When Positive Sentiment Is Not So Positive: Textual Analytics and Bank Failures

Aparna Gupta[1], Majeed Simaan[1], and Mohammed J. Zaki[2]

[1]Lally School of Management at Rensselaer Polytechnic Institute
[2]Department of Computer Science at Rensselaer Polytechnic Institute

29th April 2016

Abstract

We extend beyond healthiness assessment of banks using quantitative financial data by applying textual sentiment analysis. Looking at 10-K annual reports for a large sample of banks in the 2000-2014 period, 52 public bank holding companies that were associated with bank failures during the global financial crisis serve as a natural experiment. Utilizing negative and positive dictionaries proposed by Loughran and McDonald (2011), we find that both sentiments on average discriminate between failed and non-failed banks 80% of the time. However, we find that positive sentiment contains stronger predictive power than negative sentiment; out of ten failed banks, on average positive sentiment can identify seven true events, whereas negative sentiment identifies five failed banks at most. While one would link financial soundness with more positive sentiment, it appears that failed banks exhausted more positive sentiment than their non-failed peers, whether ex-ante in anticipation of good news or ex-post to conceal financial distress.

Electronic copy available at: http://ssrn.com/abstract=2773939

1 Introduction

Given the substantial increase in publicly available textual data along with the innovation in textual tools to analyse such unstructured information, it is an open question to what extent financial textual sentiment can play a role in predicting bank failures. To answer this question, we bridge healthiness assessment of banks using quantitative financial data with textual sentiment analysis by looking at 10-K annual reports for a large sample of banks in the 2000-2014 time period. The 52 public bank holding companies that were associated with bank failures during the global financial crisis serve as a natural experiment. Utilizing negative and positive dictionaries proposed by Loughran and McDonald [23], our findings establish a strong link between sentiment and financial soundness of banks.

Unlike previous financial crises that originated in capital markets (Long-Term Capital Management (LTCM) bailout and the dot.com bubble bust around 2000), the 2007-09 financial crisis started in the banking sector and spilled over to the broader economy. This has instigated a fresh debate about the riskiness and capitalization of banks and their ability to absorb negative shocks in economic downturns.[1] Since banks are highly leveraged and issuing equity can be costly, a sudden drop in a bank's asset value would require it to sell a large amount of its assets in order to maintain minimum capital ratios. This disproportional selling, as a result, could create a feedback loop and undermine the bank's solvency even further [4].

By regulation banks are required to maintain a minimum level of capital with respect to their risk-weighted assets. A drop below that level should be an indication of a bank's distress, and can threaten the bank's solvency. Recent research by Berger and Bouwman [12] highlights the importance of a bank's capitalization for its survival during normal or crisis periods. Therefore, a bank's financial indicators should play an important role in creating an early warning system for the bank's soundness. However, one of the main issues in the recent financial crisis is that banks were able to write off a lot of their activities from their balance sheets through securitization. This allowed banks to take excessive risk, while maintaining the same capital ratios for taking greater risk.[2]

In this paper, our focus is not on analyzing the level to which banks were able to conceal

[1] For instance, see [2].

[2] For a detailed overview on the 2007-09 financial crisis, see [1].

Electronic copy available at: http://ssrn.com/abstract=2773939

excessive risk taking leading up to the financial crisis, but rather, we study to what extent publicly disclosed textual information by banks can be used to predict financial distress. We do so by analysing the annual 10-K reports that public banks are required to file with the Securities And Exchange Commission (SEC), with respect to the sentiment dictionaries proposed by Loughran and McDonald [23] (henceforth 'LM'). According to the Federal Deposit Insurance Corporation (FDIC), there were 530 bank failures between 2000 and 2014, most of which (83%) took place between 2009 and 2012. While most of the failed banks were small and not publicly listed, our final universe of failed banks in this study consists of 52 publicly listed bank holding companies (henceforth 'BHCs').

Research on the prediction of corporate bankruptcy is extensive and dates back at least to the late 1960s. One of the famous measures to assess the healthiness of a company, for instance, is the Altman's Z-score [5]. Earlier empirical evidence documents that financial ratios as predictors of corporate failures can play the role of an early warning system, even up to 5 years prior to the actual failure [8, 9]. Later research has implemented artificial intelligence tools to predict corporate failure using financial data [10, 13].[3] For specifically banking, different lines of research also used diverse methodologies to predict bank failures. For example, [21] uses Cox proportional hazards model to predict bank failures, whereas [20] proposes a computer-based early warning system to predict U.S. large commercial bank's failures using logistic and trait recognition models. Moreover, [29] introduces neural networks approach to predict failures of Texas banks between 1985 and 1987.[4] On the other hand, [16] study the impact of equity on bank failures, and find that equity prices, returns, and volatility, all play an important role in identifying failed banks, in addition to the quarterly disclosed financial data. Nevertheless, none of the aforementioned look into unstructured data and study the predictive power of textual sentiment.[5]

Over the last decade more financial research has looked into financial textual data to better understand untapped information. To mention a few, [6, 22, 30, 18, 32] look into the impact of textual analysis on the equity market. [30], for instance, finds that high media pessimism predicts downward pressure on market prices and high market volume. Nonethe-

[3] For a recent review of common predictors used in the literature in predicting corporate bankruptcy, see [31].

[4] According to the FDIC, more than quarter of failed banks in 1987 were in Texas.

[5] For a recent exhaustive review on the literature of predicting financial distress and corporate failure see [28].

less, most textual analysis literature has focused on explaining stock market movements that are unexplained by fundamentals.[6] To the best of our knowledge, our paper is the first that tries to study the relationship between the textual content and bank failures. Our paper is closely related to [14], who look at the power of text in predicting catastrophic financial events related to fraud or company's bankruptcy. The authors analyse annual corporate disclosures (10-K reports) in which they look at the Management Discussion and Analysis Section (MD&A) and derive a dictionary to perform discriminant analysis. The authors report an average accuracy of 75% to discriminate fraudulent from non-fraudulent firms and 80% for bankruptcy, which is consistent with our findings. However, the degree to which public textual data contains valuable information about a bank's soundness remains an open question.

We attempt to bridge this gap in the literature by analyzing the power of textual sentiment in predicting bank failures. By looking at a large sample of textual data through the recent financial crisis and applying a bag of words approach, we extract sentiment-related features to perform discriminant analysis between failed and non-failed banks. Due to the statistical property of unigrams, our feature space consists of high dimensional data. For instance, we identify 833 negative and 145 positive terms that show up at least once across all reports. Further our complete panel dataset spans a comprehensive extraction of such features for a large number of banks for more than a decade. A common approach, as documented by LM, is to use the tf.idf weighting scheme to map the term frequencies into scores, and then equally weight all term scores within a document such that each report corresponds to a single sentiment score. When looking at the average sentiment of the system, we observe that both failed and non-failed banks expressed more negative sentiment as the financial crisis unraveled, where the failed banks expressed more negative sentiment on average. Nevertheless, while the system as a whole seems to be less positive as soon as the crisis began, the evidence from the failed banks does not indicate so. It appears that failed banks expressed more positive sentiment on average than their non-failed peers. It could be the case that failed banks tried to signal positive signs while in fact they were facing distress in order to maintain confidence among shareholders and investors.

For predicting bank failures, we utilize a similar weighting scheme as LM to give each term in the 10-K report a sentiment score. However, when looking at the document as a

[6] For a systematic review on text mining for market prediction see [26].

whole we do not equally weigh the term scores. If all terms in the report are assigned equal weights, one could neglect significant terms related to bank distress by allocating them less weight, while putting greater emphasis on terms that are of lesser significance. Such practice would result in a sub-optimal score assignment for the document, as it does not account for the state of the bank in the process. Instead of equally weighing the term scores in the document, we ascribe weights using a supervised learning model in which the term weights are assigned by utilizing maximum discriminative power between failed and non-failed banks. We serve this purpose by training Support Vector Machines (SVM) model on the term scores given the status of each bank. This, hence, results in a representative sentiment grade for each 10-K report in our sample that takes into account the bank's financial soundness. Finally, we use these optimized sentiment grades in a series of out-of-sample predictions. Depending on the conducted tests, we find that predictions based on negative and positive sentiment result in accuracy of $74\% - 94\%$ and $71\% - 83\%$, respectively.[7] However, accuracy by itself can be misleading, especially when the failed banks constitute a much smaller proportion of the sample as a whole. To control for this imbalance, we investigate the ability of our methodology to predict bank failures from actual failures. We find that positive sentiment contains stronger predictive power than negative sentiment. For instance, out of ten failed banks, positive sentiment on average can identify seven failed banks, whereas negative sentiment identifies five failed banks out of failed ones at most.

