European Journal of Operational Research 277 (2019) 770–783

Contents lists available at ScienceDirect

European Journal of Operational Research

journal homepage: www.elsevier.com/locate/ejor

Interfaces with Other Disciplines

Learning risk culture of banks using news analytics

Arvind Agarwal[a], Aparna Gupta[b,c,1,*], Arun Kumar[a], Srikanth G. Tamilselvam[a]

[a] IBM Research Lab, Delhi, India
[b] Lally School of Management, RPI, Troy, NY, USA
[c] US Securities and Exchange Commission, Washington DC, USA

ARTICLE INFO

Article history:
Received 16 December 2017
Accepted 22 February 2019
Available online 27 February 2019

Keywords:
Analytics
Banking
Financial crisis
Regulations
Risk governance

ABSTRACT

Risk culture is arguably a leading contributor to risk outcomes of a firm. We define risk culture indicators based on unstructured news data to develop a qualitative assessment of risk culture of banks. For US banks participating in an annual stress test program, we conduct a supervised learning ridge regression analysis to identify the most significant features to evaluate banks' risk culture characteristics. These features are used for unsupervised clustering to determine the high to low quality of risk culture. The distinct groups obtained from clustering define and allow monitoring changes in the quality of risk culture in banks.

1. Introduction

The 2008 global financial crisis increased concerns regarding the risk management practices of banks and their broader implications. The increasingly complex banking system poses an enormous regulatory challenge in terms of controlling banking risk and maintaining stability of the financial system. Specifically, the complexity of the banking system arises from two sources, the tighter connection among banks and the increasing size and organizational complexity of individual banks. The increasing size and complexity of banks also makes the risk of individual banks more opaque. Regulatory changes, such as the 2010 Dodd–Frank Act in the US and the US Federal Reserve's Stress Test program, were intended to improve financial stability.

The increasing opacity, however, can impede timely regulatory intervention (Gallemore, 2013), unless the regulators have improved tools to monitor and supervise banks' risks. We conjecture that to regulate the banking system well, only monitoring "hard" structured data is far from enough; unstructured textual data from various sources can provide valuable complementary insights. We explore this in this paper. Besides ensuring that banks follow the regulatory rules to control their risk exposures, supervision of the spirit of corporate governance of banks can help address the root cause of the problem. As pointed out by the Financial Stability Board (FSB), "a more intense and effective approach to oversight aims to deliver pre-emptive, rather than reactive, outcomes-based supervision" (FSB, 2014; Glazer & Rexrode, 2016). Treating operational losses as independent events, as most banks currently do, resembles treating the symptoms of a disease rather than the malaise itself (Chernobai, Jorion, & Yu, 2012; Drennan, 2004).

A common element affecting the risk taking behaviors of individuals and groups in a financial institution is its risk culture (Glazer & Rexrode, 2016). While obtaining appropriate data and measuring banks' risk culture is difficult, developing ways to measure, monitor and study risk culture is a worthwhile endeavor. Besides the quantitative financial statements reported to the regulators, there is significant textual content in reports banks are required to file with the regulators. Additionally, print news media creates large volumes of text providing information on banking, of financial and/or business interest, to general public. These qualitative and unstructured data can complement the quantitative analysis of bank risk-taking for assessing risk culture of banks.

In this paper, we utilize a large volume of print news data to identify a set of meaningful indicators for evaluating a bank's risk culture. We particularly focus on the period of Federal Reserve stress tests and consider banks that have participated in the stress test program. Following a specific risk culture framework,

* Corresponding author at: Lally School of Management, RPI, Troy, NY, USA.
E-mail addresses: arvagarw@in.ibm.com (A. Agarwal), guptaa@rpi.edu (A. Gupta), kkarun@in.ibm.com (A. Kumar), srikanth.tamilselvam@in.ibm.com (S.G. Tamilselvam).

[1] The Securities and Exchange Commission, as a matter of policy, disclaims responsibility for any private publication or statements by any of its employees. The views expressed herein are those of the author and do not necessarily reflect the views of the Commission or of the author's colleagues on the staff of the Commission; This work was supported in part by NSF Award III-1738895.

https://doi.org/10.1016/j.ejor.2019.02.045
0377-2217/

we investigate how the risk culture of this group of banks has changed over the stress testing period. We build a feasible way to extract the risk culture indicators (RCI) from these textual sources for applying the risk culture framework to the banks. Using a set of proxy target variables, we conduct a supervised learning based risk culture features selection for the banks. Thereafter, in an unsupervised assessment, we analyze the groups the banks belong to over time according to the chosen risk culture indicators. Given that the Fed's stress test program must help strengthen public confidence in the nation's banking system, evaluation of the evolution of banks' risk culture from news data assesses the effectiveness of the program in improving public confidence.

A rigorous definition of risk culture is essential to build any measure for it. The International Institute of Finance (IIF) (IIF, 2009) defines risk culture as the norms and traditions of behavior of individuals and of groups within an organization that determine the way in which they identify, understand, discuss, and act on the risks of the organization, and the risks it takes (Power, Ashby, & Palermo, 2013). Researchers emphasize that both external and internal analysis are necessary to fully evaluate a bank's risk culture (McConnell, 2013).

Risk culture literature has progressed on two themes, prescriptive and empirical. In the prescriptive thread, researchers have focused on aspects of a firm that should be reflected in the conceptualization of its risk culture, namely what is "good" risk culture and how to build it in a firm. McConnell (2013) introduced a framework with six key drivers reflecting managers' values and behavior, management system and employees' activities. Geretto and Pauluzzo (2015) emphasized the values, norms, and practices of the members of an organization that contribute to the organization's risk culture. Sheedy, Griffin, and Barbour (2015) identify four common factors for risk climate: values, managers, proactivity, and avoidance, and IIF (2009) provided four essential elements of successful risk culture. Power et al. (2013) warn of destructive pathways that the Financial Stability Board (FSB) should pay attention to when promoting risk culture (Ashby, 2014). Based on Fritz-Morgenthal et al.'s rigorous risk culture framework (Fritz-Morgenthal, Hellmuth, & Packham, 2016), our identification of risk culture will take into consideration the detailed features proposed by the studies and the considerations for soundness of risk culture. Firm culture and its broader influence on corporate decisions and relation with corporate governance has been investigated (Bae, Chang, & Kang, 2012). We are specifically interested in its influence on risk decisions.

In empirical studies of risk culture, researchers have discussed whether risk culture is measurable and reportable, what factors have an impact on risk culture, and the relation of risk culture to other quantitative and qualitative metrics. Palermo et al. (2015) explore the unique features of risk culture extractable from existing materials, such as, documents from websites of consulting firms, professional associations and rating agencies (Gherardi & Nicolini, 2000; Law & Singleton, 2005). Other researchers have designed questionnaires to collect data and calculate organizational culture score and risk culture score. Kimbrough and Componation (2009) measured the organizational culture by using the Organizational Culture Assessment (OCA) score. Based on Burns and Stalker's (Burns & Stalker, 1961) model, the OCA score contains 20 brief questions designed by Kimbrough and Componation. The OCA score is claimed to measure whether the organizational culture is mechanistic or organic. The Macquarie University Risk Culture Scale designs a questionnaires for the risk culture score using four factors, namely, whether risk management is valued in the firm, whether risk issues and events are proactively identified and addressed, whether risk issues and policy breaches are ignored, downplayed or excused, and finally, whether immediate management response is implemented (Sheedy & Griffin, 2014).

From the above summary of research on risk culture, it is evident that research on risk culture has significant challenges. Since the beginning of the Internet era, huge amounts of documents, comments and discussions are available as text from the world wide web. Taking advantage of these textual data in the finance domain is a valuable pursuit. A large fraction of work in the application of text mining in finance has been on predicting price movement of stocks and other market variables. For instance, Alfano, Feuerriegel, and Neumann (2015), Tai, Olson, and Blessner (2016), Antweiler and Frank (2004) and Wuthrich et al. (1998) use news articles to predict the stock market or FOREX market. Serrano and Iglesias (2016), Nguyen, Shirai, and Velcin (2015), Zhang, Swanson, and Prombutr (2012) and Ranco et al. (2016) focus on analyzing social media text from platforms such as Twitter, stock message boards and Yahoo!Finance message board, for implementing market prediction. Besides market prediction, text mining is also implemented for measuring financial condition, such as firm performance (Balakrishnan, Qiu, & Srinivasan, 2010; Tsai & Wang, 2017), credit rating prediction (Mengelkamp, Hobert, & Schumann, 2015; Tsai, Lu, & Hung, 2010), bank distress prediction (Iturriaga & Sanz, 2015; Rönnqvist & Sarlin, 2015a; 2015b; 2017) and systemic risk measurement (Lischinsky, 2011; Nyman et al., 2015; Tai et al., 2016).

Application of sentiment analysis and text mining in finance constitutes a rapidly growing literature (Liu, 2012; Mäntylä, Graziotin, & Kuutila, 2018; Nassirtoussi, Aghabozorgi, Wah, & Ngo, 2014), and essentially involves a few key steps. Feature extraction methods determine how researchers collect useful qualitative information from textual data. The most popular method of feature extraction is the bag-of-words approach (Alfano et al., 2015; Feuerriegel, Wolff, & Neumann, 2015; Nguyen et al., 2015). This technique breaks the text into word-level units, and treats these units as features, while ignoring the order and co-occurrence of the words (Nassirtoussi et al., 2014). Effort is then made to assign sentiment to these extracted words, often by taking support of extensive domain-specific dictionaries developed to assign sentiments (Bodnaruk, Loughran, & McDonald, 2015; Loughran & McDonald, 2011a; 2014). However, assigning sentiments based solely on presence of certain words has limitations, as this doesn't account of the context of the words, including presence of negations that may completely reverse the meaning. Beyond bag-of-words methods, more advanced supervised and unsupervised learning techniques are used to extract textual sentiments (Liu, 2012). For feature extraction, Schumaker, Zhang, Huang, and Chen (2012) apply a noun-phrase technique, in which they identify the words with a noun part-of-speech (POS) by using a lexicon and then apply syntactic rules to detect noun phrases around that noun to extract features. Tsai, Wang, and Chien (2016) have applied a continuous bag-of-words approach, a continuous-space language model, to discover finance words from firms' annual reports to predict stock volatility, abnormal trading volume, etc. Researchers have also implemented named entity recognition techniques to improve the feature extraction results (Vu, Chang, Ha, & Collier, 2012). Rekabsaz et al. (2017) use word embedding-based information retrieval models for sentiment analysis of firm disclosures for forecasting stock volatility, along with combining with quantitative market signals.

