European Journal of Operational Research 274 (2019) 743–758

Contents lists available at ScienceDirect

European Journal of Operational Research

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

Innovative Applications of O.R.

Deep learning models for bankruptcy prediction using textual disclosures

Feng Mai[a,*], Shaonan Tian[b], Chihoon Lee[a], Ling Ma[a]

[a] School of Business, Stevens Institute of Technology, Hoboken, NJ 07030, USA
[b] Lucas College and Graduate School of Business, San José State University, San José, CA 95192, USA

ARTICLE INFO

Article history:
Received 20 October 2017
Accepted 11 October 2018
Available online 19 October 2018

Keywords:
Decision support systems
Deep learning
Bankruptcy prediction
Machine learning
Textual data

ABSTRACT

This study introduces deep learning models for corporate bankruptcy forecasting using textual disclosures. Although textual data are common, it is rarely considered in the financial decision support models. Deep learning uses layers of neural networks to extract features from textual data for prediction. We construct a comprehensive bankruptcy database of 11,827 U.S. public companies and show that deep learning models yield superior prediction performance in forecasting bankruptcy using textual disclosures. When textual data are used in conjunction with traditional accounting-based ratio and market-based variables, deep learning models can further improve the prediction accuracy. We also investigate the effectiveness of two deep learning architectures. Interestingly, our empirical results show that simpler models such as averaging embedding are more effective than convolutional neural networks. Our results provide the first large-sample evidence for the predictive power of textual disclosures.

1. Introduction

Corporate bankruptcy is one of the main drivers of the credit risk and gains primary attention from creditors and investors. The financial damage inflicted by corporate bankruptcy cannot be overstated. The 2008-2010 financial crisis has also shown that in aggregate the corporate bankruptcy events have a profound influence on the economy. Corporate bankruptcy may incur a strong negative social cost and further propagate recession and thus jeopardize the economy at large (Bernanke, 1981). An accurate bankruptcy forecasting model, therefore, is valuable to practitioners, regulators, and academic researchers alike (Ding, Tian, Yu, & Guo, 2012). Regulators can use the model to monitor the financial health of individual institutions and curb systemic risks. Practitioners rely on predicted default probability to price corporate debt and for internal ratings (Schönbucher, 2003). For academics, corporate distress risk can help calibrate various theoretical models, such as explaining anomalies in the standard CAPM (Campbell, Hilscher, & Szilagyi, 2008). For these reasons, researchers search for more effective prediction models to forecast bankruptcy and financial distress.

Studies in bankruptcy prediction routinely adopt measures including firms' stock market trading information and accounting data from company's financial statements to forecast bankruptcy. Dating back to Altman (1968), research has shown that accounting-based ratios and stock market data offer signals on whether a firm is financially healthy or may step into severe trouble like bankruptcy. Given the high impact of corporate bankruptcy events, researchers in operational research (OR) and artificial intelligence (AI) further propose intelligent models to forecast bankruptcy. New modeling techniques include boosting, discriminant analysis, support vector machine, and neural networks (Alfaro, García, Gámez, & Elizondo, 2008; Bose & Pal, 2006; Geng, Bose, & Chen, 2015), data envelopment analysis (Li, Crook, & Andreeva, 2014), least absolute shrinkage and selection operator (Tian, Yu, & Guo, 2015), dynamic slacks based model (Wanke, Barros, & Faria, 2015), two-stage classification (du Jardin, 2016), to name a few.

A common element of those models is the application of market-based and accounting-based variables, which are usually constructed using numeric data in a well-structured format. Yet, there is growing recognition that text disclosure — a form of unstructured, qualitative data — plays an equally important role in how information is conveyed to the public. For example, a vast proportion of public firm's annual filings to regulatory agencies are textual disclosures. Also, policymakers and market participants consume a large amount of financial reports and news articles every day. Despite its ubiquity, effective integration of textual disclosure in financial models remains a challenging mission due to the difficulty in both obtaining and quantifying textual data (Lang & Stice-Lawrence, 2015). Recent studies have demonstrated strong

* Corresponding author.
E-mail addresses: fmai@stevens.edu (F. Mai), shaonan.tian@sjsu.edu (S. Tian), clee4@stevens.edu (C. Lee).

https://doi.org/10.1016/j.ejor.2018.10.024
0377-2217/

evidence that qualitative corporate filings contain valuable information about credit risk (Campbell, Chen, Dhaliwal, Lu, & Steele, 2014; Loughran & Mcdonald, 2011). Most studies, however, rely on simple text summarization techniques such as word count, sentiment, and readability. The information and signals in financial text go well beyond these measures (Bozanic & Thevenot, 2015). To leverage the full value of the textual disclosures, the need for more effective algorithms to extract and exploit information from textual data is higher than ever.

In this research, we shed lights on this issue by proposing a new deep learning method to forecast bankruptcy and assessing the predictive power of textual data. Deep learning is a machine learning paradigm that combines multiple layers of neural networks to learn representations of data with multiple levels of abstraction (Le Cun et al., 2015). These deeper neural networks have shown promising results in many areas including image recognition, language processing, and machine translation thanks to their ability to extract features from unstructured data such as image and text. Motivated by these observations, we aim to design a deep learning approach to predict firm bankruptcies using textual disclosures. We first construct a comprehensive database of 11,827 U.S. public traded firms over the period of 1994-2014. The database consists of numeric variables generated from accounting and stock market data. We then extract the qualitative discussion section, Managerial Discussion & Analysis (MD&A), from firm's annual filing and match with our observations. We investigate different model set-ups using varying input data. Our empirical study shows that a simple deep learning model using an average of the embedding layer outperforms other data mining models when textual information is used. More importantly, we find that the textual data can complement the traditional accounting-based and market-based variables in predicting bankruptcy. The deep learning model using both textual and numeric inputs has improved prediction accuracy over the models using a single type of input.

Our study makes several important contributions. First, to the best of our knowledge, this paper is the first large-sample analysis of bankruptcy prediction using textual disclosures. To date, there is limited empirical evidence on whether financial disclosure in textual form can be used in an intelligent system for bankruptcy prediction. Our study complements the literature by adding new insights on how textual data can signal early warning signs for corporate bankruptcy events. Second, we show that deep learning is a promising framework for predicting financial outcomes. Although using artificial neural networks for bankruptcy prediction has a long history (Wilson & Sharda, 1994; Zhang, Hu, Patuwo, & Indro, 1999), prior studies use numerical inputs combined with shallow networks (i.e., one or two layers). We provide strong evidence that a trained deep neural network system can discriminate bankruptcy and non-bankruptcy firms, especially when the input includes textual information. Third, we also contribute to the natural language processing literature by showcasing an impactful application area. Current deep learning research on natural language processing (NLP) has shown considerable progress on tasks such as parsing sentences and machine translation, but little is known as to whether financial institutions and regulators can apply this budding technique. Our research points to a new area to which deep learning research can contribute. Not only predicting financial distress is of great practical importance, but it is also an area where ground truth dataset can serve as the foundation of a common task framework (Hofman, Sharma, & Watts, 2017).

The rest of the paper proceeds as follows. We review pertinent literature on bankruptcy prediction in Section 2. We describe how we construct our samples in Section 3. We introduce deep learning models and compare our model architectures along with several other data mining benchmarks in Section 4. We report and discuss the model evaluation results in Section 5. We offer some concluding remarks and discuss future research directions in Section 6.

2. Literature review

When predicting corporate bankruptcy, researchers have routinely used accounting-based variables (e.g., profitability ratio and liability ratios) and market-based variables (e.g., stock market returns and volatility) as a gauge of default risk. We refer readers to Kumar and Ravi (2007) and Demyanyk and Hasan (2010) that provide comprehensive literature reviews on the studies before 2008. In Table 1, we curate a list of more recent studies on bankruptcy prediction. We summarize each study based on the sample selection criteria, source of database and countries, sample size, time period, models used, and variable types. We now highlight several themes that emerge from the table.

First, the methods used by recent studies align with the two major categories featured in Kumar and Ravi (2007)'s review: statistical models and intelligent models. The first category of research continues to focus on statistical properties of the model (e.g., Campbell et al., 2008, Ding et al., 2012). Recent studies focus on developing statistical models to improve the model's prediction accuracy and provide more insights in examining distress risk. For example, researchers can identify the most relevant features and their relative weights using statistical models in bankruptcy prediction. Such identification can help test bankruptcy theories and guide regulations in credit markets. Popular models include discriminant analysis, logistic regression models, and factor analysis.

Most of the studies in Table 1 fall into the intelligent category. The goal is to develop more accurate models using artificial intelligence and operations research techniques. In contrast with the statistical studies, the intelligent techniques make fewer assumptions about the data. Also, models that allow non-linear decision boundaries (e.g., neural networks, SVM with non-linear kernels) quickly gained popularity and are now widely applied. These features provide better model flexibility and improved classification performance. A trend in recent literature is studying the combinations of models. A number of studies demonstrate how to combine various models horizontally using ensemble techniques (e.g., Geng et al., 2015, Kim & Kang, 2010), or vertically (e.g., du Jardin, 2016). These hybrid models can capture more variations in the decision space and result in more stable and accurate predictions.

Second, we notice a wide diversification of data sources in recent studies. As noted before, theoretical and empirical studies have long established that accounting-based ratios and market-based variables are the main indicators of future bankruptcy. More recent studies have started to evaluate the predictive power of data sources beyond the two types of variables. For example, Liang, Lu, Tsai, and Shih (2016) examine the discriminatory power of a broad array of corporate governance indicators, including board structure, ownership structure, leadership personnel, and others. Doumpos, Andriosopoulos, Galariotis, Makridou, and Zopounidis (2017)'s model takes country characteristics into account. They show that country-level data on the economic and business environment, energy efficiency policies, as well as characteristics of markets can add value to corporate failure prediction models. Calabrese, Degl'Innocenti, and Osmetti (2017) study how the U.S. government's Troubled Asset Relief Program (TARP) impacted the probability of failure among commercial banks. Examining the effectiveness of these new data sources can expand the scope of features selections for prediction models and offer policy prescriptions.

Our study extends the bankruptcy prediction literature in two major ways. We examine the predictive power of a new form of data — firm's textual disclosure in annual reports. Despite its wide circulation and being designed as a leading indicator of future

Table 1
Recent studies on bankruptcy prediction.