Our final results are summarized in a series of tests. In each experiment, we capture the time dynamics by focusing on 10-K reports filed during a window of time prior to bank failures taking place. We observe that as we approach the bulk of bank failures, the prediction power greatly increases as the sentiment extraction becomes more indicative of imminent failures. Moreover, while we use SVMs to find term optimal weights within each textual report, the large dimensionality of the sentiment dictionaries (especially negative) can undermine the optimal solution, even though SVMs have the capability to work with high dimensional data. We apply thinning on terms by keeping only terms with the most significant sentiment discrepancy between failed and non-failed banks. This reduces the feature space by almost 70%, and as a result the prediction power of the model further improves. Additionally, on closer inspection we find that from the non-failed banks sample, 118 banks were acquired in

[7] Accuracy is captured by the number of correctly predicted bank states divided by the total number of banks in the experiment.

the study period. We control for these acquisitions by considering two different samples. In one case, we update the non-failed sample by dropping all acquired banks and then compare the modified non-failed bank set with the failed set. In the second case, we add to the failed set a subset of the acquired bank who signaled significant financial distress prior to being acquired. For the former case, the model achieves its highest prediction power as discrimination analysis is conducted between purely failed and non-failed groups. In the latter case, however, while the acquired banks showed signs of distress with respect to their Tier 1 capital, augmenting them into failed group adds more noise than contribution to the model's prediction power.

Our contributions, therefore, are twofold. First, we establish a link between textual content, extracted using sentiment dictionaries, and bank financial distress, where we provide robust evidence is support of sentiment predicting bank failure. Second, we find that positive sentiment played a more significant role in predicting bank failures over the study period than negative sentiment. We attribute our contribution, especially the second one, to the usefulness of integrating statistical learning tools to assigning sentiment scores to the 10-K reports. Such score assigning integrates the information about the state of the bank, and hence, finds the term weights within the document that enhances the supervised learning process. Despite the criticism meted out to machine learning tools in the sense that they obscure the relationship between the predictors and the outcome, when looking at financial unstructured data, we conclude that average positive sentiment per se does not necessarily imply good financial soundness. Hence, without learned weight, such positive sentiment can be inconclusive, and even misleading.

The rest of the paper proceeds in the following order. In Section 2, we provide a detailed description of our sample construction and data collection process, which yields our final universe of banks for our study period. Section 3 describes the feature space extraction process, the model we implement for 10-K sentiment scoring, and the methodology used to perform text-based prediction of bank failures. Section 4 covers the findings of our papers in different test cases, while Section 5 concludes the paper.

2 Text Analytics for Bank Failure

To serve the objective of this study, we need a large corpus of appropriately chosen data from a large set of banks. The appropriateness of the data is judged by several aspects, most important of which is that the textual data describes the condition of banks for their risks, their ability to remain solvent and profitable, while meeting their obligations. These data need to span a substantial time period prior to the time of investigation. Additionally the data availability should be sufficiently consistent both in relevance and volume across the sample of banks being studied. With all these considerations, for this study we focus on SEC filings of banks in a time period prior to and including the global financial crisis period.

Once the corpus of text data is identified and created, extraction of chosen features is performed after the necessary cleaning steps for the text data. The features are utilized in a classification methodology to help detect weak banks that may be prone to failure. Several methodological challenges must be addressed in the process, discussion of which we delegate to Section 3. For the rest of this section, we address the challenges faced in the creation of an appropriate corpus of text data.

Our data construction relies on several different sources. The major data for our analysis come from unstructured textual information collected from the SEC EDGAR system on all banks in our study. We first describe how we identify the failed banks for the period of the study and create the universe of banks. Moreover, we detail the process for establishing a link between common structured data and the unstructured textual data to construct our final dataset upon which our empirical framework is applied.

2.1 The Universe of Banks

We identify failed banks using the FDIC publicly available data on failed commercial banks. The main challenge in constructing our universe of failed banks is to find a key link between the FDIC failed bank data and their identifiers in the SEC EDGAR system. The former set identifies commercial banks with respect to their FDIC unique certificate, whereas the latter refers to the bank holding companies using the central index key (CIK). Therefore, the task is to find the link between the FDIC certificate number and the CIK.

We start by considering all bank holding companies (BHCs) reporting the ‘FR Y-9C’ form beginning from 2000-Q1 till 2014-Q4. Using the Federal Reserve Bank of New York

PERMCO-RSSD dataset, we find the corresponding CRSP's permanent company identifier (PERMCO) for each BHC.[8] Then, we link the BHCs to the CRSP-COMPUSTAT merged dataset. This allows us to identify the CIK for each BHC in the sample. Over the sample period of 2000-2014, there are in total 809 BHCs with valid CIK numbers. On the other hand, in order to link the FDIC data to the BHCs sample, we merge the FDIC set with the commercial banks data available at the Federal Reserve Bank of Chicago. Each commercial bank has a corresponding FDIC certificate number (RSSD9050) and a higher holder identification number (RSSD9364). This eventually allows us to link the FDIC to the BHCs, and hence, to the SEC EDGAR system by finding the corresponding CIK for each company, including the failed ones. Figure 1 contains a flowchart demonstrating the link between the different data sources.

Since the FDIC data refer to commercial banks, we narrow the universe of BHCs down to companies with standard industry classification (SIC) code less than 6200.[9] This matching narrows down our BHC universe to 730 companies with unique CIKs (646 non-failed and 57 failed banks). We then remove all observations with missing values for total assets or negative equity. This leaves us with 701 firms, of which 55 are failed banks. Furthermore, from the non-failed banks set, in order to account for the bank size effect, we retain only non-failed banks whose size is not larger than that of the failed banks set. This creates a more relevant control group of non-failed banks and omits too-big-to-fail (TBTF) banks, which enjoy government safety net on the verge of failure. This drops the number of non-failed banks to 593, leaving us with a total of 648 BHCs in our bank universe.

We display the time of failure distribution of failed banks in our sample over the years in Figure 2. Most failures are observed to have taken place between 2009 and 2011, a total of 45 out of 55. There is exactly one bank that failed in the early 2000s and one bank that failed later than 2014. We drop both these failed banks from our sample, since our data sample of 2000-2014 doesn't give enough data prior to the first bank failure and the period doesn't include the most recent bank failure. This leaves us with 53 failed banks with unique

[8] The data is available at `https://www.newyorkfed.org/research/banking_research/datasets.html`.

[9] This matches the approach to identify the universe of commercial banks defined by Adrian [3]. It includes all commercial banks, from small community banks to large financial conglomerates. This set does exclude larger banks that have large broker-dealer subsidiaries, such as Bank of America, Citibank, and JP Morgan Chase. While these companies lead the financial industry in size, there are of less relevance for comparison due to their diversified activities and their large size, both of which are not common characteristics of the failed group.

CIKs. We next explain how we extract textual data for the 648 BHCs in our universe. On collecting textual data from SEC filed annual reports, or 10-Ks, for the BHCs in the sample for the period of study, we lose additional banks due to poor textual data, and therefore, end up with 52 failed and 526 non-failed banks as the final universe of BHCs. We will discuss this last drop in bank sample later.

2.2 Textual Data

For guiding our data extraction, we refer to the master file provided by LM [23], which covers all public firms that file to the SEC.[10] We merge our dataset with LM's to find the url link to the corresponding 10-K reports for each BHC in our dataset, for each fiscal year in our study period. Since the last failed bank in our universe of banks failed in 2013, we collect 10-Ks for all banks up to and including 2012.

The time distribution of SEC filings over fiscal years for both failed and non-failed bank groups is summarized in Table 1. Fiscal year 2006 is the year with the largest number of filings by the failed bank group; thereafter the number of filings start to decline for this group. On the other hand, as we observe increase in filings over time for the non-failed banks, it also appears that several non-failed banks got delisted over time. This may be attributed to merger and acquisition activities among non-failed banks, an issue we will come back to in Subsection 2.4.

All 10-K reports submitted in a given fiscal represent a corpus for the BHCs. We extract the corpora covering all fiscal years in our study period, by adhering to the following steps.

- For all bank 10-K reports for fiscal year $t = 2000$,
  - Read the html content using the corresponding url link.
  - Given the html content, drop all tables and figures/images, if applicable.
  - Parse the html content into plain text using a special parser.
  - Convert the document to lowercase and save it as a text file in the folder corresponding to fiscal year t.
- Move to next fiscal year, i.e., $t \to t + 1$.

