Investigating Bank Failures Using Text Mining

Aparna Gupta
Lally School of Management
Rensselaer Polytechnic Institute
Email: guptaa@rpi.edu

Majeed Simaan
Lally School of Management
Rensselaer Polytechnic Institute
Email: simaam@rpi.edu

Mohammed J. Zaki
Department of Computer Science
Rensselaer Polytechnic Institute
Email: zaki@cs.rpi.edu

***Abstract*—We extend beyond healthiness assessment of banks using quantitative financial data by applying textual sentiment analysis. Looking at public annual reports for a large sample of U.S. banks in the 2000-2014 period, we identify 52 public bank holding companies that were associated with bank failures during the global financial crisis. Utilizing sentiment dictionaries designed for financial context, we find that negative and positive sentiments discriminate between failed and non-failed banks 88% and 79%, respectively, 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 six true events, while 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.**

I. Introduction

Given the substantial increase in publicly available textual data along with the innovation in textual tools to analyze 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 textual sentiment analysis by looking at 10-K annual reports for a large sample of U.S. banks in the 2000-2014 time period.[1] We rely on the sentiment dictionaries proposed by Loughran and McDonald [1] (henceforth 'LM') and identify 52 bank holding companies (henceforth 'BHC') that were associated with failure over the global financial crisis.[2] Our findings establish a strong link between sentiment and financial soundness of banks.

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 [2]. 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 [3], [4]. Later research has implemented artificial intelligence tools to predict corporate failure using financial data [5], [6].[3] For specifically banking, different lines of research also used diverse methodologies to predict bank failures. For example, [8] introduces a neural networks approach to predict failures of Texas banks between 1985 and 1987.[4] To best of our knowledge, none of the existing research papers 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, [10], [11], [12], [13], [14] look into the impact of textual analysis on the equity market. [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 [16], who look at the power of text in predicting catastrophic financial events related to fraud or company's bankruptcy. The authors analyze annual corporate disclosures (10-K reports) 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.[7] 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

[1]By regulations, public companies in the United States are required to disclose information on both quarterly (10-Q) and annually (10-K) basis. We mainly focus on the annual 10-K report, since it covers more relevant textual data with forward looking information. Mainly such information is concentrated in Item 7 of report titled as '*Management's Discussion and Analysis of Financial Condition and Results of Operations*'. For further discussion on this, see e.g. [1]

[2]According to the Federal Deposit Insurance Corporation, 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 BHCs.

[3]For a recent review of common predictors used in the literature in predicting corporate bankruptcy, see [7].

[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 [9].

[6]For a systematic review on text mining for market prediction see [15].

[7]In unigrams, we refer to specific words in the financial report that appear in a given sentiment dictionary. This is also known as the bag-of-words approach.

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.[8] 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 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 a support vector machine (henceforth '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 $79\% - 94\%$ and $71\% - 83\%$, respectively.[9] 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. On overall, 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 six failed banks, whereas negative sentiment identifies at most five failed banks out of true events.

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 in 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 II, 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 III 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 IV covers the findings of our papers in different test cases, while Section V concludes the paper.

II. 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 Security Exchange Commission (henceforth 'SEC') filings of U.S. banks in a time period prior to and including the global financial crisis.

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 III. 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 Electronic Data Gathering, Analysis, and Retrieval system (henceforth 'SEC EDGAR') 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

[8] tf.idf refers to the term-frequency (tf) multiplied by the inverse frequency among documents (idf), hence the term tf.idf. The intuition behind this weighting scheme is to adjust the frequency of a given term with respect to its popularity among other documents.

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

link between common structured data and the unstructured textual data to construct our final dataset upon which our empirical framework is applied.

A. The Universe of Banks

We identify failed banks using the Federal Deposit Insurance Corporation (henceforth '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.[10] 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.[11] 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.

[10]The dataset is available at https://www.newyorkfed.org/research/banking_research/datasets.html.

[11]This matches the approach to identify the universe of commercial banks defined by [17]. 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.