Study	Industry	Source of data/country	Sample size	Models	Time period	Variables type
Premachandra, Bhabra, and Sueyoshi (2009)	Firms	USA, Compustat, CRSP	200	DEA, logit	1991–2004	Accounting, market
Psillaki, Tsolas, and Margaritis (2010)	Firms, textiles; wood and paper products; computer activities and R&D	France, Bureau van Dijk – Diane	5751	DEA+logit	2000–2004	Measure of efficiency, accounting
Sueyoshi and Goto (2009)	Firms, construction	Japanese construction industry	1091	DEA-DA, PCA	2000–2005	Accounting
Li and Sun (2009)	Firms	China, specially treated (ST)	162	Electre-CBR-I, Electre-CBR-II, ANOVA feature selection	2000–2005	Accounting
Geng et al. (2015)	Firms	China, CSMAR	214	NN, DT, SVM, MV	2001–2008	Accounting
Wanke et al. (2015)	Banks	Brazilian, Economatica	640	DEA+DSBM	1996–2011	
du Jardin (2015)	Firms, retail, construction, services	France, Bureau van Dijk – Diane database	18,620	DA, logit, MLP, SA	2003–2012	Accounting
du Jardin (2016)	Firms	France, Bureau van Dijk – Diane database	17,660	Bagging, boosting, random subspace, PBM	2003–2012	Accounting
Doumpos et al. (2017)	Firms, energy	18 EU countries, Bureau van Dijk, Eurostat, IEA, OECD, and UNECE	138,387	MCDA	2012–2016	Accounting, Macroeconomic, energy markets
Liang et al. (2016)	Firms, manufacturing, service	Taiwan Economic Journal (TEJ)	478	SVM, KNN, NB, CART, MLP	1999–2009	Accounting, market, corporate governance
Calabrese et al. (2017)	Banks	U.S. Department of the Treasury, FDIC, Call Reports		LOBGEV(GEV model and D-vine copula)	2008–2013	Combination of variables
Olson, Delen, and Meng (2012)	Firms	USA, Compustat	1321	DT, logit, MLP, RBFN, SVM	2005–2009	Accounting
Serrano-Cinca and GutiéRrez-Nieto (2013)	Banks	USA, FDIC	8293	PLS-DA	2008–2011	Accounting
Ioannidis, Pasiouras, and Zopounidis (2010)	Banks	78 countries, Bankscope, World Bank	944	UTADIS, MLP, CART, KNN, Ordered logit, stacked models	2007–2008	Accounting, country-level variables
Boyacioglu, Kara, and Baykan (2009)	Banks	Turkey, Banks Association of Turkey	76	NN, SVM, MDA, K-means cluster analysis, logit	1997–2004	Accounting
Cecchini et al. (2010)	Firms	USA, Compustat, CRSP	156	SVM	1994–1999	MD&A, Altman variables
Chauhan, Ravi, and Chandra (2009)	Banks	Turkey, Spanish, US	129	DEWNN: Differential evolution + Wavelet NN, TAWNN, WNN	1975–1985	Accounting
Etemadi, Rostamy, and Dehkordi (2009)	Firms	Iran, Tehran stock exchange	144	GP, MDA	1998–2005	Accounting
Kim and Kang (2010)	Firms	Korea	1458	MLP + bagging, MLP + boosting	2002–2005	Accounting
Yeh, Chi, and Hsu (2010)	Firms, information and electronic manufacturing	Taiwan Stock Exchange	114	DEA + Rough sets + SVM	2005–2007	Accounting, efficiency
M.-Y. Chen (2011)	Firms	Taiwan Stock Exchange	100	PCA, DT, logit	2000–2007	Accounting
De Andrés et al. (2011)	Firms, manufacturing	Spain, bureau van Dijk and Informa	59,474	Fuzzy clustering + MARS	2007–2008	Accounting (5 Altman variables)
Li, Sun, and Sun (2009)	Firms	China ST	270	CBR based on outranking relations	2000–2005	Accounting
Huang et al. (2008)	Firms	Taiwan, TEJ	820	Financial analysis model + MLP	2001–2004	Accounting
Campbell et al. (2008)	Firms	USA, Compustat, CRSP	1m+	Discrete Hazard Model	1963–2003	Accounting & Market
Chandra et al. (2009)	Non-Financial Firms	USA, Compustat, CRSP	16,816 1.9m+	MLP, CART, logit, Random Forest, SVM, ensemble, boosting	1962–1999	Accounting & Market
Tian et al. (2015)	Firms	USA, Compustat, CRSP	1.5m+	Discrete Hazard Model, Logit	1980–2009	Accounting & Market
Ding et al. (2012)	Firms	USA, Compustat, CRSP	1m+	Transformation Survival Model	1981–2006	Accounting & Market
Chen et al. (2011)	Firms, small or medium size	France, Diane database	1200	GA + LVQ	2006–2007	Accounting
Sánchez-Lasheras et al. (2012)	Firms, construction	Spain, Bureau van Dijk	63,107	SOM + MARS	2007–2008	Accounting (5 Altman variables)

performance, the predictive power of such textual data cannot be automatically assumed. This is because the Securities Exchange Commission (SEC) grants firms considerable flexibility and encouraged firms to experiment with formats of conveying information (Bryan, 1997). Management may have incentives to hide bearish information or to use vague language in their disclosure. In addition, effective predictive models based on textual data require rethinking the entire modeling process; we cannot simply "plug-in" the inputs to existing data mining and operations research techniques. While two prior studies (Cecchini, Aytug, Koehler, & Pathak, 2010; Mayew, Sethuraman, & Venkatachalam, 2015) have reported that MD&A can discriminate bankrupted and non-bankrupted firms, Cecchini et al. (2010) explored only a small, static sample (156 firms), and Mayew et al. (2015) used manual coding. To this end, we fill the gap by systematically studying bankruptcy prediction modeling using large-scale textual data and compare the performance with models that use numerical data. We propose new intelligent prediction models based on deep learning; we also evaluate how existing data mining methods can be adapted for such task.

Our study is also relevant to a strand of literature that studies the role of qualitative variables in credit scoring models. The qualitative information is considered as soft facts whereas the traditional numerical information such as market information or accounting data is considered as hard facts (Lehmann, 2003). Recently, researchers started to apply text analysis methods to process loan descriptions and incorporate the soft information such as analysts' subjective evaluations in the model (Agarwal, Chen, & Zhang, 2016; Dorfleitner *et al.* 2016). Yet, among those existing models, most models only rely on descriptive statistics from the documents such as length, spelling errors, or tones. Our model is different from these studies. In particular, we propose an end-to-end machine-learning model, in which the learning algorithm goes directly from the raw textual input to the prediction. Our model attempts to incorporate all relevant information in the text for prediction.

Finally, although artificial neural network is a common technique in prior studies, the effectiveness of several deep learning techniques such as Word Embedding (Mikolov, Sutskever, Chen, Corrado, & Dean, 2013) and Convolutional Neural Network (CNN) (Krizhevsky, Sutskever, & Hinton, 2012) have not been evaluated. These new neural network models are cornerstones of the recent development of deep learning. They drive some of the impressive breakthrough results in many areas of AI by automatically finding high-level representations from textual and image data (Le Cun et al., 2015). We evaluate whether, and which deep learning architectures can enhance the model performance. That is, we identify the optimal way to combine different neural network layers.

3. Data

We construct our bankruptcy database by merging three data sources: accounting data from Compustat North America, equity trading data from Center for Research in Security Prices (CRSP), and textual disclosure data from10-K annual filings to the Securities Exchange Commission (SEC). Our primary sample includes all the publicly traded firms from 1994 to 2014. In total, our database includes 11,827 firms and 94,994 firm-years with no missing observations. Table 2 summarizes the yearly distribution of firms.

3.1. Bankruptcy indicator

To build a prediction model, we need to construct a bankruptcy indicator as the binary response variable. We define a company as a bankruptcy case if the company files for either Chapter 7 (liquidation) or Chapter 11 (reorganization) bankruptcy protection code. In particular, the bankruptcy indicator for firm i at time t is set to one if the firm was delisted due to either Chapter 7 or Chapter 11 filing at time t. There are very few cases that firms who were delisted may re-enter the database later, but we do not consider any firm-year observation after their first delisting in our analysis. Conversely, the bankruptcy indicator is set to zero if the firm either (1) stayed or survived in the database through the end of the sampling period or (2) exited from the database due to other reasons such as mergers and acquisitions. As a result, we identify a total of 477 bankruptcy filings over the 1994 to 2014 sampling period. Fig. 1 shows the distribution of bankruptcy probability for each year. The three peaks of the bankruptcy events match with the recessions following the 1997 Asian financial crisis, the burst of the Dot-com bubble in the early 2000s, and the more recent subprime mortgage crisis.

For the set of explanatory variables, we construct a time-varying panel dataset, consistent with Shumway (2001)'s work. Each firm-year in our sample period is a separate observation. It contains all the predictor variables we used in this study and a binary response variable, which indicates the firm's bankruptcy status one-year later. For other prediction horizons such as two-year or three-year, we adjust the firm-year observation by matching the predictor variables with its bankruptcy status after two years or three years. We eliminate the firm-year observations when the gap between the predictor variables and the response variables is different from the corresponding prediction horizon. The main advantage of such panel data structure is that all the historical information of a company is used in forecasting future bankruptcy. It may provide more consistent and accurate out-of-sample prediction when compared with the static model where only one year is selected to observe a firm's characteristics (Shumway, 2001).

3.2. Numerical predictors

For numeric input data, we compile a comprehensive list of 36 predictor variables based on the literature review on the bankruptcy of U.S. firms (Table 3). When predicting the probability of bankruptcy, it is common to consider the accounting information and up-to-date market information that may reflect the company's liability, liquidity, and profitability status. For this purpose, many studies in bankruptcy prediction have proposed relevant accounting-based and market-based predictor variables, for example, Altman (1968); Beaver (1966); Campbell et al. (2008) and Tian et al. (2015). In our study, all variables are obtained by merging annual accounting data from Compustat North America with daily and monthly equity data from CRSP. To construct the accounting-based predictor variables, we first align the firm's fiscal year appropriately with the calendar year. Companies usually report their accounting data with a delay. To ensure that the accounting information we used is observable to the investors at the time of prediction, we further lag all the accounting items by four months. Based on the carefully aligned calendar time, we add the corresponding monthly market-based predictor variables to the accounting-based predictor variables. We provide details of how to construct each variable using CRSP and Compustat database in the Appendix. To avoid any recording errors or outliers, we further winsorize all the numerical predictor variables at its 1% and 99% by replacing values that are lower than 1% with its first percentile and higher than the 99% with its ninety-ninth percentile.

3.3. Textual predictors

A key innovation of our study is that we consider an untapped textual data source — Form 10-K to forecast financial distress. The

Table 2
Firm distribution by year.

Year	Total Firms	Bankrupted Firms	Year	Total Firms	Bankrupted Firms
1994	1783	6	2005	4475	5
1995	3334	22	2006	4451	10
1996	6117	52	2007	4417	38
1997	6183	52	2008	4209	48
1998	5989	57	2009	4036	16
1999	5846	22	2010	3930	12
2000	5567	18	2011	3841	7
2001	5158	36	2012	3799	8
2002	4857	30	2013	3865	8
2003	4579	14	2014	4019	3
2004	4539	13			

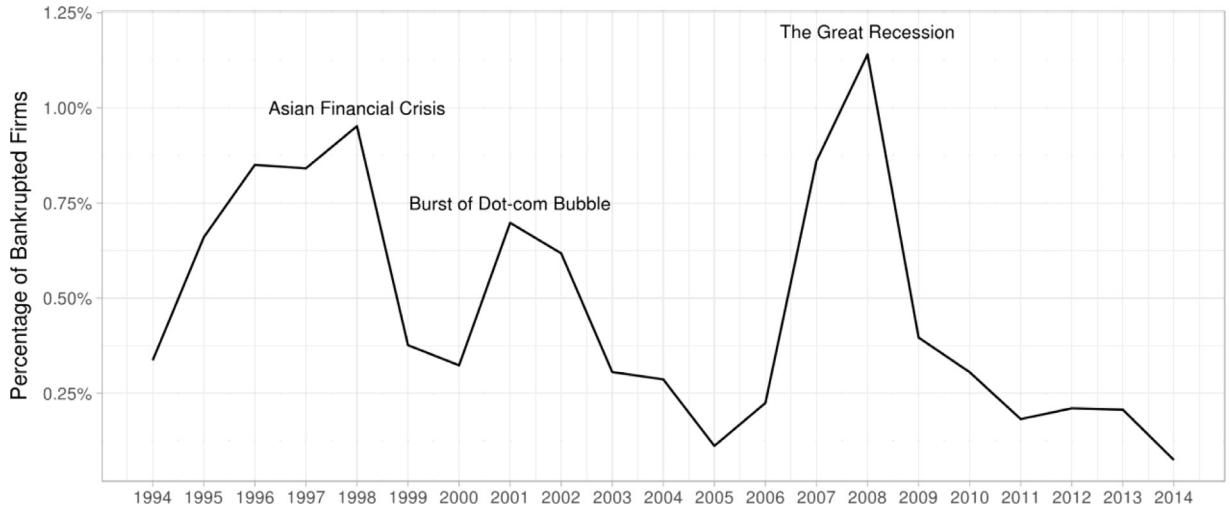

Fig. 1. Rate of bankruptcy in the sample firms (1994–2014).