After feature extraction from a corpus, text mining objective usually requires feature selection and application of a classification method to capture the required signal from the text. Various machine learning algorithms are applied to analyze the features extracted; one common method used is the Support Vector Machine (SVM) (Chiang et al., 2015; Nassirtoussi, Aghabozorgi, Wah, & Ngo, 2015; Nguyen et al., 2015). SVM is a non-probabilistic binary linear classifier that finds a hyperplane that separates two classes. Linear regression models are also used as a supervised learning

technique (Chatrath, Miao, Ramchander, & Villupuram, 2014). Additionally, Naive Bayes (Li, 2010) and Decision Rules or Trees (Rachlin, Last, Alberg, & Kandel, 2007) methods are also frequently used.

These text mining methods build the basis of the methodology used in this paper for the purpose of risk culture identification of a bank. The textual data is obtained from a large corpus of news articles in order to develop an objective, text analytics based risk culture assessment framework. We define features based on sentiment analysis applied to news articles that discuss topics related to a bank's risk culture. The sentiment assessment allows assigning a quality indication on each of the risk culture indicators. As we don't expect Federal Reserve Bank's annual stress testing scores to be perfect predictor of banks' risk culture, we use these scores as target variables to guide the identification of the most significant features. To further learn the banks' risk culture, we use the most significant features in an unsupervised clustering analysis to identify banks that group together by similar feature characteristics. We observe that quality of bank's portfolio and reputation emerge as the most important risk culture indicators, followed by bank's strategy and employee characteristics.

Considering yearly risk culture features extracted from each year's news articles corresponding to the banks participating in the Fed's stress test program allows us to examine how the risk culture changed in the banks through the years of stress testing. Even though the Dodd–Frank stress tests are a regulatory initiative, strategy, employee characteristics and reputation emerge as more important textual features than regulatory requirements, when using the stress test scores for supervised learning. The banks fall in clearly defined three clusters for their quality of risk culture, from high, medium to low quality. Although a large group of banks initially fall in the low quality risk culture group, they improve their status over the years. There are a modest number of banks in the medium and high quality groups. Matching features by stress testing years allows examining risk culture group transitions. We observe several transitions, some displaying improvement in risk culture, while others showing a deterioration. Therefore, these differentiated classes of banks may indicate a risk culture score for banks above and beyond the stress test results.

Rest of the paper is organized as follows. In the next section, we present the risk culture framework used in this paper, along with describing the data sources and text extraction methods used. Section 3 describes the supervised and unsupervised machine learning methods used to identify the most significant features for risk culture identification and clustering banks by their risk culture characteristics. Section 4 presents the results and discusses our findings. Final remarks are provided in the conclusions section.

2. Risk culture framework

Among earliest definitions of risk culture, Bozeman and Kingsley (1998) defined risk culture as "the organization's propensity to take risks as perceived by the managers in the organization," which emphasizes the risk appetite of managers. In more recent years, the concept of risk culture has been discussed in context of banks (Geretto & Pauluzzo, 2015), financial regulators (Gallemore, 2013), consultancy firms and insurers, however, it is considered hard to get a universally agreed upon definition of an organization's culture, and specifically, its risk culture. In this work, we closely follow the risk culture framework utilized in Fritz-Morgenthal et al. (2016). We use the framework for identification of risk culture indicators, construction of a risk culture dictionary and develop a risk culture measure using unstructured data.

The seven dimensional construct of the risk culture framework constitutes: (1) governance, (2) portfolio, (3) risk strategy, (4) employees, (5) regulatory requirement, (6) work culture, and (7) reputation. Governance quality dictates the tone set by the top management and firm-wide processes set in place to relay the tone through the firm. Portfolio refers to how the balance sheet reflects the firm's attitude and strategy towards risk appetite, risk exposure and management. Risk strategy emphasizes the governance in place for risk decisions in the firm. The risk strategy must be consistent with the regulatory requirements imposed on the bank.

No risk outcomes can be robust if the employees are not adequately prepared and trained, which is the next feature in the risk culture framework. In this regard, employee attrition or retention is also considered important. Employee experience, behavior and incentives at all levels of organizational hierarchy of the firm define the work culture, which specifically relates to the risk culture of the firm. Finally, governance, risk strategy, adequate response to regulatory requirements, employee competency and work culture feed into creating the reputation of the firm. Therefore, these seven dimensions holistically cover the characteristics of a firm that contribute to the firm's risk culture. We will construct a dictionary and indicators by these dimensions in the next section.

Table 1
Description of the Risk Culture Framework in terms of its seven Risk Culture Indicators (RCIs).

Category	Description
Governance	Qualifies if appropriate senior management to operate the business and an adequate supervisory authority to govern the bank are in place.
Portfolio	Evaluates selected balance sheet related figures considered as relevant indicators for the quality of a bankâs risk culture.
Risk strategy	Emphasizes appropriate risk governance, processes and personnel being in place and how the different risks are managed as relevant to a specific bank.
Regulatory requirement	Level of compliance with regulatory requirements related to risk management.
Employees	Measures average training hours completed by employees and employee retention.
Work Culture	Behavioral indicators and attitudes identifiable.
Reputation	Banks' statements made regarding their reputation and related threats or risks, where litigations and their transparent disclosure is considerd.

2.1. Risk culture indicators & dictionary

To estimate the quality of risk culture, Fritz-Morgenthal et al. (2016) defined a number of risk culture indicators (RCIs) and discussed their relationship with risk culture. We leverage these RCIs, namely regulatory requirement, governance, portfolio, employees, risk strategy, reputation and work culture for our systematic regression analysis for risk culture identification of banks. A brief description of these risk culture indicators is provided in Table 1. Based on the additional discussion of risk culture in the literature (Bowman, 1984; Mihet, 2013; Power et al., 2013), we identify a set of words that represent each of the risk culture indicators. We use these RCI keywords as seed words to expand the RCI dictionary based on their synonyms. We identify co-occurring words of these seed words, such as, verbs like 'demand', 'report' etc., from the news articles, and based on inputs from a subject matter expert filter these co-occurring words for inclusion in the dictionary. This process helps finalize the complete set of keywords for the risk culture indicators. We refer to these keywords as the dictionary for the risk culture indicators. A sample of these keywords are summarized in the Appendix.

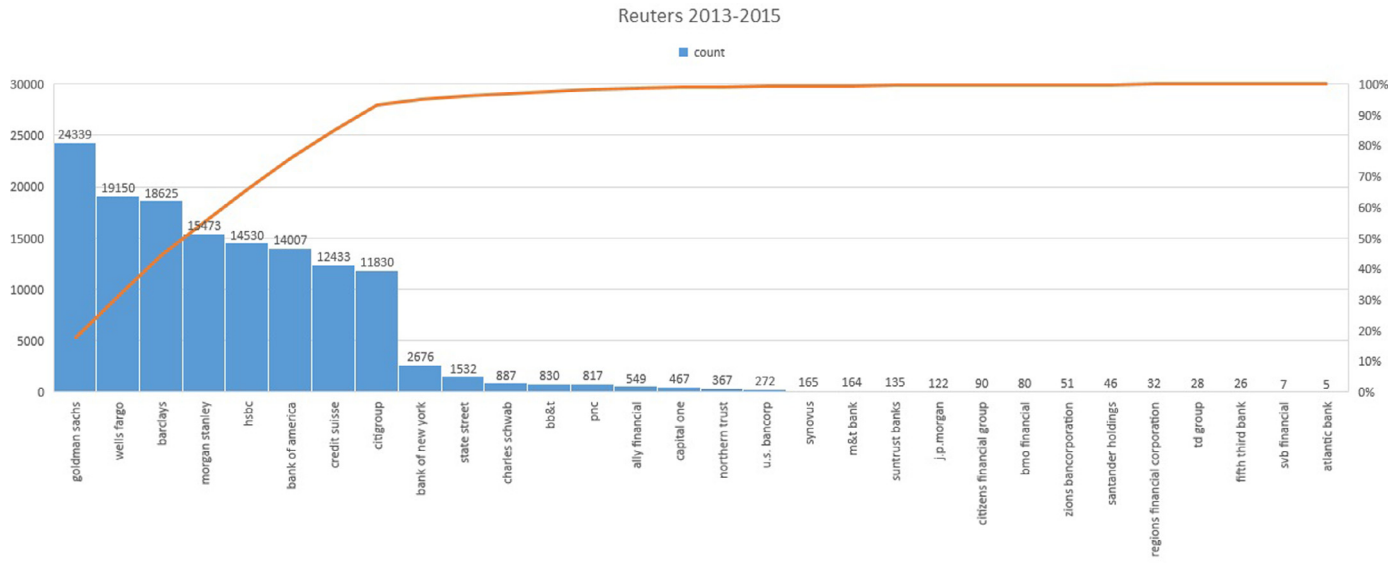

Fig. 1. Pareto chart of the filtered article distribution for each of the bank in the years 2013–2016. Only articles that mention any of the RCIs keywords are considered.

Table 2
Unique articles for banks considered discussing at least one of the RCIs.

Year	Unique articles
2013	40506
2014	29115
2015	29616

Table 3
Table discussing articles discussing upto 4 banks and the frequency of the appearing in the article and in the same sentence.

Num banks	Appear in the same article	Appear in the same sentence
1	42977	42977
2	21633	10276
3	9258	4122
4	4248	1708

2.2. Dataset and data sources

In this work, we aim to examine the risk culture indication obtainable from publicly available news resources for the financial institutions participating in the Fed's annual stress tests. We utilize print news data obtained from Reuters News Archive[2] for this task. News articles from this resource in general discuss a wide range of topics, from politics, banking, sports, entertainment, etc. For our focus on financial institutions and their risk culture, we restrict using news data that identifies with banking related articles. We filter articles by whether the articles mention at least one of the banks included in our study, which are the banks that have undergone the Federal Reserve's annual stress tests in the past years. A complete list of banks is provided in the Appendix[3]. Table 2 shows the number of unique articles obtained for the years 2013–2015 that mention at least one of the banks in our study group and a risk culture indicator. This is identified by checking if the articles contain any word(s) from the RCIs dictionary discussed in Section 2.1. Federal Reserve Bank also makes annual stress test reports and test scores available for each of these banks. These stress test scores are utilized for supervised learning of the significant features for a risk culture assessment.