[10] The data are public and available at http://www3.nd.edu/~mcdonald/Data/LoughranMcDonald_10X_2014.xlsx.

- If $t > 2012$, end process.

Parsing the html content into plain text yields our master corpora of all filings over all fiscal years in the study period. By relying on the dictionaries provided by LM, we map the corpora into a panel dataset of term frequencies for unigrams. Construction of our final panel dataset is, hence, achieved by executing the following steps on each corpus in the corpora:

1. Replace all '-' characters in the corpus with a blank space.

2. Remove punctuations, numbers, and English stop words.

3. Keep terms that show up in the specified dictionary.

4. Perform stemming.

5. Map the corpus into term frequency table using the chosen sentiment dictionary.

We mainly focus on the negative and positive sentiment words for the rest of our analysis. Therefore, for both dictionaries, of positive and negative sentiment words, we represent the related corpora by a corresponding unbalanced panel dataset of term frequencies, where columns refer to the stemmed dictionary term frequencies and rows to company i's report for fiscal year t. While this panel data represents our main textual data for discriminant analysis, we apply a term weighting scheme from which we extract our final feature space. We discuss this in Section 3.

2.3 Financial Data

We consider a number of financial variables as controls, which are commonly used in the 'CAMEL' system for banking. For bank capital, we consider Tier 1 capital and impaired assets ratios along with leverage. On assets quality and management, we consider return on assets (ROA) and return on equity (ROE), respectively. For earnings we relate interest expenses to liabilities, whereas for liquidity, we consider the proportion of short-term borrowing to total liabilities. The definition of these variables is summarized in Table 2. When we merge the financial data with the corresponding panel dataset constructed for textual analysis, the universe of banks further drops by one bank for the non-failed set, which leaves us with 52 failed and 525 non-failed banks.[11] We winsorize financial characteristics at the 1% and 99% level and summarize the financial data for the universe of banks in Table 3.

We observe that all banks are highly leveraged ranging between 74% and 97.2%, which is consistent with the empirical evidence that banks are highly leveraged.[12] Nevertheless, it appears that failed banks were more highly leveraged than the non-failed group. The same observation follows for capital ratios (using common equity or Tier 1). Failed banks were less capitalized than the non-failed ones on average, consistent with the findings of [12, 25]. On the asset quality and management consideration, we discern that failed banks on average have larger proportion of impaired assets, and a lower ROA and ROE. In fact, the average ROE for failed banks is negative.

From the quantitative summary, we also see that the failed banks were associated with greater interest expense ratio than their non-failed counterparts. On the other hand, the liquidity indicator proxied by short-term borrowing over total liabilities does not show much difference between the two groups. This could be explained by the illiquidity of the banking system as a whole that was building up till the unravelling of the financial crisis, as documented by [17].[13] Having observed these financial condition distinctions between failed and non-failed banks in our sample, we will investigate how much additional light textual analysis would be able to shed on the distinction.

2.4 Merger and Acquisition

So far we have distinguished failed banks from non-failed banks, without addressing the financial soundness of the non-failed ones. We identified bank failures with respect to the FDIC filings, however distressed banks could also have been acquired during the time of distress without ever reaching the point of bankruptcy. For instance, from the non-failed bank set, we observe that only 257 banks were active during all fiscal years between 2005 and 2012, while the number of active non-failed banks in fiscal year 2012 alone is 318.

Looking at merger and acquisitions (M&A) activities among BHCs, we identify all banks

[11] In our main results, which rely solely on textual data, we retain the original universe of banks which covers 52 failed and 526 non-failed banks.

[12] For discussion on banks leverage, see [7, 11].

[13] [17] estimate the illiquidity of banking system using the 100 largest BHCs, where they find that illiquidity of the system increased steadily from 2001-Q1 up till 2007-Q4. The authors imply that this estimate of the system's vulnerability could have been useful as an early indicator of the crisis.

that were acquired in our dataset and ceased to exist for each calendar year.[14] It appears that around the financial crisis (between 2006 and 2013 calender years), there were 118 acquisitions, 60% of which took place before 2010. As such acquisitions should not necessarily indicate a bank being in financial distress, but it can be the case that in an environment of scarce capital, banks choose to acquire underpriced assets of other institutions rather than engage in conventional lending activities [27].

If banks were acquired due to financial distress, then their Tier 1 capital should indicate a drop beyond which banks were unable to meet regulatory requirements. We looked at the time series of Tier 1 capital for each of the 118 banks in order to determine whether an acquisition of the bank took place due to financial distress. In all we find 27 (respectively, 9) banks whose last observation of Tier 1 ratio dropped more than one standard deviation (respectively, two standard deviations) below the time series mean. Figure 3 illustrates this drop by plotting the Tier 1 capital ratio of the flagged banks. Additionally, the average Tier 1 ratio for the 27 flagged banks is 6.75%, with median around 7.1%, while these statistics are 1% lower for the group with two standard deviations drop. In our discriminant analysis in Section 4, we will need to pay special attention to this group.

3 Empirical Framework and Methodology

We now describe our main empirical framework and methodology to implement bank failure prediction using textual sentiment analysis. We will need to first extract features from the textual data described in Section 2 for all BHCs over the fiscal years in the study period. To these features, we will apply appropriate weighting scheme before we present our model to map the extracted sentiment features into the classification methodology. The classification approach is designed to determine whether a certain bank is failed or not given the positive and negative sentiment attributes extracted from the corpora. Finally, we outline our prediction framework along with its performance metrics.

[14] Information on M&A activities for BHCs is available at https://www.chicagofed.org/banking/financial-institution-reports/merger-data.

3.1 Feature Extraction

As discussed in Section 2, we parse the html content of all corpora and extract the negative and positive unigrams using the dictionaries proposed by LM [23]. This results in panel data with respect to bank-fiscal years. For the negative (positive) terms, we identify 836 (148) terms that appear at least once for each bank-fiscal year observation. The panel dataset represents a high-dimensional sparse matrix of term frequencies. Instead of frequencies, we rely on term weighting scheme that maps frequencies into scores based on the uniqueness of terms across all documents and other terms. To illustrate the weighting scheme, we provide some notation.

Let Q denote the set of features that we extract with respect to a given dictionary. We denote w_q as the weight of term $q \in Q$, such that

$$w_q = \log\left(\frac{N}{df_q}\right), \tag{3.1}$$

where N is the number of reports in the data and df_q is the number reports containing the term q. This is the term weighting scheme described by [24], which attributes the score of term q with respect to proportion of documents containing the same term. However, this does not account for other terms in the same document. Hence, we adopt a similar weighting scheme used by [23], such that the score of term q in report i is given by

$$w_{i,q} = \begin{cases} [1 + \log(tf_{i,q})w_q] \,/\, [1 + \log(a_i)] & \text{if } tf_{i,q} > 0, \\ 0 & \text{otherwise,} \end{cases} \tag{3.2}$$

where $tf_{i,q}$ is the frequency of term q in report i and a_i is the number of terms that show up in report i.

The weighting scheme in Equation (3.2) implies that the score of term q in report i is determined by its relative frequency with respect to the number of words extracted from report i and the proportion of reports containing the same term. Unlike term frequencies, this weighting scheme is more indicative of the dictionary terms that show up in the corpora. For instance, the term “loss” is defined as negative, but since it is a common term in financial reports it should not have much discriminatory power, and hence, on average it should have a low score.

For all terms and reports in our panel data, we map the term frequencies into weighted scores with using Equations (3.1) and(3.2). In Table 4, we report the mean score of negative and positive terms across failed and non-failed banks. The mean scores are reported with respect to the top ten terms of each sentiment that exhibits greatest discriminatory power, i.e., largest difference in the mean scores between failed and non-failed banks. For instance, in fiscal year 2005, we observe that the negative term “stolen” received higher mean score among failed banks than it did for the non-failed banks. It appears that there are positive words that receive greater average scores among the failed group. The same applies to fiscal year 2008. However, the terms with the greatest average score difference in 2005 are not necessarily the same as in fiscal year 2008, an evidence demonstrating the time dynamics of sentiments.