Table 3
Description of numeric variables.

Variable	Description	Variable	Description
ACTLCT	Current Assets/Current Liabilities	LTMTA	Total Liabilities/(Market Equity + Total Liabilities)
APSALE	Accounts Payable/Sales	LOG(AT)	Log(Total Assets)
CASHAT	Cash and Short-term Investment/Total Assets	LOG(SALE)	Log(Sale)
CASHMTA	Cash and Short-term Investment/(Market Equity + Total Liabilities)	MB	Market-to-Book Ratio
CHAT	Cash/Total Assets	NIAT	Net Income/Total Asset
CHLCT	Cash/Current Liabilities	NIMTA	Net Income/(Market Equity + Total Liabilities)
(EBIT+DP)/AT	(Earnings before Interest and Tax + Amortization and Depreciation)/Total Asset	NISALE	Net Income/Sales
EBITAT	Earnings before Interest and Tax/Total Asset	OIADPAT	Operating Income/Total Asset
EBITSALE	Earnings before Interest and Tax/Sales	OIADPSALE	Operating Income/Sales
EXCESS RETURN	Excess Return Over S&P 500 Index	PRICE	Log(Price)
FAT	Total Debts/Total Assets	QALCT	Quick Assets/Current Liabilities
INVCHINVT	Growth of Inventories /Inventories	REAT	Retained Earnings/Total Asset
INVTSALE	Inventories/Sales	RELCT	Retained Earnings/Current Liabilities
(LCT-CH)/AT	(Current Liabilities – Cash)/Total Asset	RSIZE	Log(Market Capitalization)
LCTAT	Current Liabilities/Total Asset	SALEAT	Sales/Total Assets
LCTLT	Current Liabilities/Total Liabilities	SEQAT	Equity/Total Asset
LCTSALE	Current Liabilities/Sales	SIGMA	Stock Volatility
LTAT	Total Liabilities/Total Assets	WCAPAT	Working Capital/Total Assets

Note: The table provides the description of the 36 numerical bankruptcy predictors.

U.S. SEC requires all public firms to file 10-K[1] at the end of each fiscal year. Our prediction models focus on the Management Discussion and Analysis (MD&A) section of 10-K. Since 1980, the SEC mandates public companies to include an MD&A section in the annual report. The section contains a narrative explanation of the firm's operations in a way that an average investor can understand. The rationale, according to SEC[2], is that

> "... a numerical presentation and brief accompanying footnotes may be insufficient for an investor to judge the quality of earnings and the likelihood that past performance is indicative of future performance. MD&A is intended to give the investor an opportunity to look at the company through the eyes of

[1] A complete 10-K consists of fourteen Items that provide a comprehensive summary of a company's business in the previous year. Common items include business description, financial performance, organization structure, executive compensation, equity, among others.

[2] Securities Act Release No. 6711.

management by providing both a short and long-term analysis of the business of the company."

The statement makes it clear that the SEC intends MD&A to serve as a qualitative disclosure for investors to make more accurate projections of future financial and operating results. Although studies in accounting and finance have conducted content analyses of MD&A (see for example Cole & Jones, 2005; Li, 2008; Loughran & McDonald, 2011), most focus on extracting the tone and readability of the section. These correlational studies do not resolve how, if at all, we can use this qualitative section for corporate bankruptcy prediction. In other words, can algorithms understand this section as the SEC intends investor to do?

We download all 10-K forms and its variants 10-K405, 10KSB, and 10KSB40 forms from the SEC Electronic Data Gathering, Analysis, and Retrieval (EDGAR) system. We link 10-K forms to CRSP and Compustat database. Firms with missing links are excluded from the analysis. The firms with missing links are real estate, non-operating, or asset-backed partnerships or trusts (Loughran & McDonald, 2011). For each linked 10-K filing, we remove the HTML tags, tables, and exhibits. We then extract the MD&A section using Perl scripts. This section usually appears as Item 7, but sometimes also appears in the annual shareholder letters attached at the end of 10-K. We consider both cases. For our study, we exclude all empty MD&A sections and those with fewer than 1000 characters.

4. Model

4.1. Overview and text pre-processing

Predictive models based on textual data pose challenges to the modeling process for several reasons. First, textual data are natural languages; they cannot be directly used as inputs in many mathematical models. We need to use natural language processing (NLP) to transform the textual data into numerical units that a mathematical model can understand. Second, typical textual databases are much greater in size compared to numerical datasets. Therefore, information extraction, which locates the most relevant pieces of information, becomes a crucial step in the modeling process. Third, coping with the 'curse of dimensionality' and using effective feature selection to overcome the curse is ever more crucial when using textual data as model inputs. The simplest text model is 'bag of words,' which treats each unique word as a feature; it can increase the number of variables to over tens of thousands (the size of vocabulary). Fourth, feature selection is compounded by the problem of synonymy in natural language. In our context, there are many ways for firms to convey that they are doing well or in trouble. For prediction models to be generalizable, we need to make use of the semantic rather than the syntactic of the text. In other words, a good model should *understand* the meaning of words.

With these challenges in mind, we now describe our modeling process. Fig. 2 summarizes our analysis procedure, including the database construction, data processing, feature extraction, and model training and evaluation. The flowchart's left branch shows how we analyze the numerical data following standard procedure in prior literature. We focus on how we process and analyze the text data (the right branch).

In the data processing stage, we transform the raw MD&A section from 10-K annual filings to clean plain-text documents in three steps: (1) We tokenize each MD&A section into individual words using the Natural Language Toolkit (NLTK) (Bird, Bird, & Loper, 2016); (2) We also use NLTK to lemmatize each word and remove the inflectional forms of words and return them to basic forms. For instance, *paid* and *paying* become *pay*; (3) We remove the low-frequency words and only include 20,000 most frequent words. Such filtering procedure is a common practice in natural language processing as it can help reduce the dimensionality of downstream statistical models.

The next stage, feature (variable) extraction, converts unstructured text to numerical representations. This stage is where traditional intelligent models and deep learning models diverge. We therefore combine the discussion of feature extraction with the description of models. Section 4.2 presents the model for textual data. We first describe how we use word embedding, a new deep learning layer for textual data, to extract meaning from the texts and turn words into real-valued vectors. Next, our model uses the outputs from the word embedding layer as inputs to generate bankruptcy predictions. We compare two deep learning model architectures: average embedding model and convolutional neural network. In Section 4.3, we briefly describe deep learning models for numerical data. Section 4.4 covers the implementation of the deep learning models. In Section 4.5, we show how traditional data mining models handle text features and describe these benchmark models.

4.2. Deep learning architectures for text

4.2.1. Word embedding

As noted earlier, a fundamental challenge in using natural language as predictors of future events is to understand the underlying semantics. This is because any word can have many synonyms; there are also infinite possible combinations of words that can express the same meaning. A model that focuses on learning the syntactic of language needs to deal with the large amount variations, each with its own parameter. A "bag-of-words" model is an example, in which case the discriminant power of each word in the vocabulary is computed. Such model often suffers from the 'curse of dimensionality' and runs the risks of being not generalizable. As the dimensionality (number of words in the vocabulary) increases, the amount of data required for the model to have acceptable variance (a component of predictive error) increases rapidly (Stone, 1985).

In this research, we make an important methodological advancement over the extant methods of using text to predict financial events. We start with the word embedding model (Mikolov et al., 2013, *aka* word2vec), which is among the hallmarks of the recent development of deep learning. It is based on a simple and old idea in linguistic: words with similar meanings tend to occur with similar neighbors. To operationalize this idea, the embedding model summarizes the contextual information of each word by predicting its surrounding context words using neural networks. The features used for such prediction are in a lower-dimensional vector space (the embedding space) that preserves as many properties of the original data as possible. For each word, the estimated coefficients for its features can represent the semantics of the word well. We can then represent the meaning of a word using a real-valued vector. As a result, the dimensionality of the textual model can be reduced significantly: from the size of the vocabulary to the dimension of the word vector.

Specifically, we adopt the skip-gram model (Mikolov et al., 2013b) to calculate the word embedding vectors. Our goal is to represent a word w using a d dimensional vector v_w. To achieve this, the skip-gram model first seeks to predict each word's surrounding words by maximizing the log probability:

$$\frac{1}{|V|}\sum_{t=1}^{|V|}\sum_{-k\le j\le k,\ j\neq 0}\log p(w_{t+j}|w_t) \tag{1}$$

where k is the "window size" of the context (for demonstration purpose, we let $k=5$), w_t is a word at location t, $|V|$ is the size

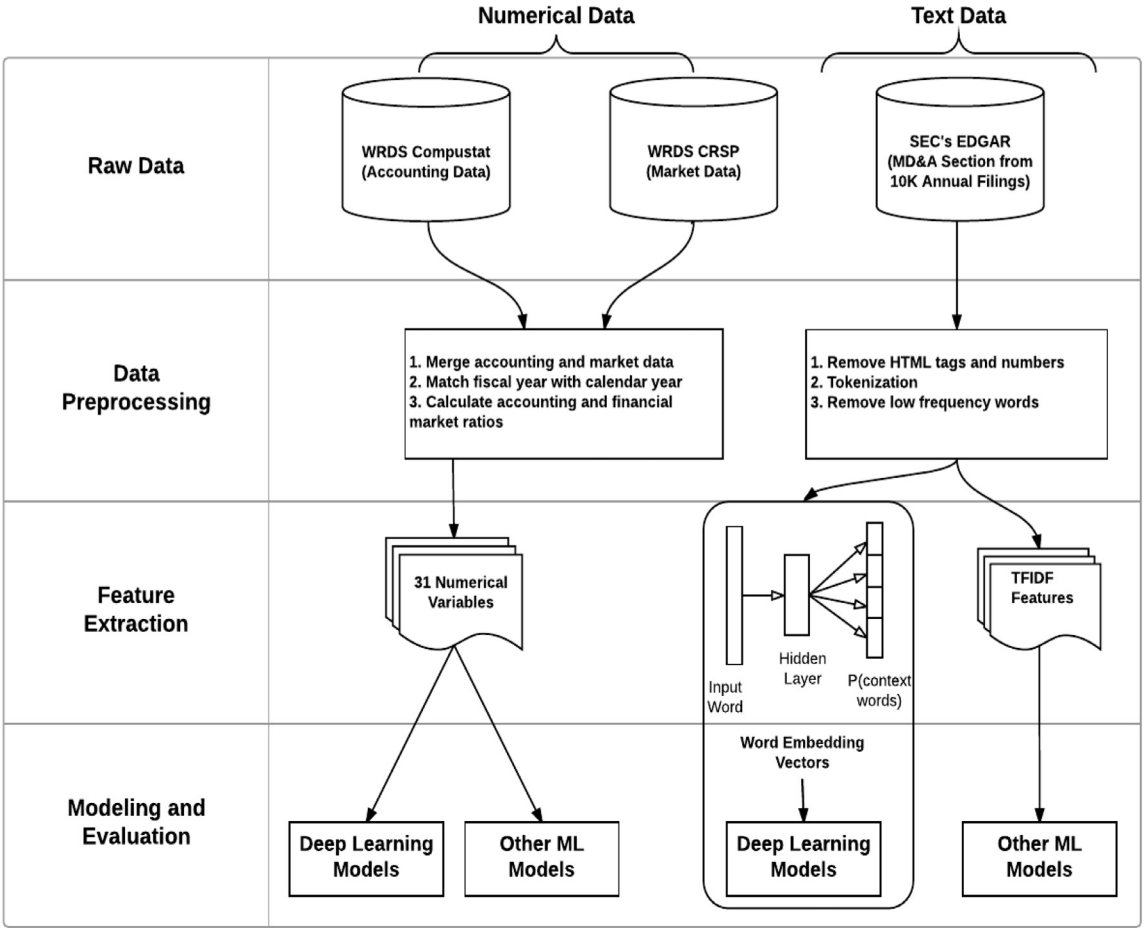

Fig. 2. Flow chart of analysis.

of the vocabulary. To train this prediction model, note that each word can be naturally represented using a $|V|$ dimensional one-hot row vector[3]. A single-hidden-layer neural network, parameterized by a $|V| \times d$ weight matrix W, first projects[4] a word w to a vector v_w in ^{d}, where v_w is simply the corresponding row in W. The network's output softmax layer, parameterized by a second $d \times |V|$ weight matrix W', uses the v_w as the input to predict the probability of observing each context word c in the context of w. The corresponding column in W' is denoted as v_c. That is:

$$p(c|w) = \frac{\exp\left(v_c^{\mathrm{T}} v_w\right)}{\sum_{c' \in C} \exp\left(v_{c'}^{\mathrm{T}} v_w\right)} \tag{2}$$

Putting it together, the log-likelihood of the entire model is computed by summing over all (w, c) combinations:

$$\begin{aligned} &\arg\max_{W,\ W'} \prod_{w \in V} \prod_{c \in c(w)} p\left(c|w;\ W,\ W'\right) \\ &= \arg\max_{W,\ W'} \sum_{\text{All } (w,c)} \log p\left(c|w; W,\ W'\right). \end{aligned} \tag{3}$$

In (3), $c \in c(w)$ is the set of all contexts for word w. The learning of word vectors v_ws is achieved when the log-likelihood is maximized.