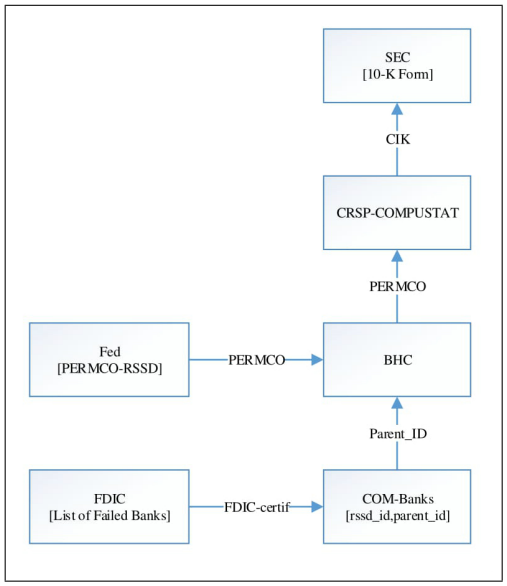

Fig. 1. **Data Construction**

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 does not include the most recent bank failure. This leaves us with 53 failed banks with unique 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.

B. Textual Data

For guiding our data extraction, we refer to the master file provided by LM [1], which covers all public firms that file to the SEC.[12] 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.

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.

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

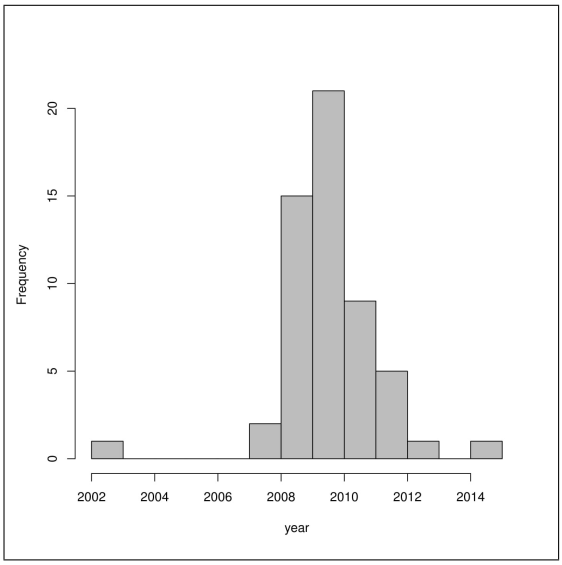

Fig. 2. **Distribution of Public Banks Failure**

  - 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., $t \rightarrow t+1$.
- 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 III.

III. 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 II 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.

A. Feature Extraction

As discussed in Section II, we parse the html content of all corpora and extract the negative and positive unigrams using the dictionaries proposed by LM [1]. 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 a 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

$$w_q = \log\left(\frac{N}{df_q}\right), \tag{1}$$

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 [18], 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 [1], such that the score of term q in report i is given by

$$w_{i,q} = \begin{cases} \frac{1+\log(tf_{i,q})w_q}{1+\log(a_i)} & \text{if } tf_{i,q} > 0 \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

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, while w_q follows from Equation (1).

The weighting scheme in Equation (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 using Equations (1) and(2). In Table I, 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 five 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" receives higher mean score among failed banks than it does 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 I
TOP FIVE TERMS WITH LARGEST DIFFERENCE BETWEEN FAILED AND NON-FAILED BANKS

Rank	Term	Failed	Non-Failed	Difference
		Negative Sentiment		
		Panel (a) Fiscal Year 2005		
1	stolen	0.27	0.15	0.12
2	complic	0.21	0.14	0.08
3	annul	0.22	0.14	0.07
4	laps	0.25	0.17	0.07
5	aberr	0.18	0.12	0.06
		Panel (b) Fiscal Year 2008		
1	injunct	0.23	0.16	0.07
2	interfer	0.23	0.16	0.07
3	counterclaim	0.21	0.14	0.07
4	closur	0.20	0.14	0.06
5	assert	0.18	0.13	0.05
		Positive Sentiment		
		Panel (a) Fiscal Year 2005		
1	perfect	0.26	0.19	0.07
2	impress	0.22	0.15	0.07
3	tremend	0.22	0.15	0.07
4	conclus	0.27	0.20	0.07
5	popular	0.22	0.16	0.05
		Panel (b) Fiscal Year 2008		
1	progress	0.26	0.18	0.08
2	dilig	0.24	0.17	0.07
3	proactiv	0.23	0.19	0.05
4	regain	0.21	0.16	0.05
5	confid	0.20	0.16	0.04

Table I 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 look at the average negative and positive sentiment across all failed and non-failed banks over time in Figure 3. 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.

B. Support Vector Machines

We use an 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 a 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 an SVM model with linear kernel function and fixed margin cost. The linearity assumption simplifies our findings and makes the prediction easier to implement manually.

