Computerized Retrieval and Classification: An application to Reasons for Late Filings with the Securities and Exchange Commission

Ronen Feldman *†‖ Benjamin Rosenfeld † Ron Lazar‡ Joshua Livnat§
Benjamin Segal¶

∗ Department of Computer Science, Bar Ilan University, Ramat Gan 52900, Israel
† ClearForest Ltd., 6 Yoni Netanyahu St., Or Yehuda 60376, Israel
‡ University of Houston, 334 Melcher Hall Houston, TX 77204-6021
§ Stern School of Business Administration,New York University, 311 Tisch Hall, 40 W. 4th St. NY NY 10012
¶ Graduate School of Management, University of California at Davis, One Shields Avenue, AOB IV Room 134, Davis, CA 95616-8609

Abstract

This study explores a system to retrieve and classify the reasons for late mandatory SEC (Securities and Exchange Commission) filings. From the source documents, the system identifies the reasons for the late filing and classifies them into one or more of seven categories. The system can be used by potential investors who have to track a large number of filings concentrated within a day or two.

Our results indicate that the SEC filings may be quite ambiguous, with experienced raters disagreeing on one category for a training sample of 600 filings in about 30% of the cases. However, allowing classifications into more than one category using document level information yields accuracy of about 90% in a test sample of 200 filings. We also show that the stock market reactions to over 9,000 late filings vary in an intuitive way according to the classified reasons.

Keywords: Computerized Text Classification, Computerized Categorization, Late Filings, Accuracy of Categorization Algorithms

1 Introduction

As part of its overseeing of capital markets, the Securities and Exchange Commission (SEC) requires firms with publicly traded shares to issue periodic reports to shareholders. These reports include Form 10-Q, quarterly filings that typically contain condensed financial statement information and Management Discussion and Analysis (MDA), and Form 10-K, annual filings that contain detailed business description, a full set of audited financial statements and detailed MDA. Until recently, the filing deadlines were 45 days for Form 10-Q and 90 days for Form 10-K, but beginning with fiscal years ending on or after December 15, 2003, many companies are subject to a three year phased-in accelerated filing schedule of 35 and 60 days respectively[1].These filings, as well as other SEC filings, are part of the SEC's Electronic Data Gathering, Analysis, and Retrieval system (EDGAR)[2], a large online database

‖**Corresponding author:** feldman@cs.biu.ac.il

[1] http://www.sec.gov/rules/final/33-8128.htm

[2] ftp://ftp.sec.gov/edgar/

maintained by the SEC which includes all SEC filings. The SEC began collecting the data in 1993 with a restricted sample of firms, and made it mandatory for all firms in 1996. Empirical evidence shows that the vast majority of firms file their quarterly and annual reports with the SEC on the last day or two of the required filing period [4].

When a company realizes that it will not be able to file by the SEC deadline, it must file another form, Form 12b-25, Notification of Late Filing, within one business day of the required filing date, in which it needs to state the reasons why Form 10-Q or Form 10-K cannot be filed within the prescribed time period[3]. The reasons for the late filing are supposed to appear under the caption: "PART III — NARRATIVE", but some companies disclose additional information at the end of Form 12b-25 in an appendix. Form 12b-25 is also gathered by the SEC and included in the EDGAR database with the marking of NT (non-timely) 10Q or 10K. Firms then have a period of up to 15 calendar days in the case of 10K and up to 5 days in the case of 10Q to file their final SEC forms. Every filing in EDGAR has an individual path or URL, and EDGAR also contains index files by quarter and year that include these paths. Searching the index files, we found URL's for 17,170 NT-10K and 25,066 NT-10Q, of which we were able to download 17,139 NT-10K and 25,043 NT-10Q individual SEC filings.

Firms file NT forms for various reasons ranging from inability to get all the necessary information on time to major transactions that consume too much managerial time. Some of these reasons may clearly be perceived as negative, such as a looming bankruptcy, whereas others may be perceived as positive, such as a newly acquired subsidiary that needs to be integrated into the firm. A careful investor is likely to be interested in finding the reason for an NT filing immediately after filing. Unfortunately, due to the concentration of fiscal year-ends (about 60% of the firms have December fiscal year-end), and due to firms' tendency to file on the last day of the SEC filing period, there are many filings that take place all at the same time [4]. Hence, investors may find it very difficult to be able to capture information from both regular filings and also to concentrate on NT filings. A computerized tool that will read the NT filings, access the reasons for the filings, classify the reasons and report them to investors as a headline can be very attractive to potential investors and other users of financial statements.