A naïve estimation using iterative optimization techniques on the neural networks can be computationally impractical for a large collection of texts. We use an efficient approximation algorithm for the skip-gram model, known as negative sampling (Gutmann and Hyvärinen, 2012). It trains high-quality models without using any dense matrix multiplications. The difficulty of solving the word embedding model using an iterative optimization procedure lies in the computation of $\nabla p(c|w; W,\ W')$ from Eq. (3) due to the size of all the contexts C. Note that $\log p(c|w; W,\ W')$ in Eq. (3) can be written as

$$\begin{aligned} \log p\left(c|w; W,\ W'\right) &= \log \frac{\exp\left(v_c^{\mathrm{T}} v_w\right)}{\sum_{c' \in C} \exp\left(v_{c'}^{\mathrm{T}} v_w\right)} \\ &= \log \exp\left(v_c^{\mathrm{T}} v_w\right) - \log \sum_{c' \in C} \exp\left(v_{c'}^{\mathrm{T}} v_w\right). \end{aligned} \tag{4}$$

Negative sampling replaces (4) using the expression

$$\log \frac{1}{1 + \exp\left(-v_c^{\mathrm{T}} v_w\right)} + \sum_{i=1}^{n} \log \frac{1}{1 + \exp\left(v_{c_i'}^{\mathrm{T}} v_w\right)} \tag{5}$$

where c_i's are n negative samples, i.e., the words that never appear in the contexts of w, randomly generated from a "noise distribution". The idea is that if the model is trained correctly, it should be good at distinguishing correct (w, c) pairs (which we can observe from review data) from the randomly generated (w, c_i') pairs.

Given a vectorized presentation of each word, we effectively transform each document to a matrix of $n \times d$ document, where n is the document length. In practice, each MD&A document varies by length. We normalize each document to length $n = 7500$ by truncating longer documents and adding vectors of 0's to shorter

[3] A one-hot vector is a vector with a single 1 and the others 0. Since there are $|V|$ unique words, each word can be represented using a one-hot vector with a unique entry being 1.

[4] A one-hot row vector with the wth entry being 1 multiplying W outputs the wth row of W.

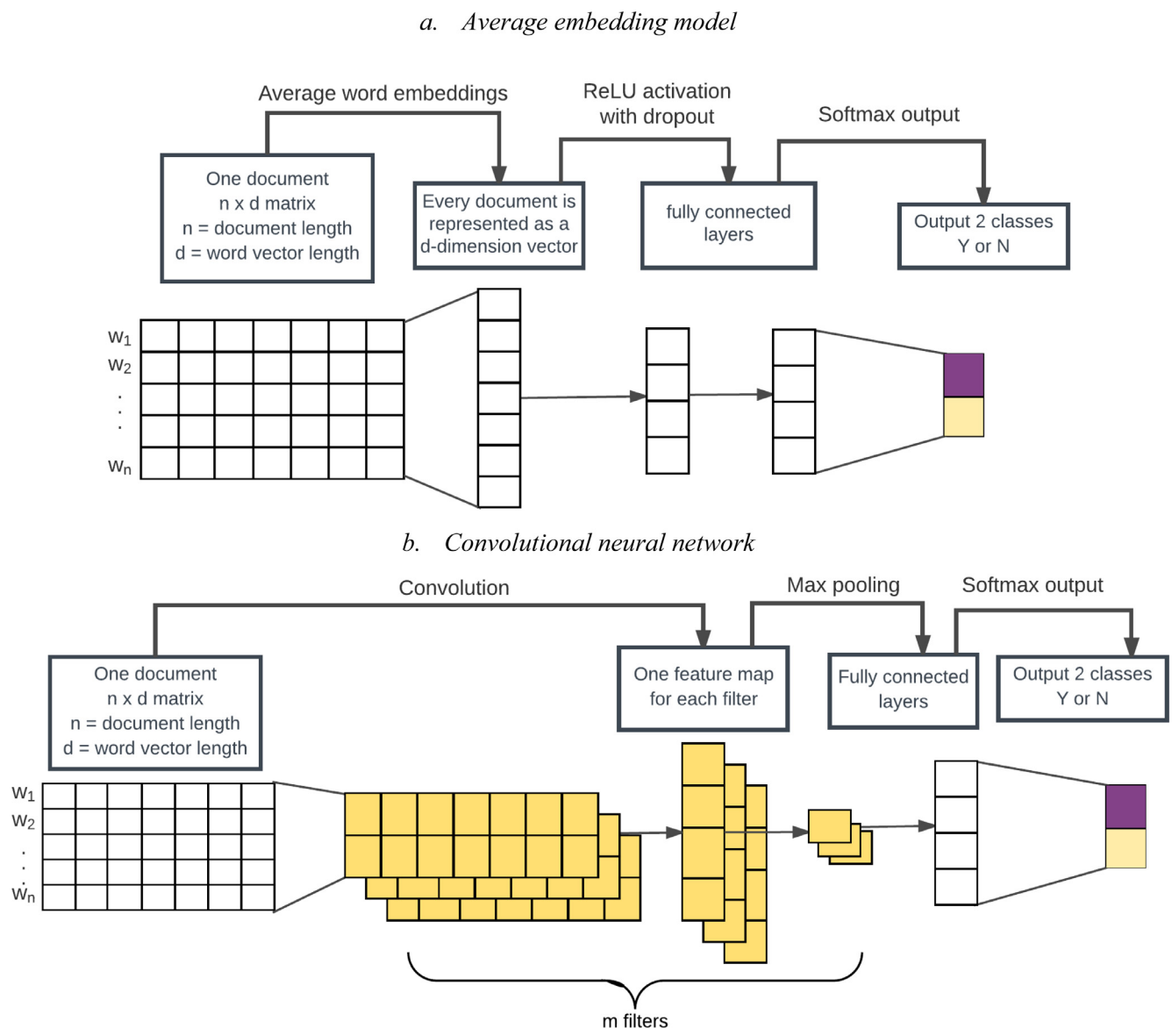

Fig. 3. Deep learning architectures.

documents. Such operation, called *padding*, is needed because the tensor (i.e., data represented as multi-dimensional arrays) used for neural network training must consist of matrices of the same dimension.

Thus far, firms' outcomes (bankruptcy or not) are not needed for the training of the embedding model. A key design decision in creating our deep learning system is defining its subsequent network architecture, or how the various processing layers can be added to the word embeddings to reach a prediction. We now compare two deep learning architectures that take pre-trained word embedding based on MD&A corpus as input features. The first is the average embedding model, and the second is a Convolutional neural network (CNN).

4.2.2. Average embedding model

Fig. 3(a) illustrates the average embedding model. Intuitively, we can view the average embedding model as identifying the most important latent themes (each as a dimension in the word vector) that can predict future bankruptcy. As its name suggests, this simple model architecture takes the average of every word vector dimension for all the word in a document. That is, the model calculates the mean of each column of the $n \times d$ document matrix and represents each document using a d-dimensional vector. Then, two layers of hidden neurons with rectified linear unit (ReLU) activation function are added. ReLU is defined as $f(x) = \max(0, x)$. Compared to conventionally used sigmoid neurons in bankruptcy prediction literature, ReLU layer is advantageous in that it can help train models faster and may yield better performance on unstructured data (Glorot, Bordes, & Bengio, 2011). Finally, a sigmoid output unit is added to classify the binary outcome.

4.2.3. Convolutional neural network

The convolutional neural network (CNN) architecture, shown in Fig. 3(b), is a variant of the model proposed by Kim (2014). The idea behind CNN is that it can train m convolving filters to detect local features. In our context, these features can be key phrases in MD&A that are helpful for bankruptcy prediction. The bag-of-words models in conventional methods, as we describe later in Section 4.3, are similar in a sense that they also assign higher weights to the most important words. However, a key difference here is that we apply CNN on the outputs of the embedding model. Thanks to the embedding model, semantically close phrases

(e.g., *growth* and *gain*) can share parameters in the lower dimensional vector space and lead to more generalizable models.

In a CNN, we define a filter B_f as a matrix $B_f \in \mathbb{R}^{h \times d}$. When the filter applies a convolution operation to a phrase of length h starting at the t-th word, it outputs a scalar:

$$c_t = \text{ReLU}(B_f \cdot W_{t:t+h-1} + b_0). \tag{6}$$

The $W_{t:t+h-1}$ matrix is the consecutive subset of rows from word vector t to word vector t+h-1 in the $n \times d$ embedding matrix of a document. We use $\cdot$ to denote the sum of the element-wise product of two matrices, and $b_0 \in \mathbb{R}$ is a bias parameter (intercept). ReLU is the aforementioned activation function. If we treat each filter as a key phrase detector, we can view the output scalar c_t as the extent the t-th word to (t+h-1)th word is semantically close to the phrase that the filter is trying to detect. In other words, each filter is trained to look for a set of predictive phrases that have similar meanings.

In our CNN architecture, we use 100 filters on our MD&A dataset. Since each filter is applied to every $h = 3$ consecutive word vectors, the convolution layer generates 100 n-dimensional feature map vectors, where n is the document length. When predicting the outcome of a firm, we are more interested in whether, rather than where, a key phrase occurs in the document. Hence, we apply a max pooling operation by taking the maximum value of each feature map vector, resulting in a 100-dimensional vector output. Finally, two layers of hidden neurons with ReLU activation function followed by a sigmoid output unit are added to classify the binary outcome.

4.3. Models for numerical data

Neural network has been a popular method for bankruptcy prediction using numerical data. We refer readers to Wong and Selvi (1998), Zhang et al. (1999) and the relevant studies in Table 1 for a review. We test three different models for benchmarking. The inputs of the models are variables from Table 3. In DL-1 Layer model, we add a single hidden layer of 4 neurons. In the DL-Deep model, we stack 4 layers of hidden units in the network, each layer constituting 4 neurons. The DL-Wide model constitutes a single hidden layer of 16 neurons. Such choice of these hidden parameters used in our model are consistent to the similar studies such as Lacher, Coats, Sharma, and Fant (1995) and Lee et al. (1996).