Table 4 shows that there are certain terms that exhibit greatest discriminatory power between failed and non-failed banks. In order to obtain a perspective on the system level average sentiment over time, we now look at the average negative and positive sentiment across all failed and non-failed banks over time in Figure 4. We observe that on average failed banks exhibit greater sentiment score than their non-failed counterparts, and surprisingly the failed banks indicate greater positive sentiment than the non-failed ones. This suggests that, while facing distress, the failed banks were more optimistic than the non-failed banks. This raises questions about the information disclosure by the management of the failed banks. On one hand, it could be the case that managers were trying their best to uplift their companies from distress. On the other hand, it could be a case of agency problem [19], where the managers were concealing information from the shareholders and the investors in order maximize their consumption of perks before the bank finally failed, which the managers discerned to be inevitable.

3.2 Support Vector Machines

We use a Support Vector Machine (SVM) model to perform discriminant analysis between the failed and non-failed banks. We rely on an SVM approach for two main reasons. The first reason is the high dimensionality of features extracted for textual analysis. Since we are extracting sentiment with respect to LM dictionaries, our extracted feature space for the negative dictionary consists of as many as 833 terms. As a cross-section, we have relatively small number of banks compared with the size of this feature space. SVMs have successfully demonstrated capability of dealing with large feature spaces. The second advantage of the SVM methodology is its out-of-sample prediction robustness. SVM avoids over-fitting by imposing a certain margin for classification. By training, SVM takes into account deviation from the estimated model, which allows for more flexibility in the out-of-sample prediction. We relate this as the margin cost. In our analysis, we rely on SVM with linear kernel function and fixed margin cost. The linearity assumption simplifies our findings and makes the prediction easier to implement manually.[15]

We let $X_{i,t}^Q$ denote the feature space of BHC i covering fiscal year t. The feature space consists of the scores extracted from the 10-K reports with respect to the specified sentiment dictionary, Q. The scores are assigned to each term and bank as per Equation (3.2). Moreover, let $y \in \{-1, +1\}$ denote the status of certain bank, where $y = +1$ is the failed bank label and $y = -1$ is the non-failed label. The objective of our model is to find a linear function that discriminates between the two labels, given an input of the feature space. More formally, we need to find a function g that maps the feature space of $X_{i,t}^Q$ into $y_{i,t} \in \{-1, +1\}$ for bank i and fiscal year t. Such linear function is described by

$$g(X_{i,t}^Q) = \text{sign}(\mathbf{w}' X_{i,t}^Q + \rho), \tag{3.3}$$

where sign$(\cdot)$ is a sign function, $\mathbf{w}$ is the vector of weights allocated to each term score in the feature space, ρ is a constant, and '$\prime$' is the transpose operation.

Equation (3.3) implies that if we know $\mathbf{w}$ and ρ, then we can classify bank i from fiscal year t as failed, if $g(X_{i,t}^Q) = +1$. This implies that determining the state of bank i from fiscal year t depends on finding the optimal parameters, $\mathbf{w}$ and ρ. This is where SVM comes into the picture. In this regard, a linear SVM uses a linear kernel function and finds the optimal weights that discriminate between failed and non-failed banks with respect to a given margin cost.

We use linear kernel for two main reasons. First, the resulting mapping of the original feature space is more tractable and less obscure when using linear kernel than the case of non-linear mapping. Second, for linear kernel, the model is tuned using one input, the margin cost, which can be determined arbitrarily. Since the model's tuning is determined by the

[15] For more information on SVM, see [15].

margin cost alone, then tuning is a less of a concern than the case for non-linear kernels that depend on other inputs. Hence, given the limited number of failed banks in our sample, performing cross-validation leaves the model with smaller set of failed banks for training purpose and should not necessarily increase its predictive power in the test simple. For these reasons, we focus solely on linear kernel and avoid issues with model's tuning.

3.3 Training and Testing

Prediction of bank failures using sentiment relies on training the SVM model and summarizing its performance out-of-sample. We describe the steps of the experiment conducted as follows:

1. Split the full panel into training and testing sets, such that from each bank group 75% unique CIKs are randomly picked for training, while the rest are kept for testing.

2. To avoid data snooping, use the weighting scheme described in Equation (3.2) separately on the training and the test sets.

3. Estimate the SVM model parameters, $\mathbf{w}$ and ρ, from Equation (3.3) using the training set.

4. For each observation x in the test set, classify the bank as failed if $\hat{g}(x) = \hat{\mathbf{w}}'x + \hat{\rho} > 0$, i.e. $\text{sign}(\hat{g}(x)) = +1$. Otherwise, classify the bank as non-failed.

While failed banks show up across different fiscal years in our sample, in practice their true state is only realized ex-post. Nonetheless, we treat all failed banks as failed across all fiscal years regardless of their actual year of failure. That is, if a certain bank, for instance, fails in calender year 2009, the model considers the bank to be failed across all available fiscal years. This approach increases the model's learning process, but it is also likely to result in less emphasis on important distress features that would only show up in the later reports, near the bank's actual year of failure. For this reason, we do not consider reports prior fiscal year 2005, as the information content of these reports are likely to contain more noise than relevant features about the bank's distress. Moreover, since the last failed bank in our set takes place in calender year 2013, reading reports beyond fiscal year 2012 is irrelevant.

Therefore, the training and testing process is focused on all 10-K reports covering all fiscal years between 2005 and 2012 (included).

One of the caveat of the experiment, nonetheless, that it still regards failed banks as failed across all years, which is not the case in practice since as banks fail they drop and cease to exist. We deal with this issue by shrinking the experiment window so it becomes more focused on the cases during which banks filed their very last reports before eventually failing. To serve this purpose, we repeat the experiment multiple times, where each time we drop the earliest fiscal year from the data. We repeat this until the experiment is conducted on the most recent fiscal years, 2009-12.

Since failed banks account for a small proportion of the data, a prediction model that returns high accuracy is not necessarily conclusive. It could be that the model assigns all banks as non-failed, which yields high accuracy due to the weight imbalance between the two groups. Therefore, we consider a number of performance metrics to capture the overall prediction performance:

1. *Accuracy* is the proportion of correctly classified banks regardless of how many failed banks were identified.

2. *Precision* is the proportion of correctly classified failed banks out of the number of failed banks that the model predicts.

3. *Recall* is the proportion of correctly classified failed banks out of the number of actually failed banks.

4. F_1 is a weighted score of *Precision* and *Recall*, give as

$$F_1 = 2 \cdot Precision \cdot Recall/(Precision + Recall). \tag{3.4}$$

One can think of *Precision* and *Recall* in the context of definition of Type II and Type I errors, respectively, of hypothesis testing. Low values of *Precision* could be due to Type II error, where non-failed banks are identified as failed. On the other hand, low *Recall* values imply that the model is assigning failed banks as non-failed. Obviously, Type I error is of greater concern than Type II. If a certain bank is identified as failed while it does not eventually fail, the associated cost is much lower than the other case when a failed bank is misclassified. In the former case, misclassification would result in an increase in the cost of capital and higher premium paid by the bank to the FDIC. Nonetheless, if a failed bank is misclassified as non-failed, then the costs are much greater, which would have repercussions on the economy on the economy, especially when the failed entity is TBTF bank, in which the bank gets bailed out by tax-payers money. Therefore, while we consider all metrics, we put greater emphasis on the model's performance with respect to the *Recall*.

4 Results and Findings

We apply the methodology developed in Section 3 to run multiple models with respect to sentiment dictionaries, banks samples, and feature spaces. First, we start by looking at the complete universe of BHCs in our data with the full feature space extracted using either dictionary. This forms the baseline results from which refinements done thereafter are compared. We then focus on a subset of feature space that exhibits significant discrimination power between failed and non-failed banks. This also helps in dimensionality reduction, which is beneficial for classification accuracy. Third, given the extracted subset of features, we control for mergers and acquisitions by dropping all acquired banks from the non-failed group.[16] Finally, for additional robustness, we add to the failed group the set of acquired banks that had experienced significant decline in their Tier 1 capital to total assets ratio before they were acquired.[17]

4.1 Baseline Results

We build the baseline model in which we consider all failed and non-failed banks. The results are reported with respect to the negative and positive sentiment dictionaries, separately and combined. Table 5 summarizes the baseline results. Panel (a) from Table 5 summarizes the performance metrics with respect to the negative dictionary terms. We note that while accuracy is high across all rows, *Recall* is low. This undermines the predictive ability of the

[16] In this case, we expect the model to achieve its highest discrimination power as we are comparing between failed and surviving banks, instead of the more noisy set of non-failed banks that contain acquired banks and other delisted banks that were not considered failed according to the FDIC.

