WORKING PAPER

DANMARKS NATIONALBANK

15 NOVEMBER 2018 — NO. 130

Predicting Distresses using Deep Learning of Text Segments in Annual Reports

Casper Hansen
c.hansen@di.ku.dk
UNIVERSITY OF COPENHAGEN

Christian Hansen
chrh@di.ku.dk
UNIVERSITY OF COPENHAGEN

Rastin Matin
rma@nationalbanken.dk
DANMARKS NATIONALBANK

Pia Mølgaard
pim@nationalbanken.dk
DANMARKS NATIONALBANK

WORKING PAPER — DANMARKS NATIONALBANK
15 NOVEMBER 2018 — NO. 130

Predicting Distresses using Deep Learning of Text Segments in Annual Reports

Abstract
Corporate distress models typically only employ the numerical financial variables in the firms' annual reports. We develop a model that employs the unstructured textual data in the reports as well, namely the auditors' reports and managements' statements. Our model consists of a convolutional recurrent neural network which, when concatenated with the numerical financial variables, learns a descriptive representation of the text that is suited for corporate distress prediction. We find that the unstructured data provides a statistically significant enhancement of the distress prediction performance, in particular for large firms where accurate predictions are of the utmost importance. Furthermore, we find that auditors' reports are more informative than managements' statements and that a joint model including both managements' statements and auditors' reports displays no enhancement relative to a model including only auditors' reports. Our model demonstrates a direct improvement over existing state-of-the-art models.

Resume
Konkursmodeller anvender typisk kun de numeriske finansielle variable i virksomhedernes årsregnskaber. Vi udvikler en model, der også anvender de ustrukturerede data i regnskaberne, nærmere bestemt revisorerklæringerne og ledelsesrapporterne. Vores model består af et convolutional recurrent neuralt netværk der, når det sammenkædes med de numeriske finansielle variable, lærer en deskriptiv repræsentation af teksten, der er velegnet til at forudsige virksomhedskonkurser. Vi finder, at de ustrukturerede data bidrager til en statistisk signifikant forbedring af konkursforudsigelsernes nøjagtighed, i særdeleshed for store virksomheder, hvor nøjagtige forudsigelser er særligt vigtige. Endvidere finder vi, at revisorerklæringerne er mere informative end ledelsesrapporterne og at en model, der både inkluderer ledelsesrapporterne og revisorerklæringerne, ikke udviser en forbedring relativt til en model, der kun inkluderer revisorerklæringerne. Vores model udviser en direkte forbedring i forhold til nuværende state-of-the-art modeller.

Key words

Credit risk; Risk management

JEL classification
C45, C55, G17, G33

Acknowledgements
The authors are grateful to Benjamin Christoffersen for helpful comments. Casper Hansen and Christian Hansen are financially supported by the Innovation Fund Denmark through the DABAI project.

The authors alone are responsible for any remaining errors.

Predicting Distresses using Deep Learning of Text Segments in Annual Reports*

Casper Hansen

Department of Computer Science, University of Copenhagen, DK-2100 Copenhagen Ø, Denmark, c.hansen@di.ku.dk

Christian Hansen

Department of Computer Science, University of Copenhagen, DK-2100 Copenhagen Ø, Denmark, chrh@di.ku.dk

Rastin Matin

Danmarks Nationalbank, DK-1093 Copenhagen, Denmark, rma@nationalbanken.dk

Pia Mølgaard

Danmarks Nationalbank, DK-1093 Copenhagen, Denmark, pim@nationalbanken.dk

Tuesday 13th November, 2018

Abstract

Corporate distress models typically only employ the numerical financial variables in the firms' annual reports. We develop a model that employs the unstructured textual data in the reports as well, namely the auditors' reports and managements' statements. Our model consists of a convolutional recurrent neural network which, when concatenated with the numerical financial variables, learns a descriptive representation of the text that is suited for corporate distress prediction. We find that the unstructured data provides a statistically significant enhancement of the distress prediction performance, in particular for large firms where accurate predictions are of the utmost importance. Furthermore, we find that auditors' reports are more informative than managements' statements and that a joint model including both managements' statements and auditors' reports displays no enhancement relative to a model including only auditors' reports. Our model demonstrates a direct improvement over existing state-of-the-art models.

Keywords: corporate default prediction, discrete hazard models, convolutional neural networks, recurrent neural networks

JEL: C45, C55, G17, G33

*We are grateful to Benjamin Christoffersen for helpful comments. Casper Hansen and Christian Hansen are financially supported by the Innovation Fund Denmark through the DABAI project.