We let $X^Q_{i,t}$ 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 (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 $X^Q_{i,t}$ into $y_{i,t} \in \{-1, +1\}$ for bank i and fiscal year t. Such a linear function is described by

$$g(X^Q_{i,t}) = \text{sign}(\mathbf{w}' X^Q_{i,t} + \rho), \tag{3}$$

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) implies that if we know $\mathbf{w}$ and ρ, then we can classify bank i from fiscal year t as failed, if $g(X^Q_{i,t}) = +1$. 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 a 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 margin cost alone, then tuning is 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,

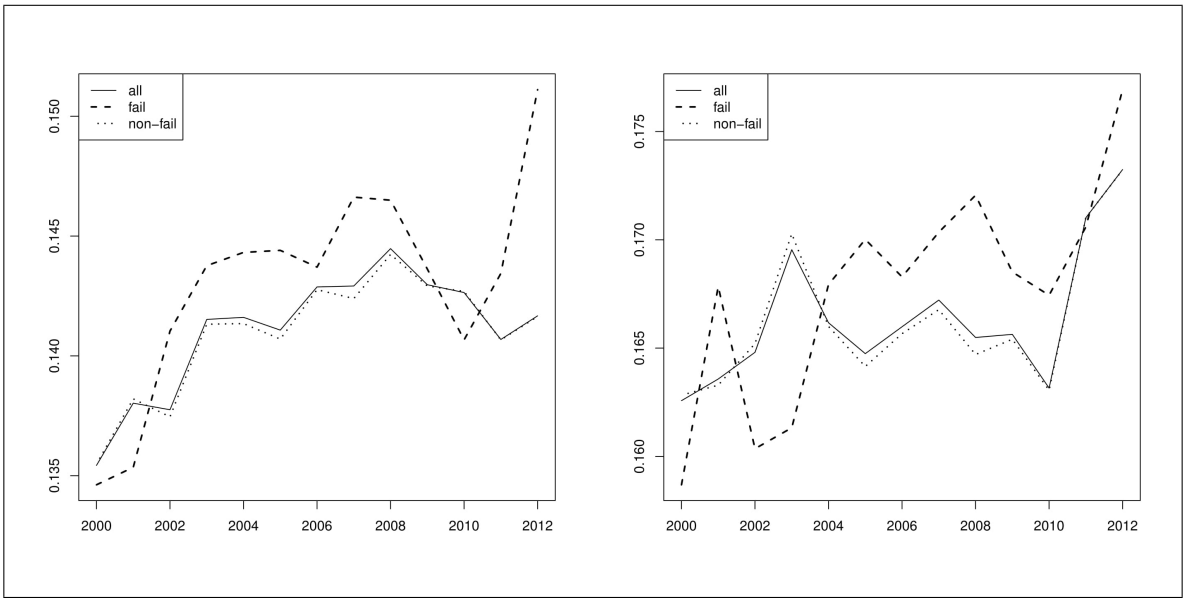

Fig. 3. **Banks Aggregate Sentiment Over Fiscal Years**

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.

C. 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 (2) separately on the training and the test sets.
3) Estimate the SVM model parameters, $\mathbf{w}$ and ρ, from Equation (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 to 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 caveats of the experiment, nonetheless, is that it regards failed banks as failed across all years, which is not the case in practice. We only observe banks to be failed ex-post, after they actually fail. To mitigate this issue, we shrink the experiment window so that the experiment sample becomes more concentrated around the period in which failures take place. 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, i.e. 2009-12.

In practice, prediction of bank's i state at year $t+1$ should be conducted using feature space $X_{i,t}$ or all information available ex-ante. Nevertheless, given the small number of actual events in our data, which mostly occur in 2009 and 2010, such approach would lack statistical significance, since the sample size of the training and testing samples becomes much smaller, which translates into low degrees of freedom. Another possible remedy is to use the leave-one-out approach as conducted by [16], who try to predict corporate (non-banks) catastrophic events using small data sample. The leave-one-out utilizes more training information and provides additional robustness. Nevertheless, it is still unclear how such approach deals with longitude data (time series and cross section). We leave this for future research.

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

$$F_1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}. \quad (4)$$

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, 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 $Recall$.

IV. Results and Findings

We apply the methodology developed in Section III to run multiple models with respect to sentiment dictionaries and feature spaces. Our main findings are summarized in Tables II and III.