[17] While this extends the set of failed banks by adding failed candidates, it also adds noise to the model, as these banks did not actually eventually fail.

model using negative sentiment to identify failed banks. We ascribe this poor performance to the high dimensionality of the feature space for the negative dictionary, as we shall discuss in the following subsection.

Looking at Panel (b) from Table 5, we find that the accuracy of the model with respect to the positive dictionary is lower than that for the negative one. However, the *Recall* is much greater, and it ranges between 34% and 60%. Moreover, it is worth noting that all performance metrics increase as the data becomes more concentrated around the financial crisis (moving down in the rows).

Comparison between Panels (a) and (b) implies that positive sentiment has greater power in predicting bank failure than negative sentiment. Hence, a combination of the two dictionaries should yield a better performance than the negative dictionary alone, but worse performance than the positive dictionary alone. This explains the results in Panel (c) where the performance metrics range between their peers in Panels (a) and (b). The feature space for the positive dictionary is much smaller than that for the negative dictionary (145 positive terms versus 833 negative terms). We need to, therefore, consider the dimensionality difference between the two in order to reach a fairer conclusion about the prediction power of each dictionary.

4.2 Dimensionality Reduction

While the SVM model is capable of dealing with high dimensional data, we need to investigate whether the performance of the two dictionaries can be improved by relying on only a subset of the original feature space. In order to accomplish this reduction in dimensionality, we extract terms that show significant score difference between failed and non-failed banks. This creates a trade-off. On one hand, reducing the dimension of the feature space should mitigate over-fitting of the model and increase its out-of-sample prediction reliability. On the other hand, dimension reduction comes at the cost of dropping possible important out-of-sample features.

Given the training data, we conduct two-tails T-test for mean difference between failed and non-failed banks given each term score in the feature space. We keep all features for which the T-test p-value is smaller than 0.01. This, as a result, cuts down feature space dimension almost by 70% for each dictionary. Using this thinner feature space, similar to

Table 5, we report the results with respect to the feature sub-space in Table 6. Interestingly, we observe that the model's performance for the negative dictionary is much better than for the original feature space. This implies that the poor performance of the negative dictionary in Table 5 Panel (a) can be attributed to greater noise in the full feature space rather than the non-informativeness of the negative dictionary. On average, we observe that *Recall* increases significantly when we focus on a feature subset instead of the entire feature space.

For the positive dictionary in Table 6 Panel (b), it appears that the improvement due to dimensionality reduction is trivial. This is due to the fact that the dimension of the original positive feature space is not as large as that for the negative dictionary. Hence, the gain from the reduced feature space does not outweigh the loss of forgoing the larger information in the original feature space that the SVM model is able to utilize. When comparing between Panels (a) and (b) in Table 6, we still observe that the positive dictionary achieves a better performance with respect to *Recall* than the negative dictionary, except in one case (third row). On the other hand, when considering the weighted score between *Precision* and *Recall*, we find that negative sentiment achieves a higher F_1 score than the positive one.

4.3 Controlling for Mergers & Acquisitions

In the previous subsections, we considered the full sample of the non-failed banks regardless of whether these banks were delisted, and therefore, stopped filing 10-Ks over the course of the study period. While considering the full sample should support the robustness of our findings, focusing on the set of non-failed banks that were not delisted should provide us a cleaner perspective on the model's ability to discriminate a failed bank from a non-failed one. Towards this objective, from the non-failed bank set, we drop all banks that were acquired via M&A (in total 118 banks). Therefore, with this modification, the non-failed set now consists of only those banks that were present and filing through out the study period, and in total the universe of non-failed banks reduces to 318 banks.

We repeat the analysis as before and summarize the results in Table 7. Overall, we observe an increase in the performance metrics with respect to all dictionaries. For instance, when the model is trained near the financial crisis (fourth row), the *Recall* increases by 5% and 10% for the negative and positive dictionaries, respectively. We also observe an overall increase in *Accuracy* and *Precision*. The increase in the model's predictability is consistent with the fact that the non-failed set becomes more representative of bank survivorship. In this case, we do expect the model to achieve greater discrimination power than the previous cases summarized in Tables 5 and 6.

For the failed banks, we examine possible failed candidates from the acquired set. As described in Section 2.4, we consider targets that suffered more than two standard deviations drop in their Tier 1 to assets ratio before being acquired. In total we find 27 banks that fit this description, which we add to the universe of failed banks. This increases the failed bank set to 79 banks. We repeat the SVM analysis as before and summarize the results in Table 8. It appears that the discrimination power of the model overall does not improve on adding the failed candidates. This implies that the candidate set does not contain features that are consistent with the failed banks, and hence does not improve the model's prediction power. After all, the suspected targets did not fail, even though they experienced distress in comparison with their acquired peers. One explanation could be that while distressed targets signaled similar sentiment as the failed group, it did convey different content in expectation of acquisition.

5 Conclusion

In this paper we propose a novel framework for assessing a bank's soundness using textual sentiment analysis. Looking at 10-K reports filed by publicly listed BHCs, we study the link between the disclosed sentiment in these filings and the BHCs performance during the study period, which includes the 2007-09 financial crisis. We mainly focus on negative and positive sentiments, where the performance of the prediction is captured by whether a BHC actually failed or not. On average, we find that both type of sentiments discriminate between failed and non-failed banks 80% of the time. Additionally, out of ten failed banks, on average positive sentiment can identify seven true events, while negative sentiment identifies five failed banks at most.

We look at the recent crisis as a natural experiment during which large number of public banks failed. However, our framework should not be constrained solely to a crisis epoch, or necessarily to the recent financial crisis experience. A future research could extend our framework to study beyond the recent financial crisis and utilize other sources of textual information, i.e. incorporate different text sources beyond that contained in annual 10-K reports. Furthermore, most online filings start during the early 1990s. Hence, expanding our sample to incorporate the 1980s Savings and Loan (S&L) crisis, which also originated in the banking sector and resulted in large number of bank failures, would be significant.

Another possible line of research could investigate the difference in sentiment between banks and non-banks. For instance, a recent research by [11] tries to explain why banks are more leveraged than non-banks, where the authors attribute this to asset risk held by banks. Nonetheless, the lesson from the recent financial crisis is that banks were manufacturing tail risk that was systematic in nature [1]. Since we find that positive sentiment played a stronger role in predicting financial distress (i.e., failure), how would this be different for other non-bank firms. The question to ask would be, was it a systematic practice that failed bank pursued while facing distress. We leave these investigations for future research.

References

[1] Viral V Acharya, Thomas Cooley, and Matthew Richardson. *Manufacturing tail risk: A perspective on the financial crisis of 2007-2009.* Now Publishers Inc, 2010.

[2] Anat R Admati, Peter M DeMarzo, Martin F Hellwig, and Paul C Pfleiderer. Fallacies, irrelevant facts, and myths in the discussion of capital regulation: Why bank equity is not expensive. *MPI Collective Goods Preprint*, (2010/42), 2011.

[3] Tobias Adrian, Nina Boyarchenko, and Hyun Song Shin. The cyclicality of leverage. *FRB of New York Working Paper No. FEDNSR743*, 2015.

[4] Tobias Adrian and Hyun Song Shin. Liquidity and leverage. *Journal of financial intermediation*, 19(3):418–437, 2010.

[5] Edward I Altman. Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The journal of finance*, 23(4):589–609, 1968.

[6] Werner Antweiler and Murray Z Frank. Is all that talk just noise? the information content of internet stock message boards. *The Journal of Finance*, 59(3):1259–1294, 2004.

[7] Viral V Archarya, Hamid Mehran, Til Schuermann, and Anjan V Thakor. Robust capital regulation. *Current Issues in Economics and Finance*, 18(4), 2012.

[8] William H Beaver. Financial ratios as predictors of failure. *Journal of accounting research*, pages 71–111, 1966.

[9] William H Beaver. Market prices, financial ratios, and the prediction of failure. *Journal of accounting research*, pages 179–192, 1968.

[10] Timothy B Bell, Gary S Ribar, and Jennifer Verchio. Neural nets versus logistic regression: a comparison of each modelâĂŹs ability to predict commercial bank failures. In *Proceedings of the 1990 Deloitte and Touche/University of Kansas Symposium of Auditing Problems, Lawrence, KS*, pages 29–58, 1990.

[11] Tobias Berg and Jasmin Gider. What explains the difference in leverage between banks and non-banks? *Journal of Financial and Quantitative Analysis (JFQA), Forthcoming*, 2016.