The ability to mechanically access the information in the SEC EDGAR database, and to identify the text of the reason is fairly simple, given the structure of Form 12b-25, where the narrative containing the reason appears in Section III. However, the classification of accessed reasons is more difficult. In many cases, firms may use similar language to describe different circumstances, and careful analysis is needed for accurate classification. In other cases, firms may include more than one reason in their NT forms. Thus, advanced text categorization techniques must be used for narrative classification.

To decide on an initial method of classification, we obtained a sample of 300 NT 10-Q forms and another 300 NT 10-K forms. These forms were separately and independently classified by two individuals with extensive investment experience according to the classification proposed in [1] for NT 10-K. An immediately apparent problem of these classifications was the low levels of consistency among the two individuals – fewer than 70% of the cases were classified consistently by the two raters. Consequently, some categories were combined and two new categories were created, resulting in a list of seven fairly solid and unambiguous categories. The manually classified forms were subsequently used as a training data for a Machine Learning text classification algorithm, based upon logistic regression. The trained system was then used to automatically classify the entire database of NT filings.

To test the classification, a sample of 200 new NT filings was again classified by an experienced investor, who indicated whether the reason suggested by the classification tool is in agreement with the rater's assessment or not. The results of this test indicated that in 93% of the cases we managed to classify the documents into the right category. We also tested the classification by examining stock returns around NT filings. We find that the largest average (negative) returns are for the bankruptcy and financial condition reasons, consistent with intuition. We find that the lowest (negative) average returns occur for the auditor changes, likely because those are typically announced previously and separately in other SEC filings. We also find a relatively low average (negative) return for the largest

[3] http://www.sec.gov/about/forms/form12b-25.pdf

category, the delayed information category, which may be due to the heterogeneity of cases in this category, and for the restructuring and reorganization category which may include cases with favorable news. Thus, we feel that our results provide the necessary evidence in support of the computerized classification scheme.

2 Prior Classification and Examples of Reasons

In a pre-EDGAR period study ([1]) examine 182 12b-25 forms and classify the reasons provided by firms for their inability to file 10-K forms on time[4].

Their grouping comprised of four categories with nineteen subcategories. The first category, financial distress, included debt negotiations/restructuring, bankruptcy/reorganization and poor financial condition. The second category, accounting and auditing issues, included accounting issue/problems, delay in obtaining information, information needed from 3rd party, audit-related delay and investigating numbers. The third category, asset acquisitions and dispositions, included dispositions of businesses/assets, acquisitions of businesses/assets and business combinations/liquidation. The fourth and last category, other, included changes in top executives, printing delays, litigation/regulation-related, review/signature of officers/directors, need registration statement approval, labor/employee-related, staff reduction and miscellaneous.

Attempting to use this classification scheme showed us that too many cases can be classified into more than one category, and many others cannot be unambiguously classified into just one category. Consequently, we have formed seven broad categories that are sufficiently distinct from each other to allow classifications that are less subjective.

The following are examples of phrases and wordings used by companies in their NT filings based on each of these seven categories:

1. **Audit Related Delay**
   - *"Due to unforeseen delays encountered in completing the Company's audited Financial Statements. . . . . ."*
   - *"The registrant is completing the audit of its subsidiary in the United Kingdom and has not yet received the audit report from the auditors in the United Kingdom."*
   - *"The audited financial statements for the Company have not yet been finalized."*
   - *"Additional time is needed for the Certified Public Accountants'. . . ."*
   - *"The Company's auditors are awaiting final evidential matter in order to complete the audit."*
   - *"The SEC initially expressed the view that reversal of the accrual would require EchoStar to restate its results for 2001, which would have required a re-audit of those financial statements."*
2. **Auditor Change**
   - *"For reasons unrelated to the Registrant, CPAs elected to cease SEC client audits for inclusion in Form 10-K. The Registrant's management was required to replace them as auditors and the additional time is required to permit the new auditors to complete their audit."*
   - *"The Registrant has yet to retain a new auditor, and as a result, will be unable to complete its Annual Report."*
   - *"As a result of the recent change in auditors, it is not possible to complete the audit by . . . "*

[4] In a more recent study, Griffin [4] identifies out of a comprehensive sample between the years 1996-2001, 742 NT 10-K and 680 Form NT 10-Q filings in the EDGAR database.