1 Introduction

Statistical corporate distress prediction is a binary classification task that was pioneered by Altman (1968) and Ohlson (1980) among others. They used a limited number of financial ratios as input and employed simplistic models such as linear discriminant analysis and logistic regression for the classification, where the financial ratios enter the model in a linear combination. Since then a range of advanced statistical methods ("machine learning") have been applied to the problem such as gradient boosting (e.g. Caruana and Niculescu-Mizil (2006)) and neural networks (e.g. Atiya (2001), Tsai and Wu (2008)) including convolutional neural networks (Hosaka (2019)). Traditionally, distress models have only employed the numerical financial variables of the firms' annual reports, i.e. structured data. However, annual reports also contain unstructured data in the form of text segments (auditors' reports and managements' statements), which is potentially a rich source of information for distress prediction.

Since 2013 Danish regulators have required firms to provide annual reports in the open data standard for financial reporting known as eXtensible Business Reporting Language (XBRL) from which these two text segments can be easily extracted. Motivated by recent advances within natural language processing, we propose a deep learning approach for predicting corporate distresses that incorporates these text segments in addition to numerical financial variables. Using annual reports of corporate firms in Denmark from 2013 to 2016, corresponding to a total of 278 047 firm years, our tests reveal that the auditors' reports, and to a lesser extent the managements' statements, increase the prediction accuracy compared to common state-of-the-art baseline classifiers that are based solely on structured data. This demonstrates that the unstructured data contains a signal that can enhance corporate distress prediction models. The readily availability of the data makes this study particularly valuable as current state-of-the-art can be augmented straightforwardly.

We investigate a model employing auditors' reports, a model employing managements' statement, and a model employing both auditors' reports and managements' statements. For each of the three models, we first apply standard preprocessing techniques to the text followed by pattern extraction and recognition by using a convolutional recurrent neural network. The output from the convolutional recurrent neural network is then concatenated with numerical financial variables and the final model is estimated using two fully-connected layers. Our models further utilize an attention mechanism, which increases the model interpretability by being able to highlight words that are important for the final prediction.

We compare performance of these three models to three competitive distress prediction models based solely on the structured data: A logistic regression, gradient boosted trees, and a neural network with the same architecture as the network that employs text. The models employing text outperform all other models. Specifically, we find that including the auditors' reports, managements' statements, and both text segments

in the neural network improves prediction accuracy measured by AUC by 1.9, 1.1, and 1.8 percentage points, respectively. The performance of the model including auditors' reports is significantly better than that of the model including managements' statements, demonstrating that the auditors' reports are more informative. Including both text segments yields the same results as including only auditors' reports, illustrating that, in our sample, managements' statements do not contain information useful for distress predictions beyond what is already contained in the auditors' reports. Finally, we run the same analysis on a subsample of large firms which comprise 95% of the debt in the economy, and find even stronger model enhancements when including auditors' reports. Given that the test is done on Danish data, and that Denmark is a relatively small economy, we believe that the gain from our textual analysis should be viewed as a lower bound relative to other larger economies, where greater amounts of data allow for improved model training, especially for data hungry models such as neural networks.

In the following section we review related works. The data and methods are described in Sections 3 and 4, respectively, and in Section 5 we demonstrate the applicability of our method in predicting corporate distresses. In Section 6 we illustrate heat maps of selected word blocks, and we draw conclusions and outline future work in Section 7.

2 Literature Review

Traditionally, textual analysis in financial research has consisted of simple semantic analysis based on word counts (see Loughran and McDonald (2011) and references herein). A recent example of this is Buehlmaier and Whited (2018) who use a naïve Bayes algorithm to model the probability of firms being financially constrained by using the word count in each management's statement as input.

A small string of literature most related to our work is dedicated to textual analysis in corporate distress prediction. Hájek and Olej (2013) categorize annual reports into six different semantic categories based on specific words found in the text. They then show, using a variety of models, that sentiment indicators improve the models' ability to predict corporate distress. Rönnqvist and Sarlin (2017) develop a deep learning model to analyze financial news with the aim of identifying financial institutions in distress, and Cerchiello et al. (2017) generalize the model to include numerical financial variables as well.