A. Baseline Results

We build the baseline model in which we consider all failed and non-failed banks. The results are reported in Table II with respect to the negative and positive sentiment dictionaries, separately and combined. Panel (a) from Table II 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 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 II, 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.

TABLE II
Out-of-sample prediction using full panel data and feature space

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

B. 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 overfitting 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-tailed T-tests 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 II, we report the results with respect to the feature subspace in Table III. 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 II 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 III 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 III, 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.

TABLE III
OUT-OF-SAMPLE PREDICTION USING FULL PANEL DATA AND SUB-FEATURE SPACE

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

V. 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 negative and positive sentiments discriminate between failed and non-failed banks 88% and 79% of the time. Additionally, out of ten failed banks, on average positive sentiment can identify six 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. 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.

REFERENCES

[1] T. Loughran and B. McDonald, "When is a liability not a liability? textual analysis, dictionaries, and 10-ks," *The Journal of Finance*, vol. 66, no. 1, pp. 35–65, 2011.

[2] E. I. Altman, "Financial ratios, discriminant analysis and the prediction of corporate bankruptcy," *The journal of finance*, vol. 23, no. 4, pp. 589–609, 1968.

[3] W. H. Beaver, "Financial ratios as predictors of failure," *Journal of accounting research*, pp. 71–111, 1966.

[4] ——, "Market prices, financial ratios, and the prediction of failure," *Journal of accounting research*, pp. 179–192, 1968.

[5] T. B. Bell, G. S. Ribar, and J. Verchio, "Neural nets versus logistic regression: a comparison of each models ability to predict commercial bank failures," in *Proceedings of the 1990 Deloitte and Touche/University of Kansas Symposium of Auditing Problems, Lawrence, KS*, 1990, pp. 29–58.

[6] J. E. Boritz, D. B. Kennedy *et al.*, "Predicting corporate failure using a neural network approach," *Intelligent Systems in Accounting, Finance and Management*, vol. 4, no. 2, pp. 95–111, 1995.

[7] S. Tian, Y. Yu, and H. Guo, "Variable selection and corporate bankruptcy forecasts," *Journal of Banking & Finance*, vol. 52, pp. 89–100, 2015.

[8] K. Y. Tam and M. Y. Kiang, "Managerial applications of neural networks: the case of bank failure predictions," *Management science*, vol. 38, no. 7, pp. 926–947, 1992.

[9] J. Sun, H. Li, Q.-H. Huang, and K.-Y. He, "Predicting financial distress and corporate failure: A review from the state-of-the-art definitions, modeling, sampling, and featuring approaches," *Knowledge-Based Systems*, vol. 57, pp. 41–56, 2014.

[10] W. Antweiler and M. Z. Frank, "Is all that talk just noise? the information content of internet stock message boards," *The Journal of Finance*, vol. 59, no. 3, pp. 1259–1294, 2004.

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

[12] P. C. Tetlock, "Giving content to investor sentiment: The role of media in the stock market," *The Journal of Finance*, vol. 62, no. 3, pp. 1139–1168, 2007.

[13] S. S. Groth and J. Muntermann, "An intraday market risk management approach based on textual analysis," *Decision Support Systems*, vol. 50, no. 4, pp. 680–691, 2011.

[14] M. W. Uhl, M. Pedersen, and O. Malitius, "Whats in the news? using news sentiment momentum for tactical asset allocation," *The Journal of Portfolio Management*, vol. 41, no. 2, pp. 100–112, 2015.

[15] A. K. Nassirtoussi, S. Aghabozorgi, T. Y. Wah, and D. C. L. Ngo, "Text mining for market prediction: A systematic review," *Expert Systems with Applications*, vol. 41, no. 16, pp. 7653–7670, 2014.

[16] M. Cecchini, H. Aytug, G. J. Koehler, and P. Pathak, "Making words work: Using financial text as a predictor of financial events," *Decision Support Systems*, vol. 50, no. 1, pp. 164–175, 2010.

[17] T. Adrian, N. Boyarchenko, and H. S. Shin, "The cyclicality of leverage," *FRB of New York Working Paper No. FEDNSR743*, 2015.

[18] C. D. Manning and H. Schütze, *Foundations of statistical natural language processing*. MIT Press, 1999, vol. 999.

[19] M. C. Jensen and W. H. Meckling, "Theory of the firm: Managerial behavior, agency costs and ownership structure," *Journal of financial economics*, vol. 3, no. 4, pp. 305–360, 1976.