3. **Management change**

   - *"The delays are the result of the termination of all of the Registrant's personnel and former operators, as a result of which successor management experienced material difficulties in obtaining required bank records and other materials."*
   - *"Because of personnel changes.."*
   - *"As a result of the management turnover.."*
   - *"... departure of the Company's former chief financial officer."*
   - *"Following the merger, the Registrant began to assemble a new management team"*
   - *"... because the Registrant's Chief Financial Officer resigned"*
   - *"... temporary absence of necessary personnel."*

4. **Restructuring and Reorganization (not bankruptcy)**

   - *"The Registrant completed a significant merger transaction on ....."*
   - *"Additional time is needed to properly disclose this acquisition and the subsequent two acquisitions that occurred in January."*
   - *"Due to a recent reorganization of the Company whereby four business segments were transferred to the Registrant's parent company, the Company was unable to complete its financial statement information in time"*
   - *"..delayed pending the outcome of negotiations which have resulted in the execution and delivery by a subsidiary of the Registrant of a merger agreement which, if consummated, will result in the Registrant's acquisition of a substantial interest in ..."*
   - *".. because the Company has experienced some difficulty in accounting for the issuance of stock as a result of a merger that occurred at the end of the Company's second quarter."*
   - *"..because the Company is experiencing delays in preparing financial statements that reflect the transition of the Company from a closed-end investment company to a holding company whose primary asset is the investment management business of ..."*

5. **Financial Condition**

   - *"Due to the uncertainty surrounding the financial covenants, the Company has not been able to complete the annual financial statements and Management's Discussion and Analysis."*
   - *"The Company is currently in negotiations with its lenders for modifications to its existing credit facilities."*
   - *"The delay in filing is principally attributable to the Company's current efforts to restructure its debt agreements and/or obtain waivers from its senior subordinated lenders with respect to certain financial covenants."*
   - *"The Company's management has been devoting substantially all of its time and attention to the resolution of certain financial matters, which resolution may affect the Company's disclosure."*
   - *"The Registrant has been involved in capital-raising activities. These activities have taken time and resources that ordinarily would have been used to prepare the Registrant's Annual Report."*
   - *"..due in part to the ongoing discussions between the registrant and its bank."*

- *".. because documents relating to amendment of the registrant's financing agreements, which impact the Form 10-Q.."*
- *"The consummation in August 1995 of new financing transactions between..."*
- *"...the Company entered into an agreement with the holders of its Convertible Preferred Stock..."*

6. **Delayed Information**

- *"Company respectfully requests additional time to gather all necessary information..."*
- *"Registrant is still reconciling the financial activity for the period so the filing will be accurate."*
- *"The Company could not obtain all the required information.."*
- *"Financial Statements still in the process of completion."*
- *"..because the Registrant is unable to do so without unreasonable effort or expense."*
- *"Additional financial information necessary for filing the Financial Statements is not yet available."*

7. **Bankruptcy Related Delay (including reorganization)**

- *"While the company was in bankruptcy for the most of the prescribed period.."*
- *"The registrant and certain of its subsidiaries (the "Debtors") filed a petition (the "Chapter 11 Petition") for relief under Chapter 11 of the U.S."*
- *"On ..... (the "Registrant") announced that, due to severe cash flow problems and lack of liquidity, its French subsidiary, Sames, S.A., had determined to file for bankruptcy protection under French law."*
- *"On ..., the Registrant's stockholders approved a Plan of Complete Liquidation and Dissolution. On ..., pursuant to two separate transactions, the Registrant sold substantially all of its operating assets."*

To experiment with our classification tools, we used a randomly selected training sample composed of 300 10-Q forms and 300 10-K forms, which were separately and independently classified by two individuals with extensive investment experience. The initial classification scheme, which was close to that of [1], showed a severe lack of internal consistency, yielding the same classification among the two individuals of less than 70%. As a result, we have decided to limit our analysis to the above seven categories. The classification consistency of the training sample using the seven categories was a much better at about 70%, as we report below.