We add to the work of Hájek and Olej (2013) by applying a highly data-driven methodology for text processing based on deep learning, thereby allowing us to learn a deeper representation of the text and extract a stronger signal. Furthermore, we provide insight to which specific text segments of the annual reports contain information most relevant for distress prediction by examining auditors' reports and managements' statements separately. This data-driven methodology for textual analysis is close to that of Rönnqvist and

Sarlin (2017) and Cerchiello et al. (2017). However, we learn the textual representation end-to-end, compared to Cerchiello et al. (2017) who first learn a representation of the text, unrelated to the specific task, and then use it together with numerical financial variables. Our approach enables the textual representation to look for signals in the reports which are important for the task of distress prediction. Furthermore, we base our analysis on annual reports which are homogeneous across firms, whereas news articles tend to focus on specific stories which the public finds interesting.

More thoroughly studied is the concept of distress modelling using neural networks and other machine learning techniques based solely on numerical financial variables (see e.g. Jones et al. (2017), Zięba et al. (2016), Sun et al. (2014, 2017)). The existing literature tends to find that tree-based algorithms, i.e. random forest and gradient boosted trees, outperform neural networks when only numerical financial variables are included in the models. Hence, we benchmark our model against not only a neural network, but also state-of-the-art gradient boosted trees in addition to a more traditional logistic regression model.

3 Data

Our data set is based on the data used in Christoffersen et al. (2018). It consists of non-consolidated annual reports filed by all Danish non-financial and non-holding private limited and stock-based firms. This data is augmented with firm characteristics such as age, sector, and legal status from the Danish Central Business Register. In total the data set consists of the 50 numerical financial variables (44 continuous and 6 categorical) listed in Table 1. The list follows from application of a thresholded Lasso and the numerical variables are winsorized at 5% and 95% quantiles for enhanced performance (Christoffersen et al. (2018)).

We further include the auditors' reports and managements' statements found within the very same annual reports. The management's statement describes the management's opinion on the given fiscal year and its outlook on the firm's future. The auditor's report consists of several paragraphs, where the (presumably) most important for distress prediction contains the auditor's opinion of the annual report and summarizes the financial health of the firm. In this section the auditor will explicitly state any concerns regarding the continued operation of the firm or any disagreements with the management's statement. We include all available paragraphs of these two text segments in our model.

We formally seek to model the probability of a given firm entering into distress, where "distress" refers to "in bankruptcy", "bankrupt", "in compulsory dissolution", or "ceased to exist following compulsory dissolution". Firms that cease to exist due to other reasons and firms that enter into distress more than two years after the last annual report is made public are excluded from our sample.

Our sample period starts in 2013, which marks the point in time where statements are available in the

XBRL-format[1]. The sample ends in 2016, and marks the last year where we can observe realized distresses. As of 2006 small and newly established firms in Denmark are not required by law to include an auditor's report in their annual report.[2] As we want to directly compare the model when employing either or both of the two text segments, we therefore limit our data set to statements that contain both a management's statement and an auditor's report. This constraint removes 88 343 firm years from the data set (corresponding to 24.1%) and our final data set contains 278 047 firm years across 112 974 unique firms and 8 033 distresses.[3]

The 25%, 50% and 75% percentiles of the managements' statements and the auditors' reports are 37, 54, and 83 words and 187, 205, and 219 words, respectively. The greater length of the auditors' reports may not necessarily imply more relevant information, as auditors' reports typically contain standardized paragraphs that describe the responsibilities of the auditor and summarize the accounting practices.

3.1 Text Preprocessing

To preprocess the unstructured data we apply the following five steps to the auditor's report and management's statement of each annual report:

1. Remove punctuation marks, newlines and tabs and convert to lowercase.
2. Apply the Porter stemming algorithm (Porter 2001) with the NLTK library (Bird et al. 2009) to obtain the word stems and enable words to be evaluated in their canonical forms.
3. Remove stop words including numbers (i.e. dates and amounts of money) in order to avoid overfitting the network to a particular format. Numbers are replaced by a generic number token.
4. Perform named-entity recognition using spaCy (Honnibal and Montani 2017) in order to strip the text of any names and entities that may lead to overfitting in the training process and reduce generalizability.
5. During construction of the vocabulary we ignore words that have less than 25 occurrences across the entire data set.

Steps 1, 2, and 3 are considered standard procedures with the purpose of reducing unique tokens in the text in order to reduce the variability across the reports. The purpose of steps 4 and 5 is to create a model that generalizes well by removing all names and entities from the texts. The aim is to prevent the model from overfitting to certain characteristics such as firm names, auditor names, and locations. The pruning of low frequency words (step 5) is done explicitly as Danish word embeddings are trained on only a dump of the Danish Wikipedia, and rare words are therefore not represented well.

[1] Extracted data is delivered to us by Bisnode.
[2] We refer to the Danish Commerce