[12] Allen N Berger and Christa HS Bouwman. How does capital affect bank performance during financial crises? *Journal of Financial Economics*, 109(1):146–176, 2013.

[13] J Efrim Boritz, Duane B Kennedy, et al. Predicting corporate failure using a neural network approach. *Intelligent Systems in Accounting, Finance and Management*, 4(2):95–111, 1995.

[14] Mark Cecchini, Haldun Aytug, Gary J Koehler, and Praveen Pathak. Making words work: Using financial text as a predictor of financial events. *Decision Support Systems*, 50(1):164–175, 2010.

[15] Nello Cristianini and John Shawe-Taylor. *An introduction to support vector machines and other kernel-based learning methods*. Cambridge university press, 2000.

[16] Timothy J Curry, Gary S Fissel, and Peter J Elmer. Can the equity markets help predict bank failures? 2004.

[17] Fernando Duarte and Thomas M Eisenbach. Fire-sale spillovers and systemic risk. *FRB of New York Staff Report*, (645), 2015.

[18] Sven S Groth and Jan Muntermann. An intraday market risk management approach based on textual analysis. *Decision Support Systems*, 50(4):680–691, 2011.

[19] Michael C Jensen and William H Meckling. Theory of the firm: Managerial behavior, agency costs and ownership structure. *Journal of financial economics*, 3(4):305–360, 1976.

[20] James Kolari, Dennis Glennon, Hwan Shin, and Michele Caputo. Predicting large us commercial bank failures. *Journal of Economics and Business*, 54(4):361–387, 2002.

[21] William R Lane, Stephen W Looney, and James W Wansley. An application of the cox proportional hazards model to bank failure. *Journal of Banking & Finance*, 10(4):511–531, 1986.

[22] Feng Li. Do stock market investors understand the risk sentiment of corporate annual reports? *Available at SSRN 898181*, 2006.

[23] Tim Loughran and Bill McDonald. When is a liability not a liability? textual analysis, dictionaries, and 10-ks. *The Journal of Finance*, 66(1):35–65, 2011.

[24] Christopher D Manning and Hinrich Schütze. *Foundations of statistical natural language processing*, volume 999. MIT Press, 1999.

[25] Hamid Mehran and Anjan Thakor. Bank capital and value in the cross-section. *Review of Financial Studies*, 24(4):1019–1067, 2011.

[26] Arman Khadjeh Nassirtoussi, Saeed Aghabozorgi, Teh Ying Wah, and David Chek Ling Ngo. Text mining for market prediction: A systematic review. *Expert Systems with Applications*, 41(16):7653–7670, 2014.

[27] Andrei Shleifer and Robert W Vishny. Unstable banking. *Journal of financial economics*, 97(3):306–318, 2010.

[28] Jie Sun, Hui Li, Qing-Hua Huang, and Kai-Yu He. Predicting financial distress and corporate failure: A review from the state-of-the-art definitions, modeling, sampling, and featuring approaches. *Knowledge-Based Systems*, 57:41–56, 2014.

[29] Kar Yan Tam and Melody Y Kiang. Managerial applications of neural networks: the case of bank failure predictions. *Management science*, 38(7):926–947, 1992.

[30] Paul C Tetlock. Giving content to investor sentiment: The role of media in the stock market. *The Journal of Finance*, 62(3):1139–1168, 2007.

[31] Shaonan Tian, Yan Yu, and Hui Guo. Variable selection and corporate bankruptcy forecasts. *Journal of Banking & Finance*, 52:89–100, 2015.

[32] Matthias W Uhl, Mads Pedersen, and Oliver Malitius. WhatâĂŹs in the news? using news sentiment momentum for tactical asset allocation. *The Journal of Portfolio Management*, 41(2):100–112, 2015.

Figures

Figure 1: **Data Construction**
This figure demonstrates how to link the FDIC data to the SEC EDGAR system. This requires a bridge between the FDIC certificate number and the CIK which is the key identification number used by the SEC to identify public companies. We first find the corresponding identification number used by regulators to identify commercial banks, and then link the commercial banks to their parent holding company. Since the CIK number is available in the CRSP-COMPUSTAT dataset, we find the corresponding CRSP's permanent company identifier (PERMCO) for each bank holding company (BHC) from the Federal Reserve Bank of New York. Finally, by merging with the CRSP-COMPUSTAT we identify the corresponding CIK for each BHC in our sample, including the failed ones.

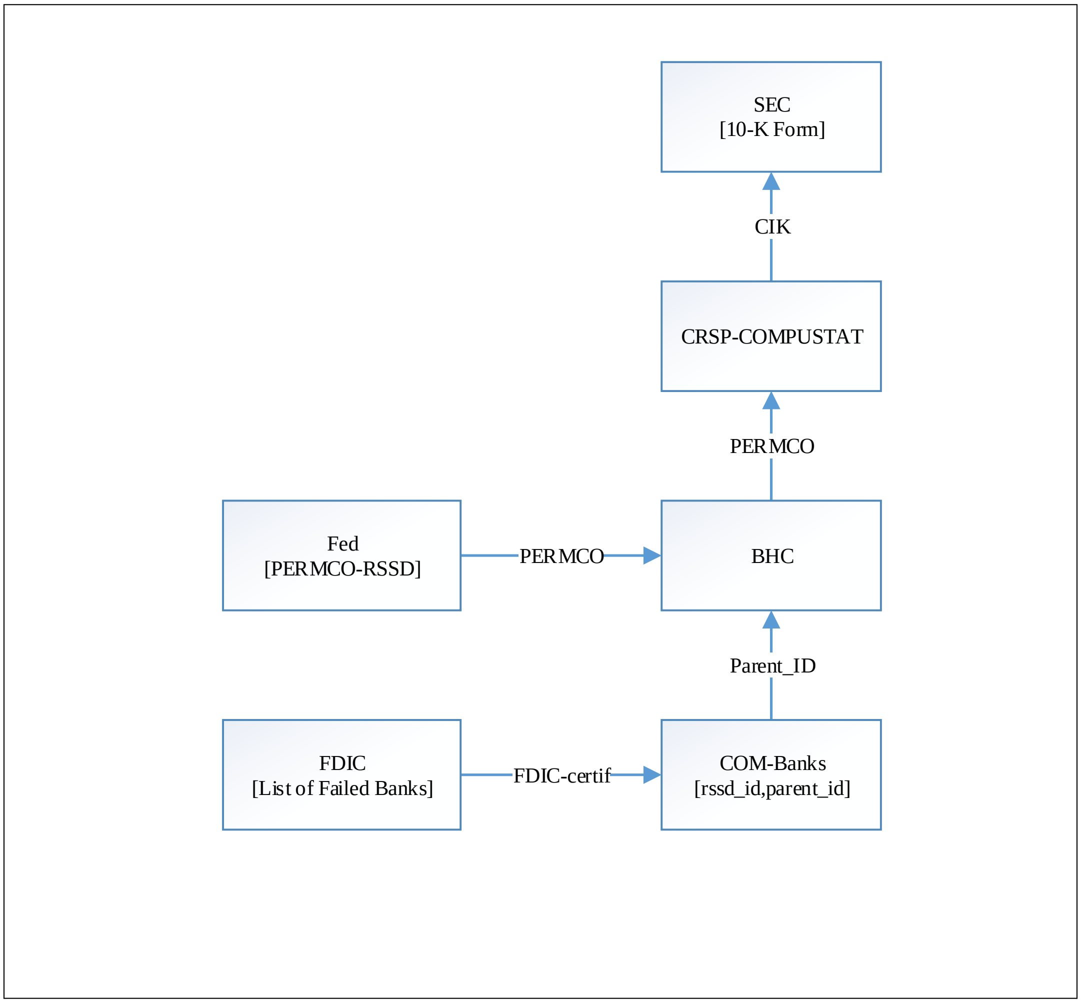


Figure 2: **Distribution of Public Banks Failure**
This histogram demonstrates the distribution of public failed banks, which we identify over the calender years starting from 2000. The earliest failure takes place in 2002 while the latest one does occur in 2015.