4.4. Model implementation

We implement our models using the Keras 2.0 package with TensorFlow backend. Our deep learning system is a feedforward model that maps inputs (numeric and text features) to a binary output (bankruptcy or not). Therefore, we use backpropagation algorithms to train the model. We use the cross-entropy as the loss function. Since the full training set is too large to fit in the memory, we use stochastic gradient descent (SGD) with the Adam update rule (Kingma & Ba, 2015). Because the objective function can be decomposed as a sum of subfunctions evaluated at different subsamples of data, SGD updates the parameters using only a subset (batch) of the training examples at a time. For each batch, all the word vectors, weight matrices, and bias vectors are updated. Hence, SGD is a much more efficient optimization method than standard backpropagation. We use a batch size of 32, and our models take less than 10 passes of the entire training sample (epochs) to train.

We employ three measures to prevent overfitting. First, we add L2 regularization to penalize the weights in the neural network[5]. In addition, we use the dropout technique (Srivastava, Hinton, Krizhevsky, Sutskever, & Salakhutdinov, 2014) by randomly omitting a subset of hidden units at each iteration of a training procedure. Lastly, we use early stopping (Bengio, 2012) by monitoring the validation set performance during the training and halt the training early when the performance stops improving.

[5] All regularization hyperparameters are chosen based on cross-validation performance of the training set.

4.5. Other benchmark models

4.5.1. Feature extraction

We now describe how we extract features for three benchmark data mining models: logistic regression, random forest, and support vector machine. For these models, the feature extraction for textual data starts with constructing the document-term matrix (DTM). We construct a DTM by converting each MD&A document as a single row in the matrix, and each of the top 20,000 words in the whole corpus is a column. Each entry in the DTM is the term-frequency (TF), which is the number of times a word (term) appears in a document. Using TF weights to represent the content of documents has a clear drawback: it grants high weights to words that are frequent across the board but lacks discriminative power. For instance, content-free words (*a, the, and*) and generic words (*finance, firm, report*) are usually not helpful in predicting the outcome of the firms. Therefore, to prioritize the important words specific to each MD&A document, we compute the term frequency-inverse document frequency (TF-IDF) for each word in each document. If a term t_j occurred in n_j of the N total documents, its TF-IDF weight in document i is calculated as

$$IDF(t_j) = -\log\left(\frac{n_j}{N}\right), \tag{7}$$

$$TF(t_{ij}) = \text{Number of times } t_j \text{ appers in document } i,$$

$$TF - IDF(t_{ij}) = TF(t_{ij}) \times IDF(t_j).$$

In other words, TF-IDF downweights a term's TF according to how frequently the term appears in the entire dataset. For the following benchmark models, each MD&A document is represented as a 20,000-dimensional TF-IDF weights vector.

4.5.2. Logistic regression

Logistic regression is one of the most popular prediction models in bankruptcy literature. The model assumes a logit link between the explanatory variables and the dichotomy default event. The model can be expressed as

$$P(Y_{i,t+1} = 1 | Y_{i,t-1} = 0, X_{i,t}) = \frac{e^{\beta_0 + \beta_1' X_{i,t}}}{1 + e^{\beta_0 + \beta_1' X_{i,t}}} \tag{8}$$

where $X_{i,t}$ is a covariate vector of time-varying firm-specific explanatory variables at time t, β is a vector of covariate effect parameters and β_0 is a scalar parameter. Mathematically, the logistic regression estimated using our sample is equivalent to a discrete hazard model (Shumway, 2001). We add an L1 penalty on the coefficients to prevent overfitting.

4.5.3. Random forest

Random forest (Schapire, Breiman, & Schapire, 2001) is a machine learning technique for both classification and regression. It is a variant of the Bagging (Breiman, 1996) ensemble learning method. For our classification problem, random forest constructs many decision tree classifiers trained on bootstrap replicates of original samples by randomly choosing k independent variables. The model outputs the classification class based on majority voting from the decision trees. Through randomization of both training samples and feature space, random forest can improve the generalization performance due to a reduction in variance while maintaining or slightly increasing bias.

4.5.4. Support vector machine

Support vector machine, developed by (Vapnik & Vapnik, 1998), has gained popularity in the bankruptcy prediction problem thanks to its generalization ability and distribution-free property towards underlying data set (Chandra, Ravi, & Bose, 2009). Plus, SVM has robust performance over a variety of text classification problems (Joachims, 1998). Simply put, SVM is a generalized linear model that finds an optimal hyperplane. The hyperplane maximizes the margin between itself and the nearest training examples while ensuring the accuracy of correct classification. The training examples that are closest to the hyperplane are called support vectors. All other examples are irrelevant for solving the classification problem. The support vectors are determined by solving a quadratic programming problem. For non-linear separable data, SVM uses non-linear kernel functions (RBF in our case) to transform training data to a higher dimensional feature space in which the data become more separable.

5. Empirical results

5.1. Model evaluation

It is crucial to choose a bankruptcy prediction model with an accurate out-of-sample prediction power. Out-of-sample prediction is also in line with the current BASEL III practice for default model validation purpose. In this work, we split the data into training and testing dataset and evaluate each model's out-of-sample prediction performance. We use two splitting method. First, we randomly partition the dataset by selecting 80% of the data as the training data set and the remaining 20% as the testing set. This method is commonly used in many of the previous research (e.g. Doumpos et al., 2017, Geng et al., 2015, du Jardin, 2016). Based on the nature of the data, we also test our models' performance by splitting the observations prior to 2007 as the training data and observations after 2008 as the testing data set. This method better resembles the forecasting scenarios in practice, and thus allows us to conduct a real assessment of the forecasting ability. For both partitioning methods, we report the model's out-of-sample AUC (area under the receiver operating characteristic curve), accuracy ratio, and cumulative decile-ranking to measure each model's prediction ability.

AUC is a popular measure of a model's overall discriminatory power (Bradley, 1997). Because bankruptcy is a rare event, using the classification accuracy to measure a model's performance can be misleading. This is because the classification accuracy score assumes that type I and type II errors are equally costly. In reality, the cost of false negatives is much heavier than that of false positives. Although it is possible to assign a higher cost to false negatives (du Jardin, 2016), such cost structure is still context specific. Also, decision-makers are interested more than a dichotomous bankruptcy prediction. In many contexts, the probability of bankruptcy can be used to construct credit portfolios or determine interest rates of loans (Hillegeist, Keating, Cram, & Lundstedt, 2004). Hence, AUC is a more flexible performance measure because it is calculated from the Receiver Operating Characteristic (ROC) curve. The ROC curve depicts the trade-off between the false positive rate and the true positive rate as the decision criterion (cutoff probability) varies. AUC, or the area under the ROC curve, can be used to evaluate a model's overall ability without assuming a relative cost structure. AUC score usually ranges from 0.5 to 1, with 0.5 indicating a baseline of random assignment of class labels, and 1 suggesting a perfect classification.

The accuracy ratio is another common gauge for corporate bankruptcy model evaluation (Engelmann, Hayden, & Tasche, 2003). It is calculated from the Cumulative Accuracy Profile (CAP), a concept similar to the ROC curve. The CAP tallies the percentage of true bankrupt firms included if we choose a varying percentage of observations using the sorted predicted probabilities generated by a model. In a baseline model that randomly assigns class labels, the CAP would be a straight line with slope one. The accuracy ratio of a prediction model is the difference in the area between the CAP of the model and the CAP of the baseline model. It captures the performance improvement of a prediction model compared with the baseline model. Ranging from 0 to 1, higher accuracy ratio value indicates better classification ability.

In addition, we also report the cumulative decile-ranking table on the testing dataset. We rank the company's predicted probabilities into deciles, where the top decile contains the companies with high default probability and the bottom decile contains firms with low default risk. The decile table is constructed by tabulating the cumulative percentage of actual bankruptcy firms in each decile. A high percentage in the high bankruptcy probability deciles implies better out-of-sample classification power.

5.2. One-year-ahead prediction performance

5.2.1. Prediction performance using textual data

Table 4 summarizes the out-of-sample prediction results using only the MD&A section of 10-K as predictor variables. We compare two deep learning architectures for text data with the three benchmark models: a linear model (logistic regression) and two non-linear models (SVM and random forest). In Table 4, DL-Embedding is the average embedding model described in Section 4.2.2, and DL-CNN is the convolutional neural network described in Section 4.2.3. We first note that all the implemented models, deep learning or not, can adequately use the textual information in MD&A for bankruptcy prediction. The AUC values are consistently above 0.7 (for a random model the AUC will be 0.5). Also, the top deciles of the predicted probabilities include at least 25% of the true bankruptcies (for a random model the value will be 10%).

Among the experimented models when we split the data randomly (Panel A), the DL-Embedding model has shown noticeably higher AUC value 0.784 compared to other models. The DL-CNN model performance is on par with the other benchmark models such as logistic regression, SVM, and random forest model, with the AUC values ranging from 0.711 to 0.716. Similarly, DL-Embedding is the only model with an accuracy ratio above 0.5. According to the decile ranking table at the bottom of Table 4, the DL-Embedding model can correctly predict 35.7% of future bankruptcy filings in the top decile and 55.9% in the top quintile (20%). When splitting the data by year, the DL-Embedding model also has the highest AUC value 0.760. The DL-CNN performed worse than the bench marking models in terms of AUC. Overall, Table 4 provides strong evidence that the average embedding model is the better deep learning architecture for bankruptcy prediction.

To formally test the performance difference between the average embedding model and the benchmark methods, we use the binomial test for algorithm comparison proposed by Salzberg (1997). Specifically, we compare the example-wise performance between the DL-Embedding model and the best-performing benchmark model (Logistic regression). In our test set, we count the number of examples for which the two algorithms give different results and denote the number as n. We then define successes (s) as the number of times that DL-Embedding got right and random forest got wrong. Under the null hypothesis that two algorithms have equal performance, we expect the probability of success in the binomial distribution to be 0.5. Hence, the p-value is equal to $\sum_{i=s}^{n} \frac{n!}{i!(n-i)!} 0.5^n$. The binomial test result on the test set performance formally confirms our findings shown in Table 4. With textual data only using the MD&A section of 10-K reports as

Table 4
One-year ahead out-of-sample performance using 10-K text.

Panel A: Random Split					
	DL-Embedding	DL-CNN	Logistic Regression	SVM	Random Forest
Accuracy ratio	**0.568**	0.428	0.434	0.422	0.433
AUC	**0.784**	0.714	0.717	0.711	0.716
1	**0.357**	0.250	0.297	0.297	0.321
2	0.559	0.440	0.487	0.499	0.464
3	0.714	0.559	0.594	0.570	0.595
4	0.821	0.738	0.736	0.724	0.690
5	0.881	0.809	0.807	0.795	0.833
6–10	1	1	1	1	1
Panel B: Split by Year					
	DL-Embedding	DL-CNN	Logistic Regression	SVM	Random Forest
Accuracy ratio	**0.521**	0.403	0.434	0.432	0.419
AUC	**0.760**	0.701	0.717	0.716	0.710
1	**0.424**	0.326	0.315	0.315	0.380
2	0.565	0.478	0.457	0.457	0.489
3	0.728	0.609	0.587	0.587	0.609
4	0.783	0.685	0.685	0.685	0.696
5	0.837	0.739	0.783	0.783	0.707
6–10	1.000	1.000	1.000	1.000	1.000

Note: The table reports the out-of-sample performance measures for the test set, including the accuracy ratio, AUC (area under the ROC curve), and the decile ranking. We use an 80-20 train-test split in Panel A and pre-2007/post-2008 splitting in Panel B. The predictors are the MD&A section of the 10 K filings. For the decile ranking, we sort firms in the testing sample equally into deciles based on their predicted default probabilities. The first decile (decile 1) contains firms with the highest predicted default probability, and the last five deciles (decile 6-10) include the firms with the lowest predicted default probability. We then tabulate the cumulative percentage of actual bankruptcy filings observed in each decile.