2.3. Sentiment analysis tools

Occurrence of risk culture indicator keywords in news articles is not sufficient for assessing the risk culture. We additionally need to determine the sentiment associated with these occurrences in order to judge the quality of reference to the risk culture discussion. Therefore, in each article, we seek out particular sentences that contain one of the risk culture indicator keywords, sample of which are listed in the Appendix. A sentence containing a keyword associated with a risk culture indicator is considered as a representative for that indicator and a sentiment analysis is then performed on the sentence towards that risk culture indicator. In Fig. 1, we plot the number of articles containing risk culture indicator keywords for all the banks. We note that the total number of articles for the three years in this plot is much higher than the total reported in Table 2. This is because Table 2 reports only unique articles that mention at least one of the banks along with at least one of the RCIs, however Fig. 1 can count an article as many times as the number of banks it mentions. A single article can contain information about more than one bank, therefore the article would get counted in the tally of multiple banks. We also observe that most articles, as shown in Table 3, tend to discuss banks in the same sentence. We generally expect a single sentence to convey a single sentiment, which we attribute to the risk culture indicator for all the banks that are mentioned in the sentence. For the challenge of co-reference resolution of banks appearing in different sentences, we consider only those articles in which mention of all the banks appears in the same sentence. This provides higher confidence for sentiment attribution for multiple banks mentioned in an article and their risk culture indication.

Sentence segmentation refers to the process of splitting a given paragraph of text into sentences, by identifying the sentence boundaries. In our case, a period punctuation is used to

[2] Reuters News: http://www.reuters.com/resources/archive/us/.
[3] Fed stress list: https://www.raalreserve.gov/bankinforeg/dfa-stress-tests.htm.

Table 4
Results comparing sentiments computed using IBM Alchemy and Loughran–McDonald methods. Column 2 is the total number of sentences that discuss the indicator. Column 3 is the number of sentences where both methods agree on the sentiments (exact match) assigned. Column 4 captures percentage of exact matches. Column 5 provides information on approximate match, where neutral and positive are merged into one sentiment or neutral and negative are merged, and then only positive and negative sentiments are considered for identifying mismatches. Column 6 provides the approximate match percentage.

Category	# Sentences	# Exact match	% Exact match	# Approx. match	% Approx. match
Regulatory req	272430	143467	53%	253895	93%
Governance	137112	60968	44%	122942	90%
Portfolio	173465	103452	60%	163612	94%
Employees	60466	28679	47%	55258	91%
Risk strategy	156156	70494	45%	140683	90%
Reputation	149812	71035	47%	134388	90%
Work culture	19703	8415	43%	17192	87%

identify the end of a sentence. Abbreviations, such as 'U.S.', are first stripped off their periods before sentence segmentation is applied. For assigning a sentiment to every sentence associated with a risk culture indicator, among the different sentiment analysis tools available, we chose to use IBM Alchemy[4] for its industry-wide recognition and Loughran–McDonald Dictionary (Loughran & McDonald, 2011b) for its popularity among finance researchers. IBM Alchemy uses advanced machine learning algorithms to learn linguistic attributes of news articles in order to assign sentiments to them. Loughran–McDonald dictionaries consist of negative and positive sentiment words (among others) that have been developed with a focus on finance specific usage. If a sentence contains more than one sentiment word, a majority voting scheme is used to arrive at a single sentiment for the sentence. A positive sentiment is assigned a value of '+1,' a negative sentiment a value '−1,' and a neutral is assigned a value '0.' Unlike dictionary based methods, Alchemy provides a single sentiment value, i.e positive/neutral/negative, for a sentence along with a confidence score.

We performed sentiment analysis using both the methods and report a comparison of results at the sentence level in Table 4. A strict agreement of negative/neutral/positive sentiments between the two methods ranges from 43% to 60% for all the risk culture indicators. However, if the neutral assignments are merged with positive or negative sentiments, the match agreement becomes uniformly above 87% (last column of Table 4). Therefore, there is significant agreement between the two sentiment assignment methods, especially strictly by positive/negative sentiments.

A closer examination of the mis-matched cases is worthwhile. Towards this end, we conducted two experiments where we employed subject matter experts to read the mis-matched sentences for a sentiment analysis and compared the outcomes with the assignments by the two tools. In the first experiment, we took 100 sentences which were labeled as positive by Loughran–McDonald sentiment assignment but negative by Alchemy tools. In the second experiment, we considered 100 sentences that were labeled negative by Loughran–McDonald assignment but positive by Alchemy. In the first case, we found Alchemy labels agreed with the human judgment in 54% of the 100 sentences and in the second case, the agreement was 64%. With Alchemy being consistently closer to human evaluation, we chose Alchemy based sentiment assignment for our study.

We observe that Alchemy is able to take the context of the keywords into consideration rather than attributing the sentiment of the entire sentence on a few words. However, IBM Alchemy has some drawbacks also. A couple of sentences with contrasting sentiments from the two methods are presented below.

> *Ukraine's debt restructuring, pro-business Mauricio Macri's election win in Argentina and hopes Venezuela will see something similar on Sunday, have seen a dramatic turnaround in investors' attitude toward all three countries.*

This sentence discusses the 'portfolio' RCI associated with the keyword "debt." IBM Alchemy identified this sentence, we believe correctly, as bearing a positive sentiment, while dictionary based method identified it as carrying a negative sentiment because of the word "restructure." An example where IBM Alchemy doesn't perform as well follows.

> *Cash equities experienced an aggressive ramp at the close of Tuesday's session, pushing the Nikkei above 20,000 points, but market participants said macro sentiment and risk appetite were cooled a little overnight by a slight strengthening of the yen against the dollar.*

This sentence discusses the 'strategy' RCI associated with the keyword "risk appetite." IBM Alchemy identified this sentence with a negative sentiment, while dictionary based method identified the sentence to have a positive sentiment due to the word "strengthen." For the uncertainty expressed in the sentence, it is not surprising that the two approaches pick a theme to conclude on the sentence's sentiment.

Using the features defined by the risk culture indicators qualified by their corresponding sentiments, we next present the risk culture assessment methodologies.

3. Methodology for risk culture assessment

A bank's performance in the Fed's stress tests can be a strong indicator of the bank's risk culture, however it can not be conclusively reliable for the bank's risk culture identification. Therefore, our approach to determining the risk culture of banks is based on first applying a supervised regression model assessment of the risk culture indicator sentiment features. This is followed by an unsupervised clustering of the banks by the important features selected from the supervised learning. The supervised regression models use training data consisting of RCI features extracted from textual data along with target values, which are the bank's Fed stress test scores. This analysis guides us to identify significant textual features extracted from news data that strongly align with bank's stress test performance. As discussed in the previous section, we use publicly available Reuters News Archives data for defining the RCI feature space.

A sentence level extraction from the relevant filtered articles are processed to assign a sentiment for each risk culture indicator. The features thus defined combine the sentiment for each of the RCIs. As discussed in Section 2, we use 7 RCIs, and for each RCI, we associate three sentiment values, i.e. positive, negative and neutral. This gives us a final feature space containing 21 variables. In order to compute these feature values, we first identify the subset of articles associated with each bank, where an article may be associated with one or more banks. The filtered articles are then processed sentence by sentence, associating a sentence with an RCI based on the presence of RCI keywords in the sentence. For example, if a sentence in an article contains a keyword *'Layoff'*, we associate this article with the RCI *'Employees*.'

We determine the sentiment of the sentences and combine them with the RCIs associated with the sentences. All the sentiment values for all RCIs and all articles associated with a bank or bank-year are aggregated to compute the feature vector. These features from the training data act as predictor variables to learn the supervised regression models. The two chosen target variables for

[4] https://www.ibm.com/watson/developercloud/alchemy-language.html.

supervised learning are the metrics available from the Fed's stress test results, namely *Tier 1 capital ratio Ending* and *Tier 1 leverage ratio Ending*. These stress test metrics are provided for two scenarios, the *adverse scenario*[5] and the *severely adverse scenario*[6]. Two metrics each for two adverse scenarios makes up for a total of 4 target variables in the supervised learning.

There are several possible models available in machine learning literature for supervised learning. For our 21 RCI feature variables and 4 target variables, we sought to find the best learning representation from among a range of model choices. We evaluated the following supervised learning regression model approaches: Ridge regression, Support Vector Ridge regression, Ridge Lasso, and Random Forest regression. These different regression models were chosen to cover models with different complexities, such as, linear models (Ridge regression), non-linear models (Support Vector Ridge regression), models with inherent feature selection (Ridge Lasso), ensemble methods (Random Forest). Our evaluation of these models was conducted for different parametric settings of the model. The best performing approach emerged to be the Ridge regression model in almost all cases of RCI feature and target variables.

3.1. Supervised regression models

In this section, we provide the details of the top three performing regression models, namely, Ridge Regression, Lasso and Random Forest.

Ridge Regression is one of the most popular regression methods. It is a linear regression method to model the relationship between a scalar response, i.e. a target variable, and the given explanatory variables. The goal of the model is to minimize the difference between the predicted values of the scalar variable and the given ground truth value. A simple linear regression method has a tendency to over-fit the data by making the weights of the explanatory variables arbitrarily large, resulting in their poor generalized performance, i.e. poor performance on the test data. In order to avoid this behavior, regularization techniques are applied to reduce over-fitting of the model to the training data. One of the most popular regularization methods is to add a penalty term based on the ℓ_2 norm of the weight vector. One can control the amount of penalty through a hyperparameter. Mathematically, let $X = \{x_1, x_2, \ldots, x_n\}$ denote the training data of size n, where each x_i is a d-dimensional vector containing values of d predictors. In our setting, the d-predictors correspond to the 21 RCI features. Let $Y = \{y_1, y_2, \ldots, y_n\}$ be the corresponding target values, where y_i is one of the 4 target values for each x_i. The objective function of the Ridge regression is as follows:

$$\min_w \|Y - Xw\|_2^2 + \alpha\|w\|_2^2, \tag{1}$$

where, w is the weight vector that will be learned during the training phase. α controls the penalty associated with the ℓ_2 regularization.