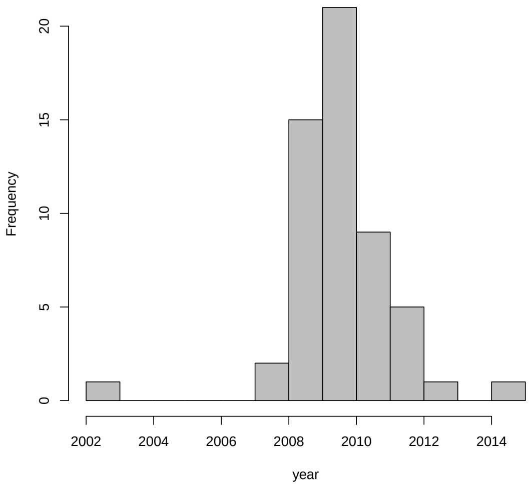


Figure 3: **Time Series of Tier 1 Capital Ratio for Acquired Banks**

This figure demonstrates the Tier 1 capital ratio over time for 16 randomly selected bank holding companies (BHCs) that were acquired between 2006 and 2013 calender years and whose last Tier 1 ratio experienced more than one standard deviation drop from the time series mean before being acquired. In total, there are 27 such companies identified in our sample. The numbers in each single plot refer to the unique identification number of BHCs as recognized in the FR Y9-C report, i.e. RSSD9001.

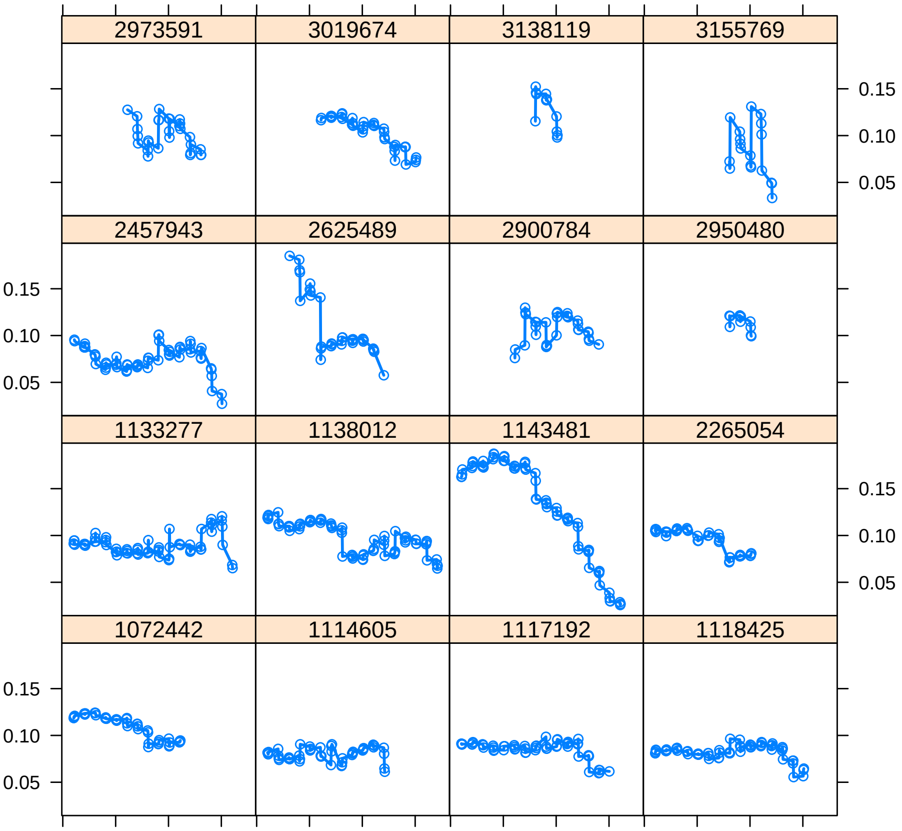
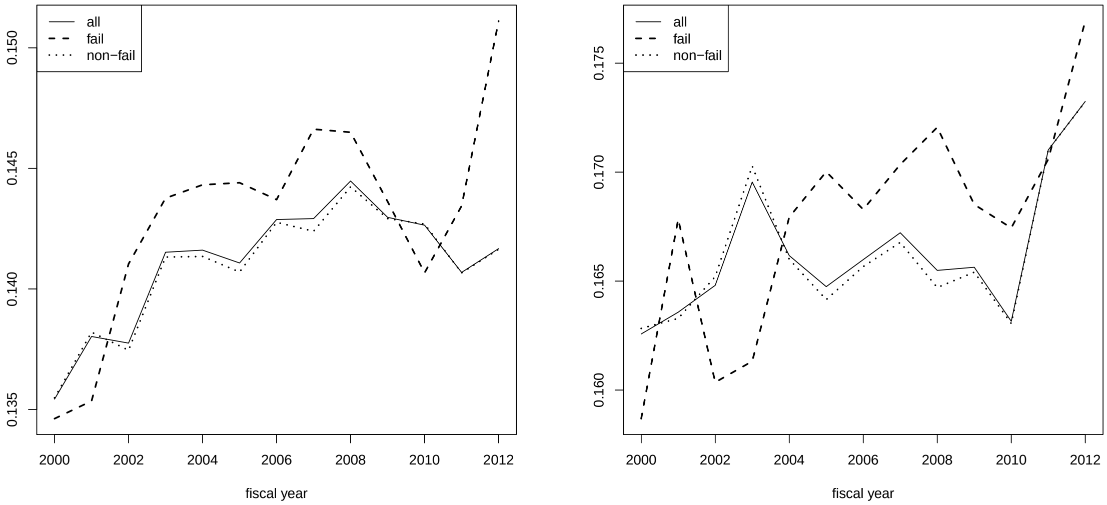


Figure 4: **Banks Aggregate Sentiment Over Fiscal Years**
This figure plots the average sentiment score for all 10-K reports in our banking universe over all fiscal years between 2000 and 2012. Terms from the reports are defined as negative or positive with respect to the dictionaries provided by Loughran and McDonald [23]. The sentiment score for each report in a given fiscal year is computed with respect to the weighting scheme described in Equation (3.2) by equally weighing the term scores in the report. For each fiscal year, the aggregate sentiment is computed by equally weighing all 10-K reports sentiment scores across banks. The aggregate sentiment score is computed for the system as a whole for all banks as well as separately for the failed and non-failed groups. Panel (a) and (b) refer to negative and positive sentiment, respectively.



(a) Negative

(b) Positive

Tables

Table 1: **Distribution of 10-K Filings over Fiscal Years**
This table reports the number of fillings distribution over fiscal years starting from 2000 till 2012. f denotes the frequency of 10-K reports that were submitted for a given fiscal year, whereas F denotes the cumulative relative frequency of submitted reports over the total fiscal years.

	Failed Banks		Non-Failed Banks	
Fiscal Year	f	F	f	F
2000	17	0.05	259	0.06
2001	17	0.10	254	0.12
2002	33	0.19	366	0.20
2003	35	0.29	392	0.29
2004	35	0.38	362	0.37
2005	42	0.50	373	0.46
2006	50	0.64	359	0.54
2007	48	0.78	339	0.62
2008	41	0.89	342	0.70
2009	27	0.97	343	0.77
2010	8	0.99	337	0.85
2011	2	1.00	330	0.93
2012	1	1.00	319	1.00

Table 2: **Financial Variables Definition**

Variable	Definition
lev	Leverage: 1 minus Capital Ratio
cap_ratio	Capital Ratio: Total Equity / Total Assets
tier1_ratio	Tier1 Ratio: Tier1 Capital / Total Assets
imp_at	Impaired Assets: Non Performing Assets / Total Assets
roa	Return on Assets
roe	Return on Equity
int_expense	Interest Expenses / Total Liabilities
short_borrowing	Short-term borrowing / Total Liabilities

Table 3: **Summary Statistics of Banks Financial Characteristics**
This table provides summary statistics for the financial characteristics of the failed and non-failed banks. Variables definition is provided in Table 2. Variables are winsorized the at %1 and %99 levels.