Table 5
One-year ahead out-of-sample performance using accounting and market data.

Panel A: Random Split						
	DL-1 Layer	DL-Deep	DL-Wide	Logistic Regression	SVM	Random Forest
Accuracy Ratio	0.633	0.603	0.597	0.616	0.619	0.636
AUC (%)	0.817	0.802	0.798	0.808	0.810	0.818
1	0.405	0.345	0.238	0.369	0.429	0.393
2	0.655	0.547	0.559	0.583	0.643	0.679
3	0.798	0.785	0.797	0.809	0.798	0.810
4	0.917	0.880	0.916	0.892	0.858	0.846
5	0.953	0.963	0.952	0.952	0.941	0.929
6–10	1	1	1	1	1	1
Panel B: Split by Year						
	DL-1 Layer	DL-Deep	DL-Wide	Logistic Regression	SVM	Random Forest
Accuracy Ratio	0.614	0.554	0.623	0.542	0.602	0.629
AUC (%)	0.807	0.777	0.811	0.771	0.801	0.814
1	0.587	0.413	0.457	0.467	0.533	0.467
2	0.717	0.576	0.707	0.685	0.750	0.728
3	0.804	0.837	0.804	0.750	0.783	0.783
4	0.848	0.891	0.870	0.804	0.815	0.815
5	0.880	0.935	0.891	0.859	0.870	0.826
6–10	1	1	1	1	1	1

Note: The table reports the out-of-sample performance measures for the test set, including the accuracy ratio, AUC (area under the ROC curve), and the decile ranking. We use an 80–20 train-test split in Panel A and pre-2007/post-2008 splitting in Panel B. The predictors are the accounting ratios and stock market data defined in Table 3.

predictors, the average embedding model significantly outperforms the logistic regression model in both test samples ($p < 0.001$). Similar comparisons between average embedding model and other benchmark models yield similar results.

5.2.2. Prediction performance using numeric data

Table 5 summarizes the prediction results using only the numerical variables listed in Table 3. We compare the three deep learning models for numerical data (described in Section 4.3) with the same benchmark model set as in Table 4, such as logistic regression, SVM, and random forest model. When splitting the train-test data randomly (Panel A), among the three deep learning models, the DL-1 Layer model delivers the highest AUC value of 0.817 and the accuracy ratio value of 0.633. The DL-1 Layer model also captures the most default events in its top decile and quintile. When the three benchmark models are included in the comparison, however, the deep learning models demonstrate similar performances as the benchmark models. For example, the random forest has the best AUC and accuracy ratio while the SVM has the highest top decile performance.

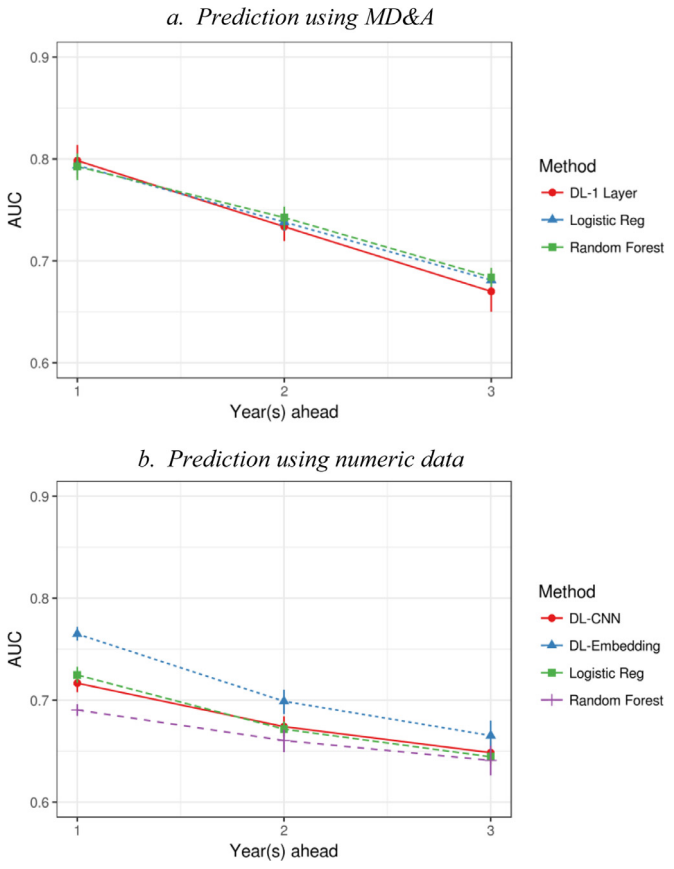

Fig. 4. Performance comparison with different forecasting horizons (10 × 10 CV).

The Salzberg binomial test confirms such comparable performances that the best-performing deep learning model (DL-1 Layer) is not significantly different from the random forest model ($p = 0.315$). Similar observations hold when we use the post-2008 data as the test set (Panel B). The DL-1 Layer has the best first-decile performance, but the random forest has the highest AUC. Overall, our results suggest that when numerical data are used as predictors, there is no compelling reason to use deep learning. Traditional data mining models, especially those that can find non-linear decision boundaries such as SVM and Random Forest, are equally capable.

5.3. Longer prediction horizons

It is critical to identify high default risk as early as possible in default risk management. Therefore, we extend our study to examine how the models' prediction performances change as we increase the prediction horizon to a longer time. Instead of mapping the firm outcomes to MD&A, market-based variables, and accounting-based ratios from 1-year prior to the event, we create new training datasets based on information from two or three years prior to the bankruptcy events. We further carry out a 10 × 10 cross-validation for each model to assess the variability of the model performance as different test samples are used. Fig. 4 depicts a summary of different models' prediction performance at different prediction horizons[6]. The error bars indicate the 95% confidence band. Fig. 4(a) shows that, consistent with the results from one-year-ahead prediction horizon, the average embedding model continues to outperform the logistic regression model and random forest model when using the MD&A data. When using market-based and accounting-based data, however, the best-performing deep learning model only delivers comparable results as the benchmark models. The conclusion holds across all the prediction horizons we studied. We also notice that the longer prediction horizon yields the inferior prediction performance (i.e., decreasing AUC values) and the smaller performance gaps between the four models. This is not surprising due to the loss of timely information used in prediction. Lastly, for the one-year-ahead and two-year-ahead prediction, the performance using numeric data is superior to that of using text data. However, for the 3-year-ahead prediction, their performance gap wanes.

[6] We pick the most representative models for visualization. DL-Deep and DL-Wide have similar performance as DL-1 Layer. SVM is excluded from the analysis because the computational time is prohibitive for 10x10 CV.

5.4. Prediction using both textual and numerical data

We now investigate the prediction performance when both textual data and numerical data are utilized. Although correlational evidence suggests that textual data contain additional informational value even when numerical data is available (Loughran & Mcdonald, 2011), whether a model can leverage such information for bankruptcy prediction is not clear. Fortunately, creating a deep learning model for mixed inputs is straightforward. We concatenate the final hidden layers (4 neurons) from the average embedding model with the hidden layer (4 neurons) from the DL-1 Layer model. Then, we connect the 8-neuron layer to a softmax output layer. Such architecture ensures that multiple levels of neural networks can first extract the best representation from the respective raw inputs. Using the higher-level learned representations, both the textual input data and the numerical input data can contribute to the prediction outcome. We can create a mixed input for logistic regression and random forest by concatenating the document-term matrix and the numeric input matrix side by side. The difference is that these traditional models can only combine the mixed inputs in their raw forms, without the extra layers of abstraction that deep learning can help find.

The prediction results of the experiment are presented in Table 6 and Fig. 5. For both train-test sample splitting methods, our deep learning model outperforms the logistic regression and random forest. Indeed, a simple concatenation of the last hidden layer from the DL-Embedding and the only hidden layer from the DL-1 Layer is sufficient for the training algorithm to find a model with the best out-of-sample performance. The 1-year-ahead out-of-sample performance of [DL-Embedding + DL-1 Layer] model yields an AUC value of 0.856 when using random split and 0.842 when splitting by year. In contrast, the AUC value of the logistic regression drops to 0.753 when using random split and 0.745 when splitting by year) when dealing with the mixed inputs. The AUC is even worse than the case with numerical data only (AUC 0.808 and 0.771, respectively), suggesting that when mixed inputs are used, the logistic regression model is incapable of self-selecting the most relevant features.

Figure 5 further confirms the superiority of the [DL-Embedding + DL-1 Layer] model over the logistic regression and random forest model. The comparison of the three corresponding ROC curves shows that the deep learning model dominates the two benchmark models across a wide spectrum of cutoff probabilities. The Salzberg binomial test also confirms that the deep learning model has delivered significantly better prediction performance compared to the two benchmark models ($p < 0.001$ for both tests). In summary, deep learning can effectively capture the most relevant features from the MD&A text to complement the numeric data.

Table 6
One-year ahead out-of-sample performance using 10-K text, accounting and market data.

Panel A: Random split			
	DL-Embedding + DL-1 layer	Logistic regression	Random forest
Accuracy Ratio	**0.712**	0.507	0.639
AUC (%)	**0.856**	0.753	0.819
1	**0.547**	0.369	0.511
2	0.725	0.571	0.677
3	0.891	0.666	0.807
4	0.926	0.725	0.890
5	0.937	0.832	0.913
6–10	1	1	1
Panel B: Split by year			
	DL-Embedding + DL-1 Layer	Logistic regression	Random forest
Accuracy Ratio	**0.685**	0.491	0.585
AUC (%)	**0.842**	0.745	0.793
1	**0.587**	0.326	0.446
2	0.750	0.554	0.62
3	0.826	0.641	0.728
4	0.870	0.728	0.837
5	0.880	0.848	0.902
6–10	1	1	1

Note: The table reports the out-of-sample performance measures for the test set, including the accuracy ratio, AUC (area under the ROC curve), and the decile ranking. We use an 80-20 train-test split in Panel A and pre-2007/post-2008 splitting in Panel B. The predictors are the 10-K text, accounting ratios and stock market data.

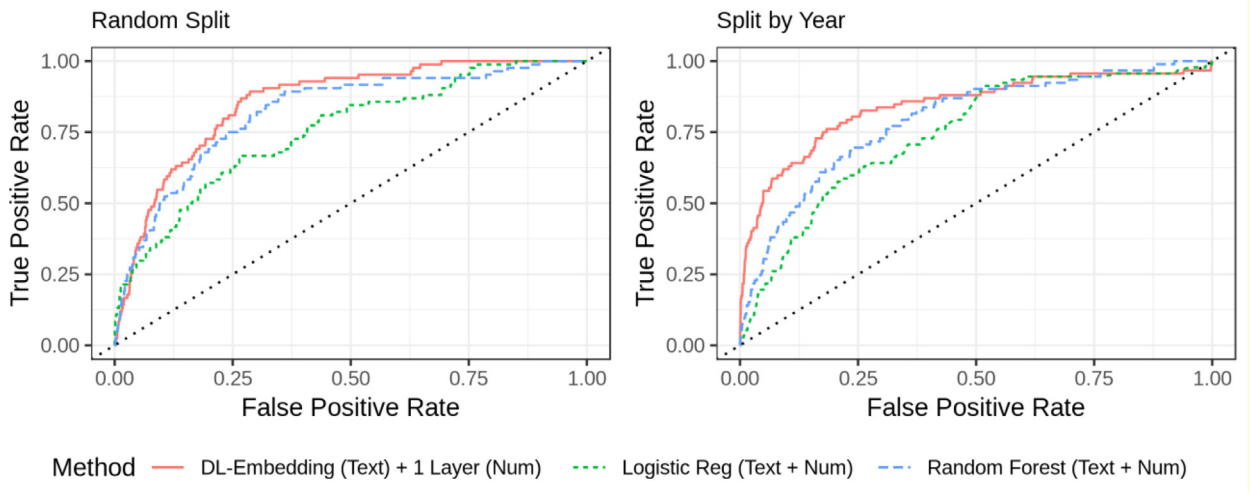

Fig. 5. Comparison of ROC curves using mixed inputs.