The population for this study includes all Form 12b-25 filings that are available and downloadable on the SEC EDGAR database through the end of 2003. These include 17,139 NT-10K and 25,043 NT-10Q SEC filings. For each of these filings, we extract the filing date, the report period, the company's SEC identifier (CIK number), and the text that contains the reasons for the late filing. The text is used to classify each filing into one of the above seven categories. Since a firm may include more than one reason for a late filing, such as a major transaction and an auditor change, we allow the classification tool to propose more than one category with probabilities attached to each suggested category.

3 Classification methods

For classifying the narratives, we used general machine learning text categorization techniques [9], [8],[11]. The general text categorization task is to classify a collection of documents into a set of predefined categories, which translates naturally to our case. The machine learning approaches to text categorization require a training set of

manually classified documents. Using the training set, a learner then proceeds to build a *classifier*, which can be used to categorize a previously unseen document.

In this section, we shall first discuss the document representation issues, then describe our learning and classification algorithm, and finally describe our experimental setup and the results.

3.1 Document representation

The general machine learning classification algorithms usually represent the data instances as sets of weighted *features*. Thus, each instance is a vector in a high-dimensional vector space, of which the features form the basis. The nature of features is problem-dependent, but for the purposes of the classification algorithms, the features are structureless atomic entities.

The most common representation model for text categorization is the *bag-of-words* model, in which the features are simply words. A document is represented by a vector that has nonzero components for those words that appear in the document. The weight of each appearing word may be simply 1 – this is called *binary* representation – or it may depend upon the frequency of appearance of the word in the document and in other documents of the collection.

We found experimentally that the classification performance in our domain can be significantly improved by using not only single words but also pairs of adjacent words as distinct features. We also tried even bigger terms as features, such as 3-word sequences, and sequences with gaps, but it did not lead to further improvement.

As usual in text classification, the number of features is very large, and most of those features are irrelevant for classification. Therefore, *feature selection* is commonly applied, reducing the dimension of the feature space by a factor of ten to hundred. During feature selection, a relevancy measure is calculated for each feature, after which the N top-scoring features are retained, and all others are dropped. The precise number N of remaining features is not very significant, as long as it is sufficiently large to include the best features. In our experiments, we used $N = 1000$. We tested several common feature relevancy measures, and found that the Bi-Normal Separation (BNS) performed the best, supporting the findings of [2].

The BNS measure $BNS(w, c)$ of a feature w for a category c is defined as follows:

$$BNS(w,c) = \left| F^{-1}\left(\frac{|\{d \in C : w \in d\}|}{|\{d \in C\}|}\right) - F^{-1}\left(\frac{|\{d \in \bar{C} : w \in d\}|}{|\{d \in \bar{C}\}|}\right)\right|,$$

where C is the set of training documents belonging to the category c, $\bar{C}$ is the set of training documents not belonging to c, and F^{-1} is the standard Normal distribution's inverse cumulative probability function (z-score).

3.2 Classification algorithm

In our experiments, we tested the SVM_{light} [5] implementation of the SVM classifier [10],[6],[7] and BBR [3] implementation of the Bayesian Logistic Regression. We found that the BBR performed much better in our case, although the precise reason is not clear. It is possible simply that the particular implementation of SVM was less suited for the particular problem.

A *logistic regression classifier* models the conditional probability $P(c \mid \mathbf{d})$ of a document $\mathbf{d}$ to belong to the category c. The logistic regression model has the form

(1) $P(c = +1|\mathbf{d}) = \varphi(\beta \cdot \mathbf{d}) = \varphi(\sum_i \beta_i d_i)$,

where $c = \pm 1$ is the category membership value for a document (± 1 is used instead of usual {0,1} for simpler notation), $\mathbf{d} = (w_1, w_2, \dots)$ is the document representation in the feature space, $\beta = (\beta_1, \beta_2, \dots)$ is the model parameters vector, and φ is the *logistic link* function

$$\varphi(x) = \frac{\exp(x)}{1 + \exp(x)} = \frac{1}{1 + \exp(-x)}.$$

The model (1) can be trained by maximizing the likelihood of the training set. However this would result in a severe overfitting problem. The Bayesian approach to the problem is to choose such prior distribution for the parameters that would assign a high probability to each β_i being at or near zero. Then, the Maximum a posteriori (MAP) estimates are used instead of the Maximum Likelihood (ML) estimates.