Statistic	N	Mean	St. Dev.	Min	Max
Panel (a) Full Sample					
lev	4,612	0.907	0.034	0.740	0.972
cap_ratio	4,612	0.093	0.034	0.028	0.260
tier1_ratio	4,423	0.013	0.013	0.001	0.093
imp_at	4,548	0.014	0.019	0.000	0.093
roa	4,612	0.006	0.011	−0.047	0.023
roe	4,612	0.053	0.189	−1.064	0.251
int_expense	4,560	0.022	0.011	0.002	0.050
short_borrowing	4,562	0.012	0.027	0.000	0.142
fail	4,612	0.073	0.260	0	1
Panel (b) Failed Banks					
lev	337	0.920	0.028	0.816	0.972
cap_ratio	337	0.080	0.028	0.028	0.184
tier1_ratio	329	0.012	0.010	0.001	0.059
imp_at	337	0.020	0.027	0.000	0.093
roa	337	0.002	0.017	−0.047	0.023
roe	337	−0.027	0.344	−1.064	0.251
int_expense	337	0.028	0.009	0.005	0.050
short_borrowing	336	0.013	0.028	0.000	0.142
Panel (c) Non-Failed Banks					
lev	4,275	0.906	0.034	0.740	0.972
cap_ratio	4,275	0.094	0.034	0.028	0.260
tier1_ratio	4,094	0.013	0.013	0.001	0.093
imp_at	4,211	0.014	0.018	0.000	0.093
roa	4,275	0.006	0.010	−0.047	0.023
roe	4,275	0.059	0.169	−1.064	0.251
int_expense	4,223	0.022	0.010	0.002	0.050
short_borrowing	4,226	0.012	0.027	0.000	0.142

Table 4: **Top Ten Words with Largest Difference Between Failed and Non-Failed Banks**
This table reports the average sentiment score of the top ten terms that exhibits greatest discrimination between failed and non-failed banks. Terms are defined as negative or positive with respect to the dictionaries provided by Loughran and McDonald [23]. The sentiment score for each report in a given fiscal year is computed with respect to the weighting scheme described in Equation (3.2). Panels (a) and (b) refer the mean term score for fiscal year 2005 and 2008, respectively. The variables neg1 and neg0 correspond to the negative term score among failed and non-failed banks respectively, whereas the neg1 - neg0 is the difference between the two. The variables pos1 and pos0 correspond to the positive term score among failed and non-failed banks respectively, whereas the pos1 - pos0 is the difference between the two. The mean scores are sorted in descending order with respect to the neg1 - neg0 and pos1 - pos0.

rank	negative term	neg1	neg0	neg1 - neg0	positive term	pos1	pos0	pos1 - pos0
				Panel (a) Fiscal Year 2005				
1	stolen	0.27	0.15	0.12	perfect	0.26	0.19	0.07
2	complic	0.21	0.14	0.08	impress	0.22	0.15	0.07
3	annul	0.22	0.14	0.07	tremend	0.22	0.15	0.07
4	laps	0.25	0.17	0.07	conclus	0.27	0.20	0.07
5	aberr	0.18	0.12	0.06	popular	0.22	0.16	0.05
6	destroy	0.23	0.17	0.06	valuabl	0.23	0.19	0.05
7	harass	0.19	0.13	0.06	outperform	0.19	0.15	0.04
8	abrog	0.18	0.12	0.06	win	0.18	0.14	0.04
9	involuntari	0.23	0.17	0.06	lucrat	0.18	0.14	0.04
10	moratorium	0.20	0.14	0.06	excit	0.19	0.16	0.04
				Panel (b) Fiscal Year 2008				
1	injunct	0.23	0.16	0.07	progress	0.26	0.18	0.08
2	interfer	0.23	0.16	0.07	dilig	0.24	0.17	0.07
3	counterclaim	0.21	0.14	0.07	proactiv	0.23	0.19	0.05
4	closur	0.20	0.14	0.06	regain	0.21	0.16	0.05
5	assert	0.18	0.13	0.05	confid	0.20	0.16	0.04
6	insubordin	0.18	0.13	0.05	superior	0.23	0.19	0.04
7	controversi	0.21	0.16	0.05	unmatch	0.18	0.14	0.03
8	suspend	0.22	0.17	0.05	creativ	0.18	0.14	0.03
9	complaint	0.22	0.17	0.05	satisfact	0.23	0.20	0.03
10	alleg	0.23	0.18	0.05	except	0.21	0.17	0.03

Table 5: **Out-of-sample prediction using full panel data and feature space**
This table reports the performance results of Support Vector Machines (SVMs) trained on the original feature space using unigrams score with weighting scheme described in (3.2) and sentiment dictionaries provided by [23]. The analysis is performed using all fiscal years between 2005 and 2012 included on a cross-temporal level. The first row uses all fiscal years, whereas each following row drops one fiscal year. Accuracy is the proportion of correctly classified observations. Precision is the ratio between the total of correctly classified failed banks and those identified by the model as failed. Recall is the ratio between the total of correctly classified failed banks and those that are actually failed. The F_1 score is a weighted score between Precision and Recall, where it holds that $F_1 = 2 \cdot \mathit{Precision} \cdot \mathit{Recall} / (\mathit{Precision} + \mathit{Recall})$. Results are reported in percentages.

Fiscal Years Dropped	*Accuracy*	*Precision*	*Recall*	F_1
Panel (a) Negative Sentiment				
none	88.09	11.11	10.71	10.91
2005-06	89.08	9.76	8.89	9.30
2005-07	91.86	13.04	9.38	10.91
2005-08	93.57	7.14	5.00	5.88
Panel (b) Positive Sentiment				
none	74.00	9.69	33.93	15.08
2005-06	75.35	10.78	40.00	16.98
2005-07	78.57	11.20	43.75	17.83
2005-08	83.33	13.79	60.00	22.43
Panel (c) Negative and Positive Sentiment				
none	86.51	10.14	12.50	11.20
2005-06	88.38	13.46	15.56	14.43
2005-07	90.03	11.11	12.50	11.76
2005-08	93.57	12.50	10.00	11.11

Table 6: **Out-of-sample prediction using full panel data and sub-feature space**
The results reported in this table follow suit with respect to Table 5. Instead of the full feature space, this table reports the performance of the model with respect to the reduced feature space as described in Subsection 4.2.

Fiscal Years Dropped	*Accuracy*	*Precision*	*Recall*	F_1
Panel (a) Negative Sentiment				
none	78.86	10.67	28.57	15.53
2005-06	80.11	12.98	37.78	19.32
2005-07	81.89	10.31	31.25	15.50
2005-08	87.75	16.39	50.00	24.69
Panel (b) Positive Sentiment				
none	70.84	10.00	41.07	16.08
2005-06	72.13	8.15	33.33	13.10
2005-07	74.09	10.26	50.00	17.02
2005-08	78.71	10.91	60.00	18.46
Panel (c) Negative and Positive Sentiment				
none	81.77	13.28	30.36	18.48
2005-06	81.37	14.52	40.00	21.30
2005-07	81.23	10.68	34.38	16.30
2005-08	84.34	10.81	40.00	17.02

Table 7: **Out-of-sample prediction using sub-panel data (i) and sub-feature space**
The results reported in this table follow suit with respect to Table 6. The table reports the performance of the model given the reduced feature space as described in Subsection 4.2. The only difference from Table 6 is the sample of non-failed banks, which excludes all banks that were acquired over the sample span, as discussed in Subsection 4.3.

Fiscal Years Dropped	*Accuracy*	*Precision*	*Recall*	F_1
Panel (a) Negative Sentiment				
none	80.17	17.02	28.57	21.33
2005-06	81.52	21.11	42.22	28.15
2005-07	82.26	19.23	46.88	27.27
2005-08	82.40	16.18	55.00	25.00
Panel (b) Positive Sentiment				
none	72.94	16.13	44.64	23.70
2005-06	73.90	15.15	44.44	22.60
2005-07	77.38	13.54	40.62	20.31
2005-08	80.00	16.87	70.00	27.18
Panel (c) Negative and Positive Sentiment				
none	81.51	23.00	41.07	29.49
2005-06	81.52	20.45	40.00	27.07
2005-07	84.48	21.21	43.75	28.57
2005-08	85.33	15.69	40.00	22.54

Table 8: **Out-of-sample prediction using sub-panel data (ii) and sub-feature space**
The results reported in this table follow suit with respect to Table 7. The only difference from Table 7 is the sample of failed banks, which includes acquired banks that experiences significant drop in their Tier 1 capital ratio before being acquired, as discussed in Subsection 4.3.

Fiscal Years Dropped	*Accuracy*	*Precision*	*Recall*	F_1
Panel (a) Negative Sentiment				
none	73.87	30.95	50.00	38.24
2005-06	74.07	27.27	41.38	32.88
2005-07	76.29	30.08	56.06	39.15
2005-08	75.12	22.68	46.81	30.56
Panel (b) Positive Sentiment				
none	71.07	25.60	41.35	31.62
2005-06	71.43	22.63	35.63	27.68
2005-07	70.93	18.49	33.33	23.78
2005-08	71.64	19.82	46.81	27.85
Panel (c) Negative and Positive Sentiment				
none	62.94	14.85	32.69	20.42
2005-06	62.76	15.42	37.93	21.93
2005-07	60.37	10.33	28.79	15.20
2005-08	70.52	15.70	40.43	22.62