5.5. Important words from MD&A

The improved predictive performance of deep learning models comes with a cost – the poor interpretability. Large parameter space and the interaction between neurons prevent us from interpreting the model coefficients directly. To find which words in the MD&A section are important, we use the representation erasure method (Li, Monroe, & Jurafsky, 2017). Representation erasure is a general method for analyzing and interpreting decisions made by a black-box model. We erase individual words from the input corpus and observe how the model performance degenerates. If the model's AUC decreases by a large amount when we remove a particular word from the entire corpus, the model considers the word to be important. In practice, we replace the word index i prior to the $n \times d$ document representation (step 1 in Fig. 3) using 0, which is the padding token. We then calculate the importance score for word i as the difference between DL-Embedding model's AUC and the AUC of the same model but with the erased input.

We present 100 words with the highest importance scores in Table 7. We separate them into two groups by comparing their relative frequencies in bankruptcy firms and non-bankruptcy firms. Intuitively, we can interpret the important words with higher frequency in bankruptcy firms as words with negative meanings and vice versa. We also cross-check the words with two widely-used sentiment dictionaries: Loughran and McDonald (2011)'s financial sentiment dictionary and the MPQA Subjectivity Lexicon (Wilson, Wiebe, & Hoffmann, 2005). The words that appear in both dictionaries are marked in Table 7. Interestingly, many words that our model considers important are not in either dictionary. This suggests that bankruptcy prediction from text is more nuanced than sentiment analysis. In addition to the words related to firm's performance, we find many words in other factors such as capital structure (*repurchase*, *dividend, tranche*), strategy (*international, exit, focus*), internal and external stakeholders (*compensation, wages, costs*, *suppliers*) also demonstrate high importance in bankruptcy prediction.

Table 7
Important words in the MD&A text.

Non-bankruptcy Firms	Bankruptcy Firms
income, increase, increased, future, revenues, rate, intangible, profit, compensation, growth, tax, percentage, goodwill, cash, value, investment, **improved**, term, compared, economic, products, intangibles, changes, revenue, repurchases, outstanding, invested, repurchased, marketable, rates, repurchase, electronic, **strong**, expenditures, construction, maturity, imaging, credit, accounts, dividend, latest, excluding, international, bank, ebitda, holdings, suppliers, well, partners, long	**loss**, services, trust, initial, decrease, fees, sale, public, extraordinary, ended, structure, managed, measurements, inception, recourse, inventory, room, accordingly, expenses, serviced, approval, prime, **restated**, incurred, stores, indebtedness, secured, certificates, **discontinued**, affiliate, convertible, exit, tranche, servicing, focus, backed, announced, disposal, mortgage, joint, reset, aggregate, conversion, generated, production, received, costs, receivables, selling, wages

Note: The table lists the 100 most important words using the representation erasure method. The **bold** words are in Loughran and McDonald (2011)'s financial sentiment dictionary. The underscored words are in MPQA sentiment dictionary (Wilson et al., 2005).

6. Conclusion

As a large amount of unstructured data is injected into the market every day, investors, regulators, and researchers demand more intelligent models to digest such information. Motivated by the successful utilization of deep learning in areas such as computer vision and speech recognition, we introduce deep learning models for bankruptcy prediction using both structured (accounting-based and market-based) and unstructured (MD&A from 10-K filings) inputs. We construct a comprehensive dataset of U.S. public firms and show that recent advancements in neural networks can extract useful representations from financial texts for prediction. Deep learning lends itself particularly well to analyzing textual data, but the improvement on numerical data is limited compared with traditional data mining models. Moreover, deep learning can effectively integrate the incremental information from textual data with numeric information and achieve better prediction accuracy than using a single form of input.

Our findings have implications beyond identifying the best prediction model. The broader concern is with the utility and information value of firms' textual disclosures. Beaver (1966), a pioneer of bankruptcy prediction research, argues that accounting data should be evaluated in terms of their utility, which in turn can be defined in terms of predictive ability. Shumway (2001) introduced the market-based predictor variables for forecasting bankruptcy. Along the two popular mainstreams using market-based and accounting-based information to predict distress, many studies such as Campbell et al. (2008) propose modifications to those variables. Different from the seminal works aforementioned, we evaluate the predictive power of textual disclosures in annual reports — a brand new data source with wide circulation — to test. Our results provide direct, large-sample evidence of textual disclosure's information value. The AUC values for data models for the 1-year ahead forecast, depending on the model we use, are between 0.711 and 0.784. When the forecasting horizon is longer (3-year), the predictive value of unstructured text is comparable to audited accounting ratios and market data. Therefore, integrating textual disclosures into risk models could provide great insights.

Our study has several limitations which prompt future investigations. First, deep learning models, like many other artificial intelligence systems that use unstructured data, is difficult to interpret. Opening the black box of a trained neural network could shed light on the nature of the disclosure that leads to future bankruptcy. Second, our study only uses MD&A as the sole source of unstructured data. Future research can investigate the value of other channels such as news reports and user-generated content. Third, many up-and-coming deep learning models are not explored in our study. Techniques such as Long Short Term Memory (LSTM) networks can exploit the time series structure in the data for prediction. Lastly, in our empirical analysis, we adopt the Area under the ROC curve (AUC) and the decile-ranking table to evaluate our model's bankruptcy prediction performance. Future research can further explore other measures such as H-measure (Hand, 2009) or Kolmogorov-Smirnov goodness-of-fit test statistics. We view our study as a potentially useful first step in applying deep learning to predict economic outcomes using large-scaled text data.

Acknowledgment

We are grateful for helpful comments from the editor, Roman Slowinski, and three anonymous reviewers. We also thank Peng Dong for his research assistance.

Appendix

In this appendix, we explain how we construct each candidate predictive variable using the CRSP and/or COMPUSTAT data items. EXCESS RETURN is a firm's log excess return on its equity relative to that on the S&P 500 index. SIGMA is the standard deviation of the daily stock return observed over the previous three months. PRICE is the equity price per share truncated from the above at the value of $15 and then takes the logarithm. MB is the ratio of the market equity to the adjusted book equity to which we add a 10% difference between the market equity and book equity. All series are available to investors in real time. Below we provide details for the other numeric variables.

ACTLCT=ACT/LCT; APSALE=AP/SALE; CASHAT=CHE/AT;
CASHMTA= CHE/(PRICE*SHROUT+LT+MIB);
CHAT=CH/AT; CHLCT=CH/LCT; (EBIT+DP)/AT=(EBIT+DP)/AT;
EBITAT=EBIT/AT; EBITSALE=EBIT/SALE;
FAT=(DLC+0.5*DLTT)/AT;
INVCHINVT=INVCH/INVT; INVTSALE=INVT/SALE;
(LCT-CH)/AT=(LCT-CH)/AT; LCTAT=LCT/AT; LCTLT=LCT/LT;
LCTSALE= LCT/SALE; LTAT=LT/AT;
LTMTA= LT /(PRICE*SHROUT+LT+MIB); LOG(AT)= log(AT);
LOG(SALE)= log(abs(SALE));
NIAT=NI/AT; NIMTA=NI/(PRICE*SHROUT+LT+MIB);
NISALE=NI/SALE;
OIADPAT=OIADP/AT; OIADPSALE=OIADP/SALE; QALCT= (ACT – INVT)/LCT;
REAT=RE/AT; RELCT=RE/LCT; RSIZE= log(PRICE*SHROUT/TOTVAL);
SALEAT=SALE / AT; SEQAT=SEQ/AT; WCAPAT=WCAP/AT.

References

Agarwal, S., Chen, V. Y. S., & Zhang, W. (2016). The information value of credit rating action reports: A textual analysis. *Management Science, 62*(8), 2218–2240 INFORMS. doi:10.1287/mnsc.2015.2243.

Alfaro, E., García, N., Gámez, M., & Elizondo, D. (2008). Bankruptcy forecasting: An empirical comparison of AdaBoost and neural networks. *Decision Support Systems, 45*(1), 110–122 Elsevier.

Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The Journal of Finance, 23*(4), 589–609 Wiley Online Library.

Beaver, W. H. (1966). Financial ratios as predictors of failure. *Journal of Accounting Research, 4*, 71 JSTOR. doi:10.2307/2490171.

Bengio, Y. (2012). Practical recommendations for gradient-based training of deep architectures. *Neural networks: Tricks of the trade* (pp. 437–478) Springer.

Bernanke, B. S. (1981). Bankruptcy, liquidity, and recession. *The American Economic Review, 71*(2), 155–159.

Bird, S., Bird, S., & Loper, E. (2016). NLTK : The natural language toolkit NLTK : The Natural Language Toolkit. *Proceedings of the COLING/ACL on Interactive presentation sessions* (pp. 69–72) (March), Association for Computational Linguistics. doi:10.3115/1225403.1225421.

Bose, I., & Pal, R. (2006). Predicting the survival or failure of click-and-mortar corporations: A knowledge discovery approach. *European Journal of Operational Research, 174*(2), 959–982 Elsevier.

Boyacioglu, M. A., Kara, Y., & Baykan, Ö. K. (2009). Predicting bank financial failures using neural networks, support vector machines and multivariate statistical methods: A comparative analysis in the sample of savings deposit insurance fund (SDIF) transferred banks in Turkey. *Expert Systems with Applications, 36*(2), 3355–3366 Elsevier.

Bozanic, Z., & Thevenot, M. (2015). Qualitative disclosure and changes in sell-side financial analysts' information environment. *Contemporary Accounting Research, 32*(4), 1595–1616 Wiley Online Library.

Bradley, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern recognition, 30*(7), 1145–1159 Elsevier.

Breiman, L. (1996). Bagging predictors. *Machine Learning, 24*(2), 123–140 Springer. doi:10.1007/BF00058655.

Bryan, S. (1997). Incremental information content of required disclosures contained in management discussion and analysis. *Accounting Review, 72*(2), 285–301.

Calabrese, R., Degl'Innocenti, M., & Osmetti, S. A. (2017). The effectiveness of TARP-CPP on the US banking industry: A new copula-based approach. *European Journal of Operational Research, 256*(3), 1029–1037 Elsevier.

Campbell, J. L., Chen, H., Dhaliwal, D. S., Lu, H., & Steele, L. B. (2014). The information content of mandatory risk factor disclosures in corporate filings. *Review of Accounting Studies, 19*(1), 396–455 Springer.

Campbell, J. Y., Hilscher, J., & Szilagyi, J. (2008). In search of distress risk. *The Journal of Finance, 63*(6), 2899–2939 Wiley Online Library.

Cecchini, M., Aytug, H., Koehler, G. J., & Pathak, P. (2010). Making words work: Using financial text as a predictor of financial events. *Decision Support Systems, 50*(1), 164–175 Elsevier B.V.. doi:10.1016/j.dss.2010.07.012.

Chandra, D. K., Ravi, V., & Bose, I. (2009). Failure prediction of dotcom companies using hybrid intelligent techniques. *Expert Systems with Applications, 36*(3), 4830–4837 Elsevier.

Chauhan, N., Ravi, V., & Chandra, D. K. (2009). Differential evolution trained wavelet neural networks: Application to bankruptcy prediction in banks. *Expert Systems with Applications, 36*(4), 7659–7665 Elsevier.

Chen, M.-Y. (2011). Predicting corporate financial distress based on integration of decision tree classification and logistic regression. *Expert Systems with Applications, 38*(9), 11261–11272 Elsevier.

Chen, N., Ribeiro, B., Vieira, A. S., Duarte, J., & Neves, J. C. (2011). A genetic algorithm-based approach to cost-sensitive bankruptcy prediction. *Expert Systems with Applications, 38*(10), 12939–12945 Elsevier.

Cole, C. J., & Jones, C. L. (2005). Management discussion and analysis: A review and implications for future research. *Journal of Accounting Literature, 24*, 135 Elsevier BV.

De Andrés, J., Lorca, P., de Cos Juez, F. J., & Sánchez-Lasheras, F. (2011). Bankruptcy forecasting: A hybrid approach using Fuzzy c-means clustering and Multivariate Adaptive Regression Splines (MARS). *Expert Systems with Applications, 38*(3), 1866–1875 Elsevier.

Demyanyk, Y., & Hasan, I. (2010). Financial crises and bank failures: A review of prediction methods. *Omega, 38*(5), 315–324 Elsevier.

Ding, A. A., Tian, S., Yu, Y., & Guo, H. (2012). A class of discrete transformation survival models with application to default probability prediction. *Journal of the American Statistical Association, 107*(499), 990–1003 Taylor & Francis.

Dorfleitner, G., Priberny, C., Schuster, S., Stoiber, J., Weber, M., de Castro, I., & Kammler, J. (2016). Description-text related soft information in peer-to-peer lending - Evidence from two leading European platforms. *Journal of Banking and Finance, 64*, 169–187 Elsevier. doi:10.1016/j.jbankfin.2015.11.009.

Doumpos, M., Andriosopoulos, K., Galariotis, E., Makridou, G., & Zopounidis, C. (2017). Corporate failure prediction in the european energy sector: a multicriteria approach and the effect of country characteristics. *European Journal of Operational Research, 262*(1), 347–360.

Engelmann, B., Hayden, E., & Tasche, D. (2003). Measuring the discriminative power of rating systems. *Risk*, 82–86 January: January.

Etemadi, H., Rostamy, A. A. A., & Dehkordi, H. F. (2009). A genetic programming model for bankruptcy prediction: Empirical evidence from Iran. *Expert Systems with Applications, 36*(2), 3199–3207 Elsevier.

Geng, R., Bose, I., & Chen, X. (2015). Prediction of financial distress: An empirical study of listed Chinese companies using data mining. *European Journal of Operational Research, 241*(1), 236–247 Elsevier.

Glorot, X., Bordes, A., & Bengio, Y. (2011). Deep sparse rectifier neural networks. *AISTATS, 15*, 275.

Gutmann, M. U., & Hyvärinen, A. (2012). Noise-contrastive estimation of unnormalized statistical models, with applications to natural image statistics. *Journal of Machine Learning Research, 13*, 307–361.

Hand, D. J. (2009). Measuring classifier performance: Acoherent alternative to the area under the ROC curve. *Machine Learning, 77*(1997), 103–123 Springer. doi:10.1007/s10994-009-5119-5.

Hillegeist, S. A., Keating, E. K., Cram, D. P., & Lundstedt, K. G. (2004). Assessing the probability of bankruptcy. *Review of Accounting Studies, 9*(1), 5–34 Springer. doi:10.1023/B:RAST.0000013627.90884.b7.

Hofman, J. M., Sharma, A., & Watts, D. J. (2017). Prediction and explanation in social systems. *Science, 355*(6324), 486–488 American Association for the Advancement of Science.

Huang, S.-M., Tsai, C.-F., Yen, D. C., & Cheng, Y.-L. (2008). A hybrid financial analysis model for business failure prediction. *Expert Systems with Applications, 35*(3), 1034–1040 Elsevier.

Ioannidis, C., Pasiouras, F., & Zopounidis, C. (2010). Assessing bank soundness with classification techniques. *Omega, 38*(5), 345–357 Elsevier.

Jardin, du (2015). Bankruptcy prediction using terminal failure processes. *European Journal of Operational Research, 242*(1), 286–303 Elsevier.

Jardin, du (2016). A two-stage classification technique for bankruptcy prediction. *European Journal of Operational Research, 254*(1), 236–252 Elsevier.

Joachims, T. (1998). Text categorization with suport vector machines: learning with many relevant features. *Proceedings of the 10th European conference on machine learning ECML '98* (pp. 137–142) Springer. doi:10.1007/BFb0026683.

Kim, M.-J., & Kang, D.-K. (2010). Ensemble with neural networks for bankruptcy prediction. *Expert Systems with Applications, 37*(4), 3373–3379 Elsevier.

Kim, Y. (2014). "Convolutional neural networks for sentence classification", pp. 1746–1751 doi: 10.3115/v1/D14-1181.

Kingma, D., & Ba, J. (2015). Adam: A method for stochastic optimization. In *Proceedings of the 3rd international conference on learning representations (ICLR2015)*.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. Proceedings of the *advances in neural information processing systems* (pp. 1097–1105).

Kumar, P., & Ravi, V. (2007). Bankruptcy prediction in banks and firms via statistical and intelligent techniques - A review. *European Journal of Operational Research, 180*(1), 1–28. doi:10.1016/j.ejor.2006.08.043.

Lacher, R. C., Coats, P. K., Sharma, S. C., & Fant, L. F. (1995). A neural network for classifying the financial health of a firm. *European Journal of Operational Research, 85*(1), 53–65 Elsevier. doi:10.1016/0377-2217(93)E0274-2.

Lang, M., & Stice-Lawrence, L. (2015). Textual analysis and international financial reporting: Large sample evidence. *Journal of Accounting and Economics, 60*(2), 110–135 Elsevier.

Le Cun, Y., Bengio, Y., Hinton, G., LeCun, Y., Bengio, Y., Hinton, G., Le Cun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature, 521*(7553), 436–444. doi:10.1038/nature14539.

Lee, K. C., Lee, K. C., Han, I., Han, I., Kwon, Y., & Kwon, Y. (1996). Hybrid neural network models for bankruptcy predictions. *Decision Support Systems, 18*(1), 63–72.

Lehmann, B. (2003). Is it worth the while? The relevance of qualitative information in credit rating. EFMA 2003 Helinski Meetings. Available at SSRN: https://ssrn.com/abstract=410186.

Li, F. (2008). Annual report readability, current earnings, and earnings persistence. *Journal of Accounting and Economics, 45*(2), 221–247 Elsevier. doi:10.1016/j.jacceco.2008.02.003.

Li, H., & Sun, J. (2009). Hybridizing principles of the Electre method with case-based reasoning for data mining: Electre-CBR-I and Electre-CBR-II. *European Journal of Operational Research, 197*(1), 214–224 Elsevier.

Li, H., Sun, J., & Sun, B.-L. (2009). Financial distress prediction based on OR-CBR in the principle of k-nearest neighbors. *Expert Systems with Applications, 36*(1), 643–659 Elsevier.

Li, J., Monroe, W., & Jurafsky, D. (2017). "Understanding neural networks through representation erasure", arXiv preprint arXiv:1612.08220.

Li, Z., Crook, J., & Andreeva, G. (2014). Chinese companies distress prediction: An application of data envelopment analysis. *Journal of the Operational Research Society, 65*, 466–479 3Springer.

Liang, D., Lu, C.-C., Tsai, C.-F., & Shih, G.-A. (2016). Financial ratios and corporate governance indicators in bankruptcy prediction: A comprehensive study. *European Journal of Operational Research, 252*(2), 561–572 Elsevier.

Loughran, T., & Mcdonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *Journal of Finance, 66*(1), 35–65. doi:10.1111/j.1540-6261.2010.01625.x.

Mayew, W. J., Sethuraman, M., & Venkatachalam, M. (2015). MD&A disclosure and the firm's ability to continue as a going concern. *Accounting Review, 90*(4), 1621–1651. doi:10.2308/accr-50983.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). "Distributed representations of words and phrases and their compositionality", pp. 1–9 doi: 10.1162/jmlr.2003.3.4-5.951.

Olson, D. L., Delen, D., & Meng, Y. (2012). Comparative analysis of data mining methods for bankruptcy prediction. *Decision Support Systems, 52*(2), 464–473 Elsevier.

Premachandra, I. M., Bhabra, G. S., & Sueyoshi, T. (2009). DEA as a tool for bankruptcy assessment: A comparative study with logistic regression technique. *European Journal of Operational Research, 193*(2), 412–424 Elsevier.

Psillaki, M., Tsolas, I. E., & Margaritis, D. (2010). Evaluation of credit risk based on firm performance. *European Journal of Operational Research, 201*(3), 873–881 Elsevier.

Salzberg, S. (1997). On comparing classifiers : Pitfalls to avoid and a recommended approach. *Data Mining and Knowledge Discovery, 328*, 317–328. doi:10.1023/A:1009752403260.

Sánchez-Lasheras, F., de Andrés, J., Lorca, P., & de Cos Juez, F. J. (2012). A hybrid device for the solution of sampling bias problems in the forecasting of firms' bankruptcy. *Expert Systems with Applications, 39*(8), 7512–7523 Elsevier.

Schapire, R. E., Breiman, L., & Schapire, R. E. (2001). Random forests. *Machine Learning, 45*(1), 5–32. doi:10.1023/A:1010933404324.

Schönbucher, P. J. (2003). *Credit derivatives pricing models: Models, pricing, and implementation*. John Wiley & Sons.

Serrano-Cinca, C., & GutiéRrez-Nieto, B. (2013). Partial least square discriminant analysis for bankruptcy prediction. *Decision Support Systems, 54*(3), 1245–1255 Elsevier.

Shumway, T. (2001). Forecasting bankruptcy more accurately: A simple hazard model. *The Journal of Business, 74*(1), 101–124 JSTOR.

Srivastava, N., Hinton, G. E., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research, 15*(1), 1929–1958.

Stone, C. (1985). Additive regression and other nonparametric models. *The Annals of Statistics, 13*(2), 689–705 JSTOR. doi:10.1214/aos/1176348654.

Sueyoshi, T., & Goto, M. (2009). DEA–DA for bankruptcy-based performance assessment: Misclassification analysis of Japanese construction industry. *European Journal of Operational Research, 199*(2), 576–594 Elsevier.

Tian, S., Yu, Y., & Guo, H. (2015). Variable selection and corporate bankruptcy forecasts. *Journal of Banking and Finance, 52*, 89–100 Elsevier B.V.. doi:10.1016/j.jbankfin.2014.12.003.

Vapnik, V. N., & Vapnik, V. (1998). *Statistical learning theory*: 1. New York: Wiley.

Wanke, P., Barros, C. P., & Faria, J. R. (2015). Financial distress drivers in Brazilian banks: A dynamic slacks approach. *European Journal of Operational Research, 240*(1), 258–268 Elsevier.

Wilson, R. L., & Sharda, R. (1994). Bankruptcy prediction using neural networks. *Decision support systems, 11*(5), 545–557 Elsevier.

Wilson, T., Wiebe, J., & Hoffmann, P. (2005). Recognizing contextual polarity in phrase-level sentiment analysis. *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing - HLT '05* (pp. 347–354) Association for Computational Linguistics. doi:10.3115/1220575.1220619.

Wong, B. K., & Selvi, Y. (1998). Neural network applications in finance: A review and analysis of literature (1990–1996). *Information & Management, 34*(3), 129–139. doi:10.1016/S0378-7206(98)00050-0.

Yeh, C.-C., Chi, D.-J., & Hsu, M.-F. (2010). A hybrid approach of DEA, rough set and support vector machines for business failure prediction. *Expert Systems with Applications, 37*(2), 1535–1541 Elsevier.

Zhang, G., Hu, M. Y., Patuwo, E. B., & Indro, D. C. (1999). Artificial neural networks in bankruptcy prediction: Generalframework and cross-validationanalysis. *European Journal Of Operational Research, 116*(1), 16–32 Elsevier.