Given the prior $p(\beta)$ and the training data $D = \{(d_1, c_1), (d_2, c_2), \dots\}$, the log-posterior distribution $l(\beta)$ of the parameter vector is

$$l(\beta) = P(\beta|D) = -\left(\sum_{(\mathbf{d},c)\in D} \ln(\exp(-c\,\beta\cdot\mathbf{d}) + 1)\right) + \ln p(\beta).$$

The common choices of prior are the Gaussian and Laplace:

$$\ln p(\beta) = -\left(\sum_i (\ln\sqrt{\tau} + \tfrac{\ln 2\pi}{2} + \tfrac{\beta_i^2}{\tau})\right), \text{ for Gaussian prior, and}$$
$$\ln p(\beta) = -\left(\sum_i (\ln 2 - \ln\lambda + \lambda\,|\beta_i|)\right), \text{ for Laplace prior.}$$

The MAP estimate $\arg\max_\beta l(\beta)$ can be calculated by any convex optimization algorithm.

4 Experimental Setup

The architecture of the system has three basic parts: the preprocessor, the categorization system, and the system to evaluate results. The preprocessor converts the original forms into the feature vectors suitable for processing by the text categorization algorithms:

[Form] $\diamondsuit$ Narrative Extractor $\diamondsuit$ Feature Extractor $\diamondsuit$ [Feature Vector]

First, the forms are passed through a manually-written Perl routine, which extracts the "Narrative" part of the form. Appendix sheets are also included if referenced in the narrative. The format of the forms is more or less uniform, so our simple extractor is at least 95% accurate. The output of the narrative extractor is a narrative text, which is passed through the feature extractor. Feature extraction and selection process is as explained above.

In order to train the classifier, we manually labeled a set of 600 narratives. In the first series of experiments, the text categorization algorithm (BBR, described above) was tested in a 10-fold cross-validation over this set of narratives. The apparent results were rather mediocre – about 71% microaveraged F-measure. However, we also found that cross-annotator consistency for the domain was very low – around 75-80%. The reason for this is that many narratives use vague terms that can be interpreted in different ways. Also, some of the narratives list several reasons for the delay.

Next, we tried annotating and classifying separate sentences from the narratives, hypothesizing that separate sentences would be easier to classify precisely. However, this turned out not to be the case – most of the errors remained. Also, the feature vectors of the sentences become too sparse, which presented a problem for the classifier.

Therefore, in our final experiment we turned back to classifying whole narratives, but changed the evaluation procedure. First, we allowed a narrative to belong to more than one category. The most probable category as given by the BBR classifier was listed as the principal category of the narrative. Other categories having probabilities within a factor of three of the principal category's probability were listed as additional categories. The system was trained upon all 600 narratives, and tested upon additional 200 narratives. The classification results of those were checked manually by an expert, who labeled the results as "right", "wrong", and "can be considered right". Evaluated in this way, the system showed a much better 93% F-measure.

4.1 Classification Results

4.1.1 Inter-Annotator Agreement:

Table 1 shows the extent of agreement between the two annotators of the training sample of 600 late filings. The table contains the Inter-Annotator classification agreement of the training sample composed of 300 late 10-Q and 300 late 10-K forms. Recall is the proportion of consistently classified documents in a given category by both annotators out of all possible documents in that category by one of the annotators. Precision is the proportion of consistently classified documents in a given category by both annotators out of all documents classified into that category by one of the annotators.

Note that the overall level of agreement is rather low, with only about 75% agreement between the two annotators. The main reason for this low agreement is the vagueness of the language used to explain the reason for the late filing, and the inclusion of several reasons in one late filing, which allowed an assignment of the reason into multiple categories. A further examination of the recall and precision by category shows that the financial condition category has consistently higher recall and precision than other categories, but also that there are categories with high recall (precision) but low (high) precision (recall). Thus, it may make sense to allow the classification algorithm to use more than just one reason.