Lasso builds on the linear regression model, except that instead of the ℓ_2 regularization, it uses ℓ_1 regularization. The ℓ_1 regularization has the inherent capability of selecting features by forcing feature weights to be zero which otherwise will have a close-to-zero value. Mathematically the Lasso minimizes the following objective function:

$$\min_w \|Y - Xw\|_2^2 + \lambda\|w\|_1, \tag{2}$$

where similar to the ridge regression, λ controls the amount of regularization. In our study, we experiment with several values of λ and choose the one that performs the best.

In both Lasso and Ridge regression, one can use the weights of the explanatory variables, i.e. w, to determine the importance of the corresponding features. Since the weights are implicitly affected by the scale of the variables, all variables are normalized to remove the scaling. We use the standard normalization such that all variables have zero mean and unit variance. Furthermore, We use the absolute value of the weights associated with each feature to determine its relative importance.

Unlike Ridge regression and Lasso, which are a single model, Random Forest is an ensemble method. In ensemble learning, the central theme is to combine multiple simple models to build a single powerful model. In this spirit, a random forest regression model is a collection of decision tree regressors, where each decision tree regressor is built on a sample of data, chosen at random (but with replacement) from the original data.

The decision trees, which are the building block of a random forest model, are built by splitting the training dataset into subsets based on some criteria for the quality of split. There are many metrics available to measure the quality of a split. A standard metric used for random forest is the Mean Squared Error (MSE) measure. MSE measures the variance reduction in the data due to splitting, and is defined as:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2. \tag{3}$$

Therefore, intuitively, it measures the goodness of fit for each feature. In order to build a decision tree, at each node, the quality of split is computed for each feature. The feature with the best quality of split is chosen for defining the split, and is removed from consideration for subsequent splits in the tree. This process of splitting is recursively repeated until a stopping criterion is met. One frequently used stopping criterion is the minimum number of samples in the leaf nodes. We also use MSE measure as our evaluation metric to compute the quality of the final regression model. While MSE provides a measure of goodness of fit for the overall model, it does not provide guidance on feature importance. However, since a random forest is a collection of regression trees operating on features one-by-one, feature importance is obtained as a by-product of the model building process. Individual regression trees intrinsically perform feature selection by selecting appropriate split points. This information can be used to measure the importance of each feature. The features that are selected first bear greater importance than those that are selected later in the model building process. Since a random forest is a collection of decision trees, this notion of feature importance can be extended to the decision tree ensemble by simply averaging the feature importance of each tree.

[5] In 2016 Fed Report, 'The adverse scenario is characterized by weakening economic activity across all countries or country blocs included in the scenario. The economic downturn is accompanied by a period of deflation in the United States and in the other countries and country blocs. The adverse scenario features a moderate U.S. recession that begins in the first quarter of 2016. Real GDP in the United States falls 1.75 percent from the pre-recession peak in the fourth quarter of 2015 to the recession trough in the first quarter of 2017, while the unemployment rate rises steadily, peaking at 7.50 percent in the middle of 2017. The U.S. recession is accompanied by a mild deflationary period with consumer prices falling about 0.50 percent over the four quarters of 2016. Reflecting weak economic conditions and deflationary pressures, short-term interest rates in the United States remain near zero over the projection period. The 10-year Treasury yield declines to 1.25 percent in early 2016 before rising gradually thereafter to 3 percent in the first quarter of 2019.'

[6] In 2016 Report, 'The severely adverse scenario is characterized by a severe global recession accompanied by a period of heightened corporate financial stress and negative yields for short-term U.S. Treasury securities. In this scenario, the level of U.S. real GDP begins to decline in the first quarter of 2016 and reaches a trough in the first quarter of 2017 that is 6.25 percent below the pre-recession peak. The unemployment rate increases by 5 percentage points, to 10 percent, by the middle of 2017, and headline consumer price inflation rises from about 0.25 percent at an annual rate in the first quarter of 2016 to about 1.25 percent at an annual rate by the end of the recession.'

For real valued target variables, y_i, the random forest algorithm consists of two main steps as follows.

1. For $t = 1, \ldots, T$, create a dataset X_t by sampling with replacement from original dataset X. Here T is a hyper-parameter denoting the number of estimators we want the random forest model to have, which determines the number of regression trees in the ensemble.
2. Build a regression tree, f_t, on X_t, $\forall t$. This generates a collection of regression trees, $f_1, \ldots, f_T$.

Once regression models (Ridge, Lasso or Random Forest) are built corresponding to the 4 target variables using the training data, the models must be evaluated on the test data. Given the test data $D_t = \{(x_1, y_1), (x_2, y_2), \ldots (x_m, y_m)\}$, where m is number of data points in the test set, the MSE of D_t is computed using Eq (3). For the test data point, $\hat{x}$, the prediction value is computed using the corresponding function. For the Ridge regression and the Lasso, it is computed as follows:

$$\hat{y}_i = f(\hat{x}_i) = w^T \hat{x}_i, \tag{4}$$

whereas for the random forest model, it is computed by taking an average of all predictions, i.e.,

$$\hat{y}_i = f(\hat{x}_i) = \frac{1}{T} \sum_{t=1}^{T} f_t(\hat{x}_i). \tag{5}$$

As discussed earlier, in our dataset we have 4 target variables that result in 4 different regression models. Each of these models gives us a ranked list of features ordered according to their importance. In order to get a unique ranked list of features, these ranked list can be aggregated using the importance scores, i.e., absolute weight for Ridge and Lasso and the value of the split in Random Forest. The scores are normalized before aggregation to avoid the bias due to scaling.

3.1.1. Specific considerations for regression modeling

The experimental setting used in this paper is built on training data set constructed from news articles obtained from Thomson Reuters and Fed's stress test results. For Thomson Reuters, although we have multiple years of data available, ground truth stress test target variables are available only for 4 years, starting from year 2013 to 2016. In the year 2013, stress test results were available for only 18 banks, and only for the *severely adverse* scenario. In other years, the stress test results are available for 20–30 banks, and for both *adverse* and *severely adverse* scenarios. Therefore, we only considered 3 years of news articles based feature variables data, from 2013–2015. This is because we match year t features data with $t + 1$ stress test results, as the stress tests in year $t + 1$ are applied to Q4 condition of the bank in year t. Therefore, the matched year of articles to the stress test results is the one prior to the year of stress test results. This gives us a a total number of 69 bank-years however from these bank-years we filtered all the bank-years that had less than 200 articles across all features. This resulted in 45 bank-years which are used for both training and test data set creation.

In our experimental setting, we normalize the data across both rows (i.e. features) and columns (across banks). For various reasons, including bank size and popularity, a bank can be mentioned in a large number of articles in the media. Therefore, it is important that we normalize the data. This is done by normalizing each bank across all its features. Similarly, we normalize the data for each feature across banks. We use standard normalization to have zero mean and unit variance after normalization.

To overcome the challenge of small data set available for learning, we used cross validation for the evaluation of regression models with 20 folds. We run each model for different parameter settings and report the best results, averaged over all folds. In Table 5, we show the results of different models with different parameter settings. We experiment with 3 kind of models i.e., Ridge Regression, Lasso and Random Forest. For Ridge and Lasso, we experiment with different values of the regularization parameter, i.e. α, $\lambda \in \{0.01, 0.1, 0.5, 1, 10, 100\}$, and for the Random Forest, we experiment with different numbers of trees (i.e. estimators) in the Forest, i.e. $T \in \{10, 50, 100, 1000\}$. All implementations are done using scikit-learn in Python. All the parameter values are set at their default values except for the ones mentioned above. In the unsupervised learning analysis discussed next, we use the results from the Ridge regression model due to its consistent performance across all 4 target variable scenarios.

Table 5
Results of different regression models with different parameters setting across 4 target variables. For each target variable, only top performing models are reported.

Target variable	Regression model	Parameter	Param value	MSE
Tier 1 capital ratio	Ridge	α	100	2.95
Ending Adverse	Lasso	λ	0.5	2.97
	Lasso	λ	1	3
	Lasso	λ	10	3
	Lasso	λ	100	3
Tier 1 capital ratio	Ridge	α	100	3.91
Ending Severely	Lasso	λ	0.5	4.11
Adverse	Random Forest	T	1000	4.21
	Random Forest	T	100	4.26
	Lasso	λ	10	4.29
Tier 1 leverage	Ridge	α	10	1.72
ratio Ending	Lasso	λ	0.1	1.85
Adverse	Ridge	α	1	1.86
	Ridge	α	100	1.91
	Random Forest	T	1000	1.91
Tier 1 leverage	Random Forest	T	1000	1.3
ratio Ending	Ridge	α	100	1.31
Severely Adverse	Lasso	λ	0.1	1.32
	Random Forest	T	50	1.32

3.2. Risk culture analysis through clustering

The ridge regression model provides, most importantly, the RCI features that emerge as significant and important in assessing the bank's quality of stress test results. A bank's ability to pass the stress tests in flying colors, in both adverse scenarios and by both performance metrics, is a strong indication that the bank possesses strong risk culture. However, it is not a perfect indicator of the bank's risk culture. Therefore, guided by the significant and important features obtained in the supervised learning, we explore an unsupervised differentiation between the banks in terms of the selected important feature variables. These differentiated classes of banks may indicate a risk culture score for banks above and beyond the stress test results.

Ridge regression model provided a ranking for feature importance. We use the aggregated ranked list based on the important score of individual target variables, and use the most important features from that analysis to conduct an unsupervised clustering analysis for further insight on banks' risk culture. Among the many different clustering methods available, we use the k-means clustering approach. The k-means clustering method is arguably one of the most popular and widely used methods for clustering. Much like any other clustering method, it is an unsupervised method to partition the data in a predefined number of k clusters. The d-dimensional vector of observed data is divided into k clusters with each data point belonging to the cluster with the nearest mean. The means serve as a cluster representative and sometimes are also referred to as data representatives. Mathematically, if we are given n data points $x_1, x_2, \ldots, x_n$, where each $x_i \in \mathbf{R}^d$ is a d-dimensional vector, k-means clustering attempts to find k clusters

$C_1, C_2, \ldots, C_k$ by solving the following optimization problem:

$$\arg \min_{C_1, C_2, \ldots, C_k} \sum_{i=1}^{k} \sum_{x \in C_i} ||x - \mu_i||^2, \quad (6)$$

where μ_i is the mean of cluster C_i. The objective function of k-means clustering is non-convex, implying that it may not have a unique minimum (solution). The solution is obtained often by running the algorithm multiple times with different initializations, and the best solution from these multiple runs is chosen to be the final solution. The k-means clustering algorithm is quite simple and works by the following steps: (1) initialize k cluster centers randomly; (2) assign each data point to one of the nearest clusters based on the distance between the data point and cluster mean; (3) re-calculate the cluster center for each cluster, by taking the mean of all the points belonging to that cluster. Steps (2) and (3) are repeated until convergence is achieved.

