,137 corporate 8-K filings filed in 2012. From the 7,363 filings that contain Item 8.01, we randomly select 300 and manually classified their contents as financial, operational, legal, administrative or HR. Using the vector of words, a multi-variate Bernoulli event model is trained with Naive Bayes assumption, and a multi-class SVM is also trained. The last part is evaluating the performance of these models using an N-fold cross validation test. If either approach performs well on this data set, it could be used to classify tens of thousands of corporate 8-K filings every year and provide valuable insight into the operations of the filing firms. Throughout the process, we use Perl and R as our primary tools for analysis.

II. Background

The section on Form 8-K on SEC's website defines the requirement to file 8-K reports as the following: "*In addition to filing annual reports on Form 10-K and quarterly reports on Form 10-Q, public companies must report certain material corporate events on a more current basis. Form 8-K is the "current report" companies must file with the SEC to announce major events that shareholders should know about.*" In 2012, 7,777 different firms filed 49,137 8-K forms. This gives us an average of 6.3 filings per reporting firm while the median number of filings for reporting firms is 5. Such a positively skewed distribution is driven by a few firms that filed a large number of filings such as *21st Century*

Fox, Inc., which filed 122 8-K reports in 2012. Due to the massive number of filings, we limit our analysis to Form 8-K's filed in 2012 and do not expand it to those from other years.

Public companies are required to file 8-K reports for events including changes in management, departure of directors, bankruptcy, and layoffs. The initial goal of the study was to classify 8-K filings according to their tones - positive, negative and uncertain. However, in the process of manual classification, we come across difficulties of losing most of the filings to the 'uncertain' category, as a large fraction of filings are merely factual. Thus, we changed our gear to classifying the content of a filing and not the tone.

Form 8-K is divided into multiple sections and items, depending on the content.

> *Section 1. Registrant's Business and Operations*
> *Item 1.01 Entry into a Material Definitive Agreement*
> ⋮
> *Section 2 Financial Information*
> *Item 2.01 Completion of Acquisition or Disposition of Assets*
> ⋮
> *Section 8 Other Events*
> *Item 8.01 Other Events*

The item of our interest is *Section 8 Item 8.01 Other Events,* which covers a range of topics not covered by the other items and hence does not have a specifically defined topic, unlike all other sections. In 2012, there are 7,363 8-K filings with Item 8.01. (Note that a single filing can contain multiple items.) Since the topic of the section of Form 8-K filed under Item 8.01 cannot be predicted by its definition, we employ machine learning techniques to classify them. We expect that filings are associated with one of following five aspects: financial, operational, legal, administrative, and HR. We limit the scope of the 'financial' category to cover only financing-related activities such as equity issuance, stock repurchases, and dividend payouts. For instance, below is the following 8-K report by *Omnicare, Inc.* filed on March 1st, 2012:

> ***Item 8.01. Other Events.*** *On March 1, 2012, Omnicare, Inc. (the "Company") announced its adoption of a Rule 10b5-1 plan under which the Company may continue to repurchase its shares at times when the Company would not ordinarily be in the market due to the Company's trading policies or the possession of material non-public information. This plan has been established pursuant to, and as part of, the Company's share repurchase program....*

To list a few frequent topics that each category covers, 'operational' covers acquisition/disposition of assets, 'legal' covers lawsuits and patent issues, 'administrative' contains announcement of meetings/conferences, and 'HR' has departure or election of executives.

III. Methodology

A. Naive Bayes Classifier

As Naive Bayes classifiers are not restricted by the number of response classes, we adopt this methodology as our primary mode of analysis. We have five content classes - financial, operational, legal, administrative, and HR - in this project, whereas Li (2010) [1] uses 12 categories for the content classification. Thus Naive Bayes classifier provides greater flexibility to future extensions of this project. To prevent any arithmetic underflow problem, we take logarithms when computing Naive Bayes labels.

$$\phi_{k|y=b} = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n_i} 1(x_j^{(i)} = k \wedge y^{(i)} = b) + \alpha}{\sum_{i=1}^{m} 1(y^{(i)} = b)n_i + \alpha|V|} \quad (1)$$

We start off with the Laplace smoothing ($\alpha = 1$ in Equation (1)) of the estimated word probabilities. Then we adjust α in Equation (1) to see whether a specific degree of smoothing yields a better performance than others given the number of training data points. If α is too big, that is, when we use more aggressive smoothing, all words end up with roughly identical conditional probabilities for all categories. Thus, the conditional probabilities' contributions to log-likelihood are roughly identical for all categories and the examples end up being classified according to the prior probabilities. As the size of the training set increases, the smoothing effect of fixed alpha decreases, as more word counts are registered. If the training set size is too small and α also too small, the few words that had appeared in the training set will have unduly high conditional probabilities, introducing high variance to the resulting model and thus poor performance on the test set. Thus, picking the right smoothing parameter α given the training set size is very important to the performance of NB models.

h

TABLE I
THREE MOST FREQUENTLY ENCOUNTERED WORDS IN EACH CATEGORY

Category	Word 1	Word 2	Word 3
ADM	Park	Standard	Annual
FIN	Share	Note	Security
HR	Director	Annual	Grant
LEG	Trust	Report	Quarterly
OPR	Statement	Energy	Press, Release

B. Multi-class SVM

SVMs are natural binary classifiers and does not have an obvious extension to the multi-class cases. There are two ways that are widely used to apply SVM to multi-class problems in practice: one-versus-all (OVA) SVM and pairwise SVM. [2]

In OVA SVM, for each class of observations, an SVM separating the given class from all other classes is built, with the convention of labeling observations in the given class with +1 and those in all other classes with -1. The most widely used approach for making predictions based on OVA SVM is to choose the class whose SVM assigns the highest margin to the point in question. With OVA SVM, non-linear kernels may be very useful, for even if the decision boundary between each pair of classes are linear, the OVA SVM for a class that is sandwiched between two different classes will not perform well under linear kernels. Hence, we will try several forms of non-linear kernels to address this issue.

In pairwise SVM, given each pair of classes i and j, an SVM separating the two classes is built. Thus, for an n-class problem, a total of $n(n-1)/2$ SVMs are constructed. To make a prediction, the class to which the point in question was most frequently assigned by the $n(n-1)/2$ pairwise SVMs is chosen. More efficient algorithms for aggregating the results of the $n(n-1)/2$ pairwise SVMs have been studied in the field, although it would not be particularly relevant for our problem since we only have five classes and thus only ten pairwise comparisons to make, which, given the size of our example, is not computationally challenging.

We will train the data using both the OVA SVM and pairwise SVM and compare their performances to that of the Naive Bayes classifier.

IV. DATA PROCESSING

We download all 49,137 8-K filings made in 2012 from the SEC's EDGAR website. For each 8-K filing, we download what is indicated as "complete submission text file," which includes not only Form 8-K but also attached documents. Then we sort out 7,363 filings that contain Item 8.01. Out of these 7,363 filings, we randomly pick 300 8-K's and manually classify them into each of five aforementioned categories. As part of our data cleaning efforts, we remove all HTML tags and tables to extract only the text language. Once we extract the 8-K text, we further examine the text as it may contain information irrelevant to the content analysis. For instance, in accordance with the Private Securities Litigation Reform Act of 1995, some filings include a disclaimer that it is a forward-looking statement along with the definition of forward-looking statements as below:

> *Forward-looking statements are made based upon management's good faith expectations and beliefs concerning future developments and their potential effect upon the Company and can be identified by the use of words such as anticipated, believe, expect, plans, strategy, estimate, project, and other words of similar meaning in connection with a discussion of future operating or financial performance...*

This will effectively increase the counts of words such as "expect" and violate the Naive Bayes assumption. Thus, we exclude the paragraph containing the definition and disclaimer. We are not aware of any other common disclaimer/definition at this point. Then the truncated text is tokenized into a set of words. As a next step, we use Perl module Lingua::StopWords to remove "stop words" such as "the," "and," and "of," which appear frequently in most documents but provide a little information about the content of filings. Lastly, before moving onto the textual analysis, we use Perl module Lingua::Stem to reduce related words to a common root form; for instance, "name," "names," and "named" are replaced with "name."

Once this process of data cleaning and parsing is completely done, we move onto to the textual analysis using machine learning techniques.

V. TEXTUAL ANALYSES

Table 1 shows three most frequently appeared words in each category when the filing category is predicted using the Naive Bayes method. As we have gone through 300 filings manually and know the content, keywords in the table give us confidence that the method is picking up the right words. For instance, *"director"* can be about director appointment while *"grant"* in HR relates to compensation grants such as stock options. "FIN" mostly

h

TABLE II
CONFUSION MATRIX

Actual \Prediction	ADM	FIN	HR	LEG	OPR
ADM	17	14	2	2	7
FIN	2	78	0	2	0
HR	2	1	10	0	0
LEG	0	5	1	7	4
OPR	7	7	1	3	22

relate to firms' financing activities, and *"share"*,*"note"*, and *"security"* are the words you would expect to see filings about equity/debt issuance, share repurchases and such. However, most terms in the table are fairly generic that there is a risk of misclassification.

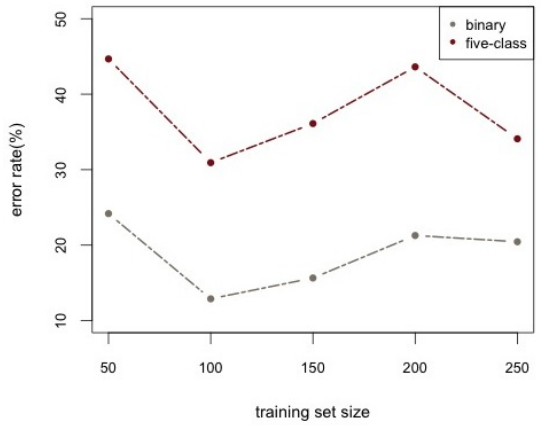

Fig. 1. Naive Bayes: 5-Class vs. Binary

Figure 1 shows when the data is classified into five categories, nearly one-third of filings are misclassified. We believe that two factors contribute to such high error rates. First, our dataset is not large enough to produce a statistically stable model. Second, almost half of filings belong to FIN, which leads to its prior probability being exceptionally high. Thus, if conditional probability for a given category is not particularly high, an observation is classified as FIN by default. The confusion matrix in *Table 2* confirms this. To resolve this issue, we examine a binary classification into financial and non-financial topics, in which setting each category contains comparable number of data points.

In case of binary classification, we test various smoothing factors as mentioned in the previous section. For a given training set size, smoothing factors can yield a difference as large as 6 percentage point in test error rate. As expected, when the training set size is small,

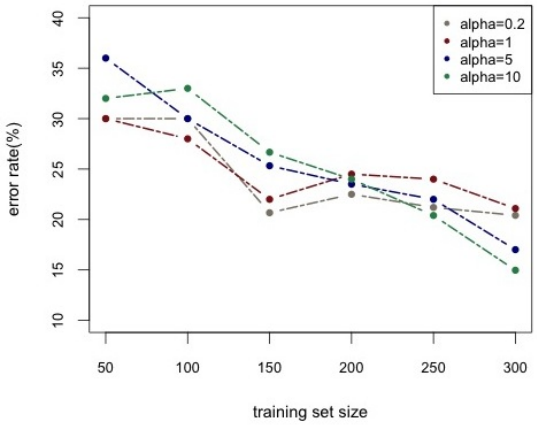

Fig. 2. Naive Bayes: 10-Fold CV with Different Smoothing Factors

smaller α (less than or equal to 1) performs better, while run on the full set, larger ones (α equal to 5 or 10) outperform the smaller ones.

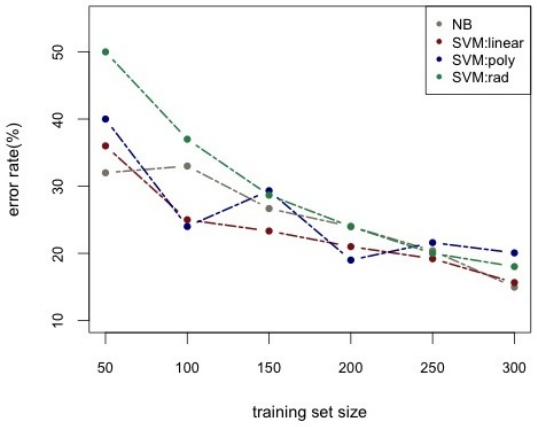

Fig. 3. Naive Bayes vs. SVM: 10-Fold CV

We then conduct 10-fold cross validation with both Naive Bayes and SVM with different kernels. (Naive Bayes in *Figure 2* indicates the one with $\alpha = 10$.) It turns out that SVM with linear kernel outperforms all other methods in most cases, although Naive Bayes with the biggest smoothing factor perform the best on the full set. We can infer that the decision boundaries supported by linear SVMs best suits our data, although differences are within 5 percentage point in all cases. When we test various cost values *C* for the objective function to make it work for non-linearly separable datasets, we get same error rates which shows that our data is not sensitive to the choice of the regularization parameter. This indicates that our data is reasonably well-behaved and does not

contain a disproportionate number of outliers.

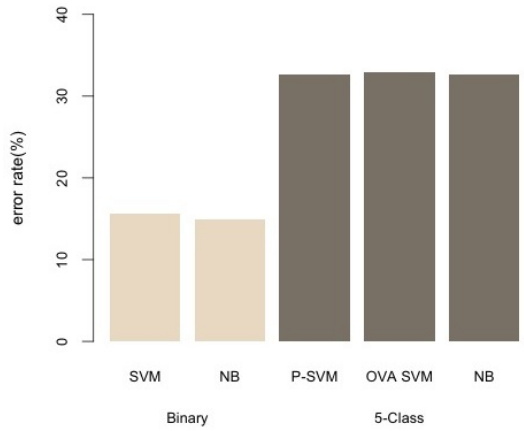

Fig. 4. Various Methods

We return to the 5-category classification problem. As there is no solid SVM method that works best for all multi-class problems, we conduct both one-versus-all (OVA) SVM and pairwise SVM (indicated as P-SVM in *Figure 4*). The results tell us that there is no significanct difference between the two SVM methods and they perform similarly to the Naive Bayes classifier.

VI. Conclusion

We use the machine learning techniques, specifically Naive Bayes and SVM, to predict the content of corporate 8-K filings. When introduce more than two categories, both methods yield error rates as high as one-third. For the purpose of this project, we reduce the number of category to two, and both techniques perform better with error rates around 15-20 percent. Naive Bayes perform slightly better in binary classification and the two methods perform very similarly in the multi-category problem. What we would be eventually interested in is multi-category classification and we are positive that once we significantly increase the size of the training set, that is, manually classify more filings, then more information would be fed into the our models and we will be able to predict the filing contents more accurately.

References

[1] Li, F. (2010). The information content of forward-looking statements in corporate filings: a Naive Bayesian machine learning approach. Journal of Accounting Research, 48. 1049-1102.

[2] Janes, G., et al. (2013). An introduction to statistical learning. Springer, New York.