Table 1: Inter-Annotator Agreement

Reason	**Recall**	**Precision**
Auditor Change	75.00	63.16
Audit- Related Delay	93.44	47.11
Bankruptcy-related Delay	78.95	57.69
Delayed Information	75.60	87.89
Financial Condition	82.22	80.43
Management Change	73.91	75.56
Reorganization and restructuring (not bankruptcy related)	55.26	79.25
All	75.29	75.29

4.1.2 Sentence-Level Model:

Table 2 shows the classification results when individual sentences are used to train the classification model. The table contains the classification results of the training sample composed of 300 late 10-Q and 300 late 10-K forms. The classification is based on using sentence level annotations to classify the reasons for the late filing. The table is based on 10-fold cross-validation; we randomly use 540 forms to estimate model parameters and classify the remaining 60 forms according to the estimated parameters. This process is repeated 10 times, with averages over the 10 experiments reported in the table. Recall is the proportion of correctly classified documents in a given category out of all possible documents in that category. Precision is the proportion of correctly classified documents in a given category out of all documents classified into that category.

As can be seen in the table, the overall accuracy of the sentence-level model is pretty low at about 63%. A possible explanation for this low level is that many documents may contain sentences that can be classified into more than one category, leading to low accuracy. In particular, note the low accuracy of the auditor change and audit-related delays, possibly because these two categories are often more difficult to interpret and assign correctly at the sentence level.

Table 2: Accuracy results For the Sentence Level Model

Reason	Recall	Precision
Auditor Change	0.00	0.00
Audit- Related Delay	38.10	37.21
Bankruptcy-related Delay	23.08	100.00
Delayed Information	84.65	69.80
Financial Condition	20.69	54.55
Management Change	63.64	62.50
Reorganization and restructuring (not bankruptcy related)	53.85	56.00
All	63.48	63.48

4.1.3 Document-Level Model:

Table 3 shows the classification results based on a document-level model. The table contains the classification results of the training sample composed of 300 late 10-Q and 300 late 10-K forms. The classification is based on using the entire document to classify the reasons for the late filing. The table is based on 10-fold cross-validation; we randomly use 540 forms to estimate model parameters and classify the remaining 60 forms according to the estimated parameters. This process is repeated 10 times, with averages over the 10 experiments reported in the table. Recall is the proportion of correctly classified documents in a given category out of all possible documents in that category. Precision is the proportion of correctly classified documents in a given category out of all documents classified into that category.

These results are more reasonable with an overall accuracy of 71.79 % for the entire training sample using a 10-fold cross-validation. Note that this accuracy is not much below the overall agreement level among the two annotators of 75%. Still, many categories show low recall or precision levels, indicating that attempting to assign a typical case into just one category is likely to result in low accuracy levels.

Table 3: Accuracy results For the Document Level Model

Reason	Recall	Precision
Auditor Change	31.57	75.00
Audit- Related Delay	63.63	74.03
Bankruptcy-related Delay	61.53	84.21
Delayed Information	93.08	70.79
Financial Condition	52.17	63.15
Management Change	48.89	73.33
Reorganization and restructuring (not bankruptcy related)	30.18	80.00
All	71.79	71.79

4.2 Classification Results based on Test Sample:

Table 4 contains the results of using the classification estimated from the training sample of 600 forms on the 200 fresh forms. The table contains the classification results of the test sample composed of 200 late filings. The classification

is based on using the entire document to classify the reasons for the late filing, with possible multiple categories for each filing. Recall is the proportion of correctly classified documents in a given category out of all possible documents in that category. Precision is the proportion of correctly classified documents in a given category out of all documents classified into that category.

In the test sample, raters assessed whether the classified category was correct or not, and supplied the correct category for each of the 200 forms. To increase the model's accuracy, a form may be assigned into more than one category, with a probability attached to each category. We assigned each form into the highest-probability category (let that probability be P) and in addition to all the categories that have a probability not less than P/K (K=3 in our experiments). As can be seen in the table, the overall accuracy of the model is an impressive 90%, showing that the classification model does well in showing the relevant categories for a particular form. Untabulated results show that the precision and accuracy was about 95% in those cases where at least one category was correct. Thus, the model tends to work fairly well when it is allowed to offer more than one category with the associated probability for each category.