In our implementation of clustering, we used 13 most important features from supervised learning to construct the data vectors for unsupervised learning and to generate the clusters. Note that the data are normalized across both rows and columns, as in the supervised learning models. In k-means clustering, the number of clusters k is an input parameter to the algorithm, and it is usually difficult to determine the best value of k in advance. Therefore, in our experiments, we tried different values of k. As such, an increasing value of k would result in more cohesive clusters, however after a point there is diminishing value in increasing k for obtaining meaningful clusters. Based on our analysis of different choices of k, we selected $k = 3$ as a trade-off between degree of cluster cohesion and a meaningful labeling of clusters.

In order to obtain further insight into the clusters, further analysis is done through visualization. An original 13 dimensional data space is harder to visualize, as it cannot be adequately displayed in 2 or 3 dimensional space. For this purpose, we use principal component analysis (PCA), one of widely used methods for dimension reduction, for reducing the dimension of the data for ease of visualization. PCA works by projecting the data into a lower dimensional space such that the variance of the original data is preserved as much as possible. In our experiments, we were able to preserve approximately 60% variance of the original 13-dimensional data.

We generated clusters for all three years, i.e. 2013, 2014 and 2015 independently, as well as all the three bank-years' data combined together for each bank. This allows us to examine both temporal risk culture groups, as well as average risk culture group each bank belongs to for our study period. Since clustering methods only provide a partitioning of the data, while the actual identity of those partitions must be assigned by a subject matter expert. In order to aid the analysis and be consistent among different clusterings, we similarly label different partitions across the years to allow comparison between the clusters. The implementation is done using the scikit-learn package available in python for both k-means clustering and PCA-based dimension reduction.

4. Results and discussion

We follow the risk culture indicators (RCIs) utilized by Fritz-Morgenthal et al. (2016) to define the features for supervised learning of risk culture. The seven risk culture indicators are: regulatory requirements, governance, portfolio, employees, risk strategy, reputation and work culture. A brief overview of the risk culture framework based on these indicators was provided in Section 2. A dictionary of words is created for each risk culture indicator, where regulatory words are designed to capture regulatory requirements with regard to risk management. Governance words identify senior management's role in business and supervisory authority. Portfolio refers to selected balance sheet related figures seen relevant to the quality of a bank's risk culture. Employees features highlight the training preparedness of the bank's employees and employee retention, while strategy is specific to appropriate risk governance and processes in place for different risks of a specific bank. Reputation identifies indications for bank's reputation, litigations and their transparent disclosure. Finally, work culture identifies the principles and behavior of an organization and its employees.

In order to identify the sentiments corresponding to each risk culture indicator, a natural language processing based sentiment identifier, Alchemy, is found to have better performance compared to Loughran–McDonald positive-negative sentiment dictionaries Loughran and McDonald (2011b). Therefore, we perform our experiments using the Alchemy sentiment assignment method. As explained in Section 3, this sentiment analysis gives us a 21 dimensional feature vector for each bank-year. Summary statistics for all the bank-years of data for each feature is shown in Table 6.

As seen in the table, portfolio and work culture are the only two feature categories that are not covered in the corpus for at least one bank-year. All the other risk culture indicators have at least some discussion in the corpus for all banks, as seen by the minima in summary statistics. Regulatory requirement, reputation and portfolio make the categories that get the maximum coverage in the news articles. Given the frequency of occurrence of these features is rather skewed towards the few largest banks, medians for the feature occurrence are more instructive than the means. Median for the work culture features are the smallest, with all the other features showing good levels of occurrence in the corpus.

As such, all the 21 features can contribute and indicate the risk culture of a bank, however it is arguable that some features are more instructive than others for assessing the risk culture of a bank. We use the bank's annual stress test results to guide feature reduction for risk culture identification. Table 7 shows the summary statistics for four key ratios, Tier 1 Capital Ratio and Tier 1 Leverage Ratio, under the adverse and the severely adverse scenarios for the sample of banks. Tier 1 capital ratio is Tier 1 capital (shareholders' equity and retained earnings) divided by total risk-weighted assets of the bank. Tier 1 leverage ratio is Tier 1 capital divided by a bank's total exposures including consolidated assets, derivative exposure and certain off-balance sheet exposures. Tier 1 Leverage Ratio is uniformly lower in all summary statistics than Tier 1 Capital Ratio, with notably the standard deviation of Tier 1 Capital Ratio also being higher. All severely adverse statistics are lower than the corresponding adverse scenario ones, except the standard deviation of Tier 1 capital ratio under severely adverse scenario is higher than the standard deviation of Tier 1 capital ratio in the adverse scenario. We use these ratios for each bank under adverse and severely adverse scenarios of Fed's stress tests as the target variables for supervised learning using Ridge, Lasso and Random Forest regression models. The assumption is how banks fair in these adverse scenarios in terms of these key ratios is indicative of their risk culture.

As the best representation of the target variables and the predictor features for risk culture is not known *a priori*, we examine the representation using different regression model approaches. Among the different approaches and their corresponding parametric choices considered and discussed in Section 3, the Ridge regression approach emerged as the best representation. Table 8 shows the target variables, the sentiment analysis method used in each model, the corresponding mean squared error (MSE) and the most significant 4 variables of the model.

The model fits are in general better for the Tier 1 Leverage Ratio target variable case. The model fit is marginally worse in the severely adverse scenario for Tier 1 Capital Ratio, but better for severely adverse Tier 1 Leverage Ratio. Examining the four most significant variables in each of these models, the portfolio feature stands out as the most common important variable. Reputation

Table 6
Summary Statistics for the Features defined using the Alchemy sentiment attribution tool. Banks with less than 200 articles across all features were removed for the above statistics, and these trimmed data were used in the supervised and unsupervised learning.

	regulatory req positive	regulatory req negative	regulatory req neutral	governance positive	governance negative	governance neutral	strategy positive
Mean	311.02	1042.18	1429.24	266.22	385.27	719.53	270.18
Median	74	179	353	67	86	243	34
Std Dev	378.77	1340.29	1714.67	314.68	465.69	811.72	347.24
Min	3	18	23	1	2	17	2
Max	1243	6073	5359	1195	1715	2861	1133
	strategy negative	strategy neutral	reputation positive	reputation negative	reputation neutral	work culture positive	work culture negative
Mean	410.33	810.84	210.31	687.29	651.13	73.24	39.4
Median	90	100	51	139	170	12	5
Std Dev	511.72	1023.72	251.75	1047.59	766.15	96.06	54.91
Min	7	14	2	8	4	0	0
Max	1754	3422	829	6011	2631	396	245
	work culture neutral	employee positive	employee negative	employee neutral	portfolio positive	portfolio negative	portfolio neutral
Mean	79.22	120.31	181.38	256.11	146.44	918.71	584.09
Median	15	34	32	56	32	170	121
Std Dev	98.46	148.78	228.49	331.9	180.53	1085.98	667.53
Min	0	1	2	4	0	12	5
Max	360	545	919	1259	726	4288	2372

Table 7
Summary statistics for the target variables.

	Tier 1 capital ratio Ending Adverse	Tier 1 leverage ratio Ending Adverse	Tier 1 capital ratio Ending Severely Adverse	Tier 1 leverage ratio Ending Severely Adverse
Mean	11.66	7.86	9.82	6.66
Median	11.4	8	9.4	6.9
Std Dev	1.63	1.44	1.98	1.2
Min	9	5	6.8	4.4
Max	17.1	10.4	16.1	8.8

Table 8
Summary of models to identify significant risk culture features.

Target Variable	Sentiment	Model	Mean Square Error (MSE)	Most significant four variables
Tier 1 Capital (Adverse Scenario)	Alchemy	Ridge	2.95	portfolio_neu; portfolio_neg; regulatory_req_neg; governance_neu
Tier 1 Leverage (Adverse Scenario)	Alchemy	Ridge	1.72	portfolio_neu; reputation_pos; reputation_neu; work culture_neu
Tier 1 Capital (Severely Adverse Scenario)	Alchemy	Ridge	3.91	portfolio_neu; portfolio_neg; portfolio_pos; strategy_neu
Tier 1 Leverage (Severely Adverse Scenario)	Alchemy	Ridge	1.31	reputation_neu; employee_pos; work culture_neu; portfolio_neu

Table 9
Ordered list of most significant risk culture features.

Ordered List of Most Significant Features (Score-based)			
1	portfolio_neu	2	reputation_pos
3	employee_neg	4	reputation_neu
5	work culture_neu	6	employee_pos
7	portfolio_pos	8	portfolio_neg
9	reputation_neg	10	strategy_pos
11	regulatory req_neu	12	strategy_neu
13	governance_pos	14	governance_neu
15	regulatory req_pos	16	regulatory req_neg
17	work culture_pos	18	employee_neu
19	strategy_neg	20	work culture_neg
21	governance_neg		

and work culture are picked up as the second level of significance across these models. Finally, regulatory requirement, governance, strategy and employee feature as important in some of the models. In order to advance the analysis for differences in risk culture among banks based on the significant risk culture indicators, we need to more formally identify the most significant indicators. As described in Section 3.1.1, in order to get a unique ranked list of features, the ranking for the model for each target variable is aggregated using the importance scores, i.e., absolute weight for the Ridge regression. The scores are normalized before aggregation to avoid bias due to scaling. Among the four Ridge regression models using the four target variables, we identify the common most significant variables ordered from most to least significant in Table 9.

All three sentiment flavors of portfolio and reputation appear in the top 13 significant variables in Table 9. Negative and positive sentiments for employee are highlighted as important. Strategy and governance with positive sentiment appear among the top 13 features, while work culture and regulatory requirement appear with a neutral reference. The governance category first appears in the 13th spot. Therefore, top 13 features are considered in the unsupervised learning analysis as this would include at least one risk culture indicator of each type and sentiment. Looking beyond the top 13 variables, the same categories continue to be important in the next few slots. It is interesting to note that, even though work culture occurrences are fewer in the corpus, it appears in the fifth spot in the ranked variable list.

Using the most significant risk culture indicators, we next implement an unsupervised learning clustering analysis to identify how these indicators help distinguish banks from one another in their risk culture characteristics. As discussed in Section 3.2, a k-means clustering methodology is utilized in our study. Based on our analysis of different choices of k, we selected $k = 3$ as a trade-off between degree of cluster cohesion and a meaningful labeling of clusters. We group banks in the study duration into 3 groups by the 13 most important RCI feature variables found from the supervised learning Ridge regression models. We conduct this analysis first at the bank level, assuming that banks' risk culture characteristics do not change through the years. For this purpose we merge the bank-year data through the years for each bank for defining the 13 feature variables for each bank, and a k-means clustering is performed for the banks. Subsequently, we conduct year-by-year clustering using bank-year data for the 13 feature variables to examine how banks' risk culture characteristics changed year-on-year.

Fig. 2 displays the 3 clusters formed of the banks included in this study. The banks form distinct groups, each group with a sizeable membership: Group 0 (blue nodes) consists of 7 banks, Group 1 (green nodes) has 13 members, and Group 2 (red nodes) has 4 banks. The largest bank members of Group 0 are: Citigroup, Wells Fargo, Bank of America, while the largest bank members of Group 1 are JPMorgan and Goldman Sachs, and finally, the largest bank member of Group 2 is State Street. In order to understand the property of each group, we need to review the features of the banks in each cluster. We plot the mean and standard deviation for each feature in a bar chart of Fig. 3 for the banks in each cluster – blue (Group 0), green (Group 1) and red (Group 2), after feature normalization. The bar chart is organized by features, with levels for each feature for the three clusters grouped together to facilitate a comparison between clusters for each feature. Comparison of normalized feature levels between different clusters is quite instructive. The bar heights for each feature corresponds to the mean level of the feature for the three clusters, with the most important feature chosen from supervised learning organized from left to right and the length of the line overlaid on each colored feature bar is the standard deviation of that feature for each cluster.

Focusing on the positive sentiment features for reputation, employee, strategy, portfolio and governance, and negative sentiment for portfolio and reputation, Group 0 (blue nodes) does very poorly compared to Groups 1 and 2. Group 0 (blue nodes) are strongly negative for positive sentiment and strongly positive on the negative sentiment. It is, therefore, safe to label Group 0 (blue nodes) as one with weakest risk culture. Comparing Group 1 (green nodes) and Group 2 (red nodes), we observe that Group 2 (red nodes) has a high level of reputation positive feature, while Group 1 (green nodes) has a high level of employee negative feature. Portfolio positive feature is negative for Group 1 (green nodes), and high and positive for Group 2 (red nodes). Based on these differences in terms of the most significant features, we label Group 2 (red nodes) to have the strongest risk culture and Group 1 (green nodes) to have a moderate risk culture.

We need to shift our attention to examine how risk culture has changed through the years of Fed's stress testing program. In Fig. 4, from top to bottom yearly clusters of banks for 2013, 2014 and 2015 are provided. In the right panel of the figure, corresponding feature summaries are provided in a similar normalized format as seen before for all years combined. Cluster labels are given so that the cluster properties by the feature space is not too dramatically different. An examination across the three years' features characteristics still suggests that Group 2 (red nodes) is the most favorable group for its risk culture characteristics. Group 0 (blue nodes) positions least favorably and Group 1 (green nodes) falls in the intermediate range.

From year to year, the profile of the un-normalized bar charts for each group's risk culture indicators shows some similarities.

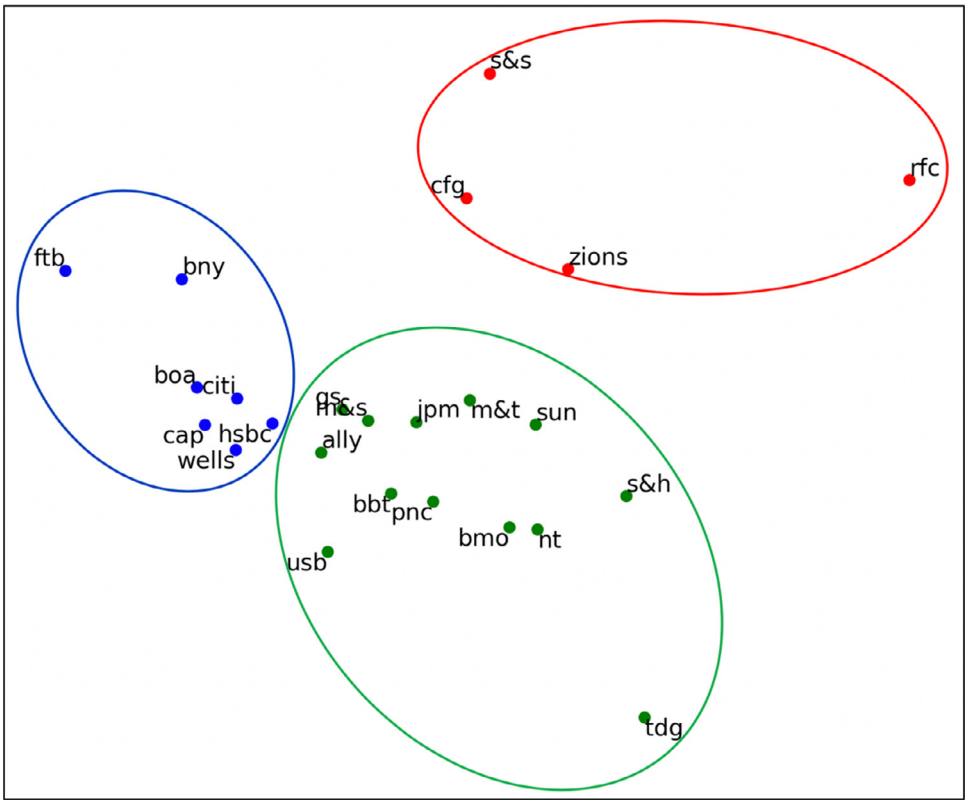

Fig. 2. Clustering of banks into 3 Clusters. (For interpretation of the references to color in this figure, the reader is referred to the web version of this article.)

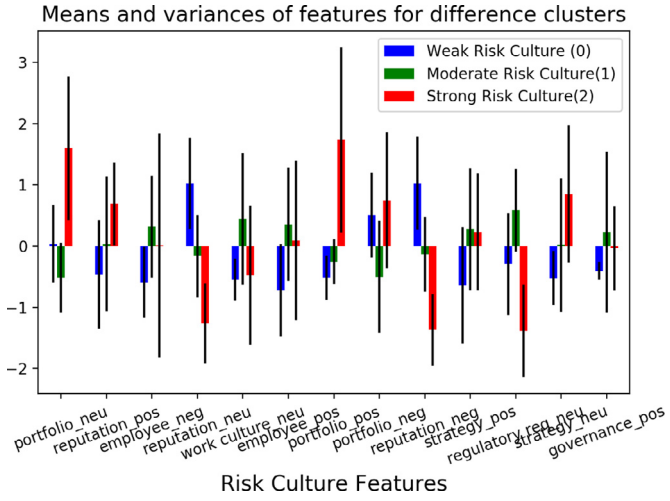

Fig. 3. Properties of each cluster of banks in the 3 Clusters. Here group 0 corresponds to blue nodes in Fig. 2, 1 to green, and 2 to red. (For interpretation of the references to color in this figure, the reader is referred to the web version of this article.)

However, in normalized plots shown in Fig. 4, the profiles differ from each other by year and by quality of risk culture. Yearly labeling by the relative properties is still possible for weakest, moderate and strongest risk culture indication. Throughout, Group 0 is labeled as the group with weakest risk culture, Group 1 with moderate and Group 2 with the best profile of risk culture characteristics. Group 0 consistently shows weak trends by reputation, employee or portfolio characteristics. In the same token, Group 2 is consistently strong in employee positive sentiment among other redeeming characteristics. One must note that the number of banks

Table 10
Number of banks in each group in each year.

Year	Group 0	Group 1	Group 2	Total
2013	14	4	3	21
2014	4	5	14	23
2015	6	11	6	23

in each year's stress testing program has changed, with the most recent set of banks undergoing the Fed's stress tests being the largest. Table 10 displays the membership size of each group in each year. Group 0 (blue nodes, weakest risk culture) was initially the largest group, while Group 2 (red nodes, strongest risk culture) was the smallest group with Goldman Sachs being a member. The increasing total number of banks going through the Fed's stress tests, as well as the changing size of the clusters each year suggests that some banks have migrated from a cluster to another over the three years. If this migration is for the better, this suggests that the stress test program was instrumental in improving the risk culture of these banks. And an improvement would be if a bank goes from Group 0 to Group 1 or Group 2, or from Group 1 to Group 2.

Tables 11 and 12 show the number and names of the banks that transition from a group to another. Along the diagonal of these transitions matrices are the banks that remained in their group in the two consecutive years. From year 2013 to 2014, there were 14 positive transitions, 2 banks moved from Group 0 to Group 1 and 11 from Group 0 to Group 2. On the other hand, there were also several negative transitions, namely from Group 1 over to Group 0 and from Group 2 to Group 1. There was 1 bank that retained its Group 1 (green node) status. There were significant, but less dramatic, transitions from 2014 to 2015, as seen in the diagonal of Table 12. There was 1 dramatic drop from Group 2 to Group 0 and

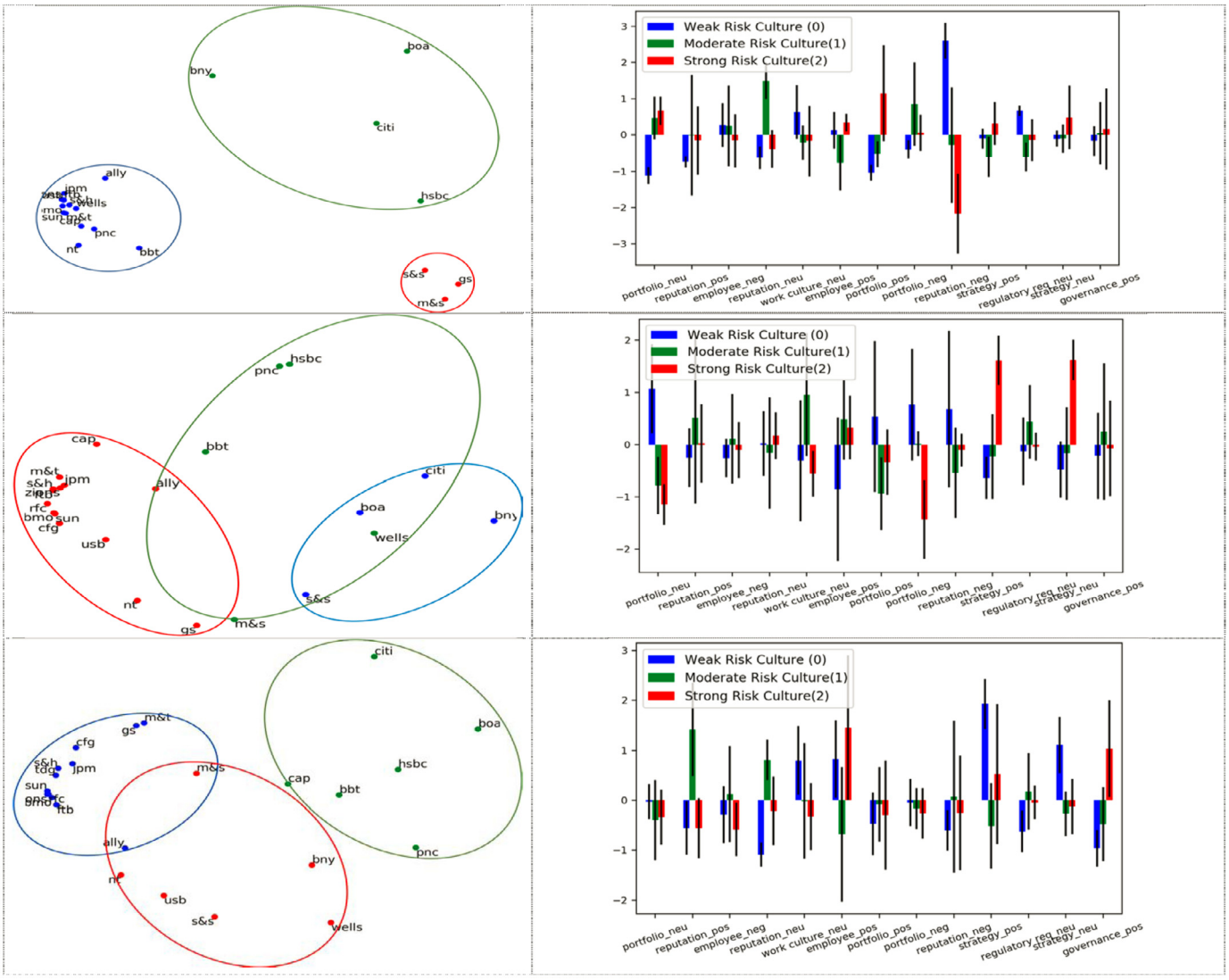

Fig. 4. Yearly clustering of banks into 3 Clusters along with each clusters feature mean and standard deviation. Years 2013, 2014 and 2015. (For interpretation of the references to color in this figure, the reader is referred to the web version of this article.)

Table 11
Number and names of banks making cluster transitions from year 2013 to year 2014.

	2014 Group 0	2014 Group 1	2014 Group 2
2013 Group 0	0	3 (‘pnc’,‘bbt’,‘wells’)	11 (‘jpm’,‘usb’,‘sun’,‘cap’,‘bmo’, ‘ftb’,‘s&h’,‘m&t’,‘zions’,‘nt’,‘ally’)
2013 Group 1	3 (‘boa’,‘bny’,‘citi’)	1 (‘hsbc’)	0
2013 Group 2	1 (‘s&s’)	1 (‘m&s’)	1 (‘gs’)

Table 12
Number and names of banks making cluster transitions from year 2014 to year 2015.

	2015 Group 0	2015 Group 1	2015 Group 2
2014 Group 0	2 (‘boa’,‘citi’)	0	2 (‘bny’,‘s&s’)
2014 Group 1	3 (‘pnc’,‘bbt’,‘hsbc’)	0	2 (‘m&s’,‘wells’)
2014 Group 2	1 (‘cap’)	11 (‘jpm’,‘s&h’,‘cfg’,‘bmo’,‘rfc’, ‘ftb’, ‘m&t’, ‘zions’, ‘sun’, ‘gs’, ‘ally’)	2 (‘nt’,‘usb’)

2 positive transitions from Group 0 to Group 2 made by Bank of New York Mellon and State Street.

Similar analysis can also be done for risk culture transitions in the future years as the Federal Reserve’s stress tests results become available. As new banks are added to the stress testing program, and as text sources become available for these banks, features can be extracted as needed for the definition of the clusters, and the risk culture of the bank can be identified based on the cluster the bank is found to belong in. If arbitrary large banks need to be classified for their risk culture based on textual news data, a re-assessment of the unsupervised learning model may be needed for its continued suitability. An updated unsupervised clustering can

then be used to identify the new banks risk culture. Moreover, one may argue that the analysis done here on the basis of news articles on the banks undergoing stress tests can be complemented by other documents on these banks, such as the banks' SEC filings of annual reports, etc. This would be a worthy analysis to extend the work done here. Our objective here was to evaluate how well an external reporting resource was able to evaluate banks' conditions year-on-year as they underwent the Federal Reserve stress tests.

5. Concluding remarks

We explore the hypothesis in this paper that news articles that cover banks on a range of topics can lead us to evaluate the bank's risk culture. Following risk culture indicators from the literature, we define features based on a large corpus of new articles to support the risk culture learning. Using two alternate sentiment identification approaches, the features are defined and evaluated. Thereafter, a supervised learning method based on Ridge regression models is used. Target variables from Fed's stress tests are used to identify the most significant variables that contribute the greatest towards prediction in the Ridge regression learning. A clustering based unsupervised learning based on the most important variables identified in the supervised learning shows how banks can be grouped by their risk culture characteristics.

Stress test scores used as target variables in supervised learning help identify the most important features for risk culture. Even though stress testing is a regulatory initiative, the most important variables are found to be those that indicate a bank's portfolio, reputation, employee characteristics and strategy. Regulatory requirement feature shows up as eleventh important variable. Moreover, one may think that employee training and retention are important for risk culture identification, and in fact, both negative and positive sentiment for it show up among the top 6 risk culture features. Work culture emerges as important at the 5^{th} place, despite having fewer data points for it in the news articles corpus.

The unsupervised learning based on the most important 13 features identified in supervised learning allows clustering of the banks in distinct groups. An examination of the feature levels of each group arguably allows assigning high, medium, and low quality of risk culture to the three groups. We similarly implement clustering on the bank-year features data to examine how, if at all, the banks' risk culture characteristics change with time. While some banks stick to their original group membership, we observe migrations that suggest improvement and deterioration in risk culture characteristics. We labeled all 24 banks for their risk culture characteristics through the 2013–2015 period.

Freezing the clusters from 2015 or considering the clusters created for all the years' data combined, future transitions of bank's risk culture can be examined as more Fed's stress test results and corresponding year's news articles corpus becomes available. Similarly, as new banks get included in the Fed's stress test program, the bank's risk culture can be examined by determining which group it belongs to based on the features defined for it using the news articles data. A reassessment and updating of the unsupervised learning model may be needed for its continued suitability, if the bank set or the textual data source is expanded significantly beyond what is considered in this article.

Appendix A

Tables A.1–A.2

Table A.1
Banks included in this study along with the symbol associated with them and used in the clustering results.

Banks
J.P.Morgan Chase & Co (jpm)
Bank of America (boa)
Wells Fargo (wells)
Citigroup (citi)
Goldman Sachs Group (gs)
Morgan Stanley (m&s)
U.S. Bancorp (usb)
Bank of New York Mellon (bny)
PNC Financial Services (pnc)
Capital One (cap)
HSBC North America Holdings (hsbc)
TD Group US Holding (tdg)
State Street Corporation (s&s)
BB&T Corporation (bbt)
SunTrust Banks (sun)
Ally Financial (ally)
Fifth Third Bank (ftb)
Citizens Financial Group (cfg)
Santander Holdings USA (s&h)
BMO Financial Corp (bmo)
Regions Financial Corporation (rfc)
M&T Bank Corporation (m&t)
Zions Bancorporation (zion)
Northern Trust Corporation (nt)

Table A.2
Sample of keywords representing each Risk Culture Indicator (RCI).

Category	Keywords
Regulatory requirements	Risk management, risk-taking behavior, ethics, compliance, Regulatory failings, regulatory demand, disclosure report, controls, risk delegation, risk education, testing, limit, proactive, report.
Governance	Asset, experience, governance, leadership, operational excellence, competent leader, professional, authority, monitor.
Portfolio	Debt, over dues, default, impairment level, off-balance positions, derivatives, below target performance, loss, non-performing asset, leverage, write-offs.
Employees	Layoff, job cut, job elimination, workforce reduction, attrition, hiring, recruitment, skilled, competent, review, performance.
Risk strategy	Takeover, buyout, acquisition, merger, risk portfolio, subsidiary, subsidiaries, new venture,dividend payout, risk strategy, holding, risk framework, risk appetite, risk perspective, priority, resources.
Reputation	Lawsuit, litigation, sue, compliance fail, compliance risk, legal issue, penalty, complaints, integrity, honest, competence, confidence.
Work culture	Process breach, procedure breach, reluctance, internal competition, centralized decision, decentralized decision, high pressure, vulnerable environment, work around, avoidance, delegation, integrity.

References

Alfano, S. J., Feuerriegel, S., & Neumann, D. (2015). *Do pessimists move asset prices? evidence from applying prospect theory to news sentiment*. Available at SSRN: https://ssrn.com/abstract=2602353.

Antweiler, W., & Frank, M. Z. (2004). Is all that talk just noise? The information content of internet stock message boards. *The Journal of Finance, 59*(3), 1259–1294.

Ashby, S. (2014). *Guidance on supervisory interaction with financial institutions on risk culture feedback on the FSBs consultative document* Ph.D. thesis. London School of Economics.

Bae, S. C., Chang, K., & Kang, E. (2012). Culture, corporate governance, and dividend policy: international evidence. *Journal of Financial Research, 35*(2), 289–316.

Balakrishnan, R., Qiu, X. Y., & Srinivasan, P. (2010). On the predictive ability of narrative disclosures in annual reports. *European Journal of Operational Research, 202*(3), 789–801.

Bodnaruk, A., Loughran, T., & McDonald, B. (2015). Using 10-k text to gauge financial constraints. *Journal of Financial and Quantitative Analysis, 50*(04), 623–646.

Bowman, E. H. (1984). Content analysis of annual reports for corporate strategy and risk. *Interfaces, 14*(1), 61–71.

Bozeman, B., & Kingsley, G. (1998). Risk culture in public and private organizations. *Public Administration Review*, 109–118.
Burns, T. E., & Stalker, G. M. (1961). *The management of innovation*. University of Illinois at Urbana-Champaign's Academy for Entrepreneurial Leadership Historical Research Reference in Entrepreneurship.
Chatrath, A., Miao, H., Ramchander, S., & Villupuram, S. (2014). Currency jumps, cojumps and the role of macro news. *Journal of International Money and Finance, 40*, 42–62.
Chernobai, A., Jorion, P., & Yu, F. (2012). The determinants of operational risk in us financial institutions. *Journal of Financial and Quantitative Analysis, 46*(06), 1683–1725.
Chiang, C., Han, C.-C., Chiang, Y.-M., Tsai, T.-C., Wu, F.-S., Seng, D., et al. (2015). *Market liquidity, funding liquidity in the news and housing price*. Available at SSRN: https://ssrn.com/abstract=2565340.
Drennan, L. T. (2004). Ethics, governance and risk management: Lessons from mirror group newspapers and barings bank. *Journal of Business Ethics, 52*(3), 257–266.
Feuerriegel, S., Wolff, G., & Neumann, D. (2015). *Information processing of foreign exchange news: Extending the overshooting model to include qualitative information from news sentiment*. SSRN, 2603435.
Fritz-Morgenthal, S., Hellmuth, J., & Packham, N. (2016). Does risk culture matter? The relationship between risk culture indicators and stress test results. *Journal of Risk Management in Financial Institutions, 9*(1), 71–84.
FSB (2014). Guidance on supervisory interaction with financial institutions on risk culture: A framework for assessing risk culture. *Technical Report*. Financial Stability Board.
Gallemore, J. (2013). Does bank opacity enable regulatory forbearance. *Technical Report*. Chicago Booth.
Geretto, E. F., & Pauluzzo, R. (2015). Knowledge management and risk culture in the banking industry: relations and problems. In *Proceedings of the European conference on knowledge management* (p. 313). Academic Conferences International Limited.
Gherardi, S., & Nicolini, D. (2000). To transfer is to transform: The circulation of safety knowledge. *Organization, 7*(2), 329–348.
Glazer, E., & Rexrode, C. (2016). As regulators focus on culture, wall street struggles to define it. *The Wall Street Journal*.
IIF (2009). *Reform in the financial services industry: Strengthening practices for a more stable system*. Institute of International Finance Report.
Iturriaga, F. J. L., & Sanz, I. P. (2015). Bankruptcy visualization and prediction using neural networks: A study of us commercial banks. *Expert Systems with Applications, 42*(6), 2857–2869.
Kimbrough, R. L., & Componation, P. J. (2009). The relationship between organizational culture and enterprise risk management. *Engineering Management Journal, 21*(2), 18–26.
Law, J., & Singleton, V. (2005). Object lessons. *Organization, 12*(3), 331–355.
Li, F. (2010). The information content of forward-looking statements in corporate filings a Naïve Bayesian machine learning approach. *Journal of Accounting Research, 48*(5), 1049–1102.
Lischinsky, A. (2011). In times of crisis: A corpus approach to the construction of the global financial crisis in annual reports. *Critical Discourse Studies, 8*(3), 153–168.
Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies, 5*(1), 1–167.
Loughran, T., & McDonald, B. (2011a). When is a liability not a liability? Textual analysis, dictionaries, and 10-ks. *The Journal of Finance, 66*(1), 35–65.
Loughran, T., & McDonald, B. (2011b). When is a liability not a liability? Textual analysis, dictionaries, and 10-ks. *The Journal of Finance, 66*(1), 35–65.
Loughran, T., & McDonald, B. (2014). Regulation and financial disclosure: The impact of plain english. *Journal of Regulatory Economics, 45*(1), 94–113.
Mäntylä, M. V., Graziotin, D., & Kuutila, M. (2018). The evolution of sentiment analysis - A review of research topics, venues, and top cited papers. *Computer Science Review, 27*, 16–32.
McConnell, P. J. (2013). A risk culture framework for systemically important banks. *Journal of Risk and Governance, 3*(1), 23–68.
Mengelkamp, A., Hobert, S., & Schumann, M. (2015). Corporate credit risk analysis utilizing textual user generated content-a twitter based feasibility study. In *PACIS* (p. 236).
Mihet, R. (2013). Effects of culture on firm risk-taking: A cross-country and cross-industry analysis. *Journal of Cultural Economics, 37*(1), 109–151.
Nassirtoussi, A. K., Aghabozorgi, S., Wah, T. Y., & Ngo, D. C. L. (2014). Text mining for market prediction: A systematic review. *Expert Systems with Applications, 41*(16), 7653–7670.
Nassirtoussi, A. K., Aghabozorgi, S., Wah, T. Y., & Ngo, D. C. L. (2015). Text mining of news-headlines for forex market prediction: A multi-layer dimension reduction algorithm with semantics and sentiment. *Expert Systems with Applications, 42*(1), 306–324.
Nguyen, T. H., Shirai, K., & Velcin, J. (2015). Sentiment analysis on social media for stock movement prediction. *Expert Systems with Applications, 42*(24), 9603–9611.
Nyman, R., Gregory, D., Kapadia, S., Ormerod, P., Tuckett, D., & Smith, R. (2015). News and narratives in financial systems: Exploiting big data for systemic risk assessment. *Technical Report*. BoE, mimeo.
Palermo, T., Power, M., Ashby, S., Jordan, S., Munro, I., & Maguire, S. (2015). *Searching for risk culture: Sites and dynamics*. Available at: http://www.hec.unil.ch/documents/seminars/dcc/1804.pdf.
Power, M., Ashby, S., & Palermo, T. (2013). *Risk culture in financial organisations: A research report*. Available at: http://eprints.lse.ac.uk/67978/1/Palermo_Rsik%20culture%20research%20report_2016.pdf.
Rachlin, G., Last, M., Alberg, D., & Kandel, A. (2007). Admiral: A data mining based financial trading system. In *Proceedings of the IEEE symposium on computational intelligence and data mining, CIDM* (pp. 720–725). IEEE.
Ranco, G., Bordino, I., Bormetti, G., Caldarelli, G., Lillo, F., & Treccani, M. (2016). Coupling news sentiment with web browsing data improves prediction of intra-day price dynamics. *PloS one, 11*(1), e0146576.
Rekabsaz, N., Lupu, M., Baklanov, A., Hanbury, A., Dür, A., & Anderson, L. (2017). *Volatility prediction using financial disclosures sentiments with word embedding-based ir models* arXiv preprint arXiv:1702.01978, Feb 7.
Rönnqvist, S., & Sarlin, P. (2015a). Detect & describe: Deep learning of bank stress in the news. In *Proceedings of the IEEE symposium series on computational intelligence* (pp. 890–897). IEEE.
Rönnqvist, S., & Sarlin, P. (2015b). Identifying bank stress by deep learning of news. In *Proceedings of the workshop new challenges in neural computation* (p. 112). Citeseer.
Rönnqvist, S., & Sarlin, P. (2017). Bank distress in the news: Describing events through deep learning. *Neurocomputing, 264*, 57–70.
Schumaker, R. P., Zhang, Y., Huang, C.-N., & Chen, H. (2012). Evaluating sentiment in financial news articles. *Decision Support Systems, 53*(3), 458–464.
Serrano, E., & Iglesias, C. A. (2016). Validating viral marketing strategies in twitter via agent-based social simulation. *Expert Systems with Applications, 50*, 140–150.
Sheedy, E., & Griffin, B. (2014). Empirical analysis of risk culture in financial institutions: Interim report. *Technical Report*. Macquarie University, Centre for International Finance and Regulation.
Sheedy, E. A., Griffin, B., & Barbour, J. P. (2015). A framework and measure for examining risk climate in financial institutions. *Journal of Business and Psychology, 32*(1), 101–116.
Tai, I., Olson, B., & Blessner, P. (2016). Unsupervised text mining approach to early warning system. *World Academy of Science, Engineering and Technology, International Journal of Computer, Electrical, Automation, Control and Information Engineering, 10*(4), 651–656.
Tsai, F.-T., Lu, H.-M., & Hung, M.-W. (2010). The effects of news sentiment and coverage on credit rating analysis.. In *Proceedings of the PACIS* (p. 199).
Tsai, M.-F., & Wang, C.-J. (2017). On the risk prediction and analysis of soft information in finance reports. *European Journal of Operational Research, 257*(1), 243–250.
Tsai, M.-F., Wang, C.-J., & Chien, P.-C. (2016). Discovering finance keywords via continuous-space language models. *ACM Transactions on Management Information Systems (TMIS), 7*(3), 7.
Vu, T.-T., Chang, S., Ha, Q. T., & Collier, N. (2012). An experiment in integrating sentiment features for tech stock prediction in twitter. In *Proceedings of the workshop on information extraction and entity analytics on social media data* (pp. 23–38).
Wuthrich, B., Cho, V., Leung, S., Permunetilleke, D., Sankaran, K., & Zhang, J. (1998). Daily stock market forecast from textual web data. In *Proceedings of the IEEE international conference on systems, man, and cybernetics: 3* (pp. 2720–2725). IEEE.
Zhang, Y., Swanson, P. E., & Prombutr, W. (2012). Measuring effects on stock returns of sentiment indexes created from stock message boards. *Journal of Financial Research, 35*(1), 79–114.