Table 4: Classification Accuracy – Test Sample

Reason	**Recall**	**Precision**
Auditor Change	100.00	100.00
Audit- Related Delay	90.91	100.00
Bankruptcy-related Delay	83.33	100.00
Delayed Information	100.00	75.30
Financial Condition	87.50	95.45
Management Change	94.44	94.44
Reorganization and restructuring (not bankruptcy related)	96.00	75.00
All	89.65	89.65

4.3 Stock Market Reactions for Various Categories:

To have an independent assessment of our classification, we also examine the stock market reactions to announcements of late filings. Presumably, the more severe is the reason for the late filing, such as in the case of bankruptcy or financial condition, the more likely is the associated stock market return around the announcement to be large and negative. However, market participants may have had the information previously from company prior announcements and press releases. In these cases, the market reaction to the Form 12b-25 filings may be less severe. Thus, the market return analysis should be taken just as an indication of the classification accuracy.

For each late filing, we match the CIK with a unique identifier used by Compustat, GVKEY, using data provided to us by Compustat which maps CIK into GVKEY. We then used the merged CRSP/Compustat database in WRDS (Wharton Research Data System) to identify the permanent CRSP number assigned to each company's securities. For each SEC filing, we calculate the cumulative daily stock return during the period [-3,+3], where day 0 is the SEC Form 12b-25filing date. To focus on abnormal returns, we subtract the return on the CRSP Index of all available equity securities using the same period as the individual returns. Since the announcement of a late filing can typically be thought of as negative news, because the company admits its failure to file on time, we expect a negative abnormal market return around the announcement. However, depending on the particular reason, the negative reaction may be more or less severe.

Table 5 contains all late filings for which we could obtain abnormal returns surrounding the SEC filing. Abnormal

returns are the cumulative return on the individual company stock during days [-3,+3], where day 0 is the SEC Form 12b-25filing date, minus the cumulative return on the CRSP value-weighted index of all stocks during the same period. The table shows the mean abnormal return within a category, as well as the statistical significance of the mean; i.e. whether the mean is statistically different from zero using a t-statistic.

As can be seen in Table 5, the mean abnormal return associated with a late filing is of -1.4%, statistically different from zero at a significance level below 0.00001 for the 9,125 cases where we could obtain abnormal returns for the filing companies. As can be intuitively expected, the most negative (severe) market reactions occur for bankruptcy-related and financial-condition-related delays, although the former is not statistically different from zero using conventional levels, possibly due to the small number of cases (only 42) in the sample. Auditor changes and reorganizations and restructurings that are unrelated to bankruptcy have the lowest (least severe) negative abnormal returns, as can be expected; auditor changes are typically announced previously through SEC form 8-K, which must be filed within four business days after the event[5], and reorganizations and restructurings include many transactions that may be perceived as good news about the firm due to its expansion. Management changes and audit-related delays have negative abnormal returns that are in the middle, likely because they signal problems, but not as sever as those related to bankruptcy or necessary financing. Thus, the mean abnormal returns for our categories seem to be in accordance with our prior expectations and intuition, providing another dimension of credibility to the classification results.

Table 5: Mean Abnormal Stock Returns in Various Categories (Sentence Level Model)

Reason	**Abnormal Return**	**N**	**Significance**
Auditor Change	-0.00023	115	0.99083
Audit- Related Delay	-0.02002	251	0.10432
Bankruptcy-related Delay	-0.04427	42	0.20544
Delayed Information	-0.01098	6717	$< .00001$
Financial Condition	-0.03413	514	0.00001
Management Change	-0.02745	381	0.00113
Reorganization and restructuring (not bankruptcy related)	-0.01654	1105	0.00087
All	-0.01391	9125	$< .00001$

5 Discussion

The purpose of this study was to examine various ways to automate the categorization of Late Filing Notification forms filed with the Securities and Exchange Commission. Since the relevant information in the forms appears as free text, the problem is not trivial. Also, the categories are ambiguous and sometimes may overlap, which presents additional problems to both human and machine classifiers. The study shows that using a careful methodology, automatic (machine learning-based) classifiers are able to achieve accuracy close to human annotators' performance.

There are many ways to translate the "real-world" problem of forms categorization into a problem of vector categorization, which can be solved by a Machine Learning (ML) classification algorithm. We performed an extensive

[5] Five days prior to August 23, 2004.

series of experiments, finding the best way. It turns out that the best performance is achieved by whole document classification, using "bag-of-pairs-of-words" document representation, with the BNS feature filtering.

Due to the ambiguousness of category assignment of forms, evaluations of the categorization results are not straightforward. This is clearly demonstrated by a significant 25% disagreement between different human annotators. In this evaluation, the automatic categorization system produces 30% disagreement, only slightly worse than human results. However, 70% accuracy is not a true assessment of system performance, just as 75% is not a correct figure of human performance. In order to get a better evaluation, we performed another experiment, in which the human annotator, instead of specifying correct categories for all forms, was asked to specify whether the ML system's decisions were correct. Evaluated in this way, the system produced an impressive 90% accuracy (95% if multiple categories per documents were allowed). We checked, as an additional test of the system and perhaps the most relevant to the practical applications, whether the extracted categories of the forms filed by companies could be used as predictors for the behavior of companies' stock on the market. It turns out that indeed the association between the various categories and stock returns corresponds to logic and intuition.

6 Further work

The application provided in this study is just one example of computerized categorization and classification in the finance sector. Many other applications with various degrees of difficulty and complexity exist in the financial sector because of the reporting requirements prevalent in it, and the systematic collection of reports and press releases. For example, we are now in the process of automating the retrieval and classification of the reasons for changes in sales (revenues) as discussed by management in SEC filings (a required disclosure). This is a much more complex problem because firms vary widely not only in the reasons they provide for the changes in sales, but also in their presentations, and do not maintain consistency of reporting from one period to another. Another example that we began addressing is the association of text in the SEC filings with various financial statement items, such as discussion of inventory linked to the inventory amount on the balance sheet. The firm may provide references to inventory in many places throughout the filing, such as the accounting policy adopted by the firm with respect to inventories, inventory writedowns, LIFO reserve, restructuring charges related to inventory, reasons for inventory increases or declines, etc. These are just two examples of projects that may use computerized retrieval and classifications of text in the financial area.

7 Conclusions

The purpose of this study is to suggest a methodology to develop a computerized system that will retrieve and classify the stated reasons for delayed SEC filings provided by firms. The system can be used by investors, who need to scan many such filings that become available on the same day, to guide them through portfolio investment decisions. Our study indicates that the machine learning text categorization systems can successfully solve the problem and that the categorization results can be used by investors as the stock market reactions associated with each of the categories differ in magnitude.

References

[1] Jones J. J. Alford, A. W. and M. E. Zmijewski. Extensions and violations of the statutory sec form 10-k filing requirements. *Journal of Accounting and Economics*, 17:229–254, 1994.

[2] George Forman. An extensive empirical study of feature selection metrics for text classification. *Journal of Machine Learning Research*, 2003.

[3] Alexander Genkin, David D. Lewis, and David Madigan. Large-scale bayesian logistic regression for text categorization. Technical report, DIMACS, 2004.

[4] P. A. Griffin. Got information? investor response to form 10-k and form 10-q edgar filings. *Review of Accounting Studies*, 8:433–460, 2003.

[5] Thorsten Joachims. Estimating the generalization performance of a svm efficiently. In Pat Langley, editor, *Proceedings of ICML-00, 17th International Conference on Machine Learning*, pages 431–438. Morgan Kaufmann Publishers, San Francisco, US, Stanford, US, 2000.

[6] Thorsten Joachims. A statistical learning model of text classification with support vector machines. In W. Bruce Croft Zobel, David J. Harper, Donald H. Kraft, and Justin, editors, *Proceedings of SIGIR-01, 24th ACM International Conference on Research and Development in Information Retrieval*, pages 128–136. ACM Press, New York, US, New Orleans, US, 2001.

[7] Thorsten Joachims. *Learning to Classify Text using Support Vector Machines.* Kluwer Academic Publishers, Dordrecht, NL, 2002.

[8] David D. Lewis. Machine learning for text categorization: background and characteristics. In Martha E. Williams, editor, *Proceedings of the 21st Annual National Online Meeting*, pages 221–226. Information Today, Medford, USA, New York, US, 2000.

[9] Fabrizio Sebastiani. Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1):1–47, 2002.

[10] V. Vapnik. *The Nature of Statistical Learning Theory.* Springer-Verlag, 1995.

[11] Yiming Yang. An evaluation of statistical approaches to text categorization. *Information Retrieval*, 1(1/2):69–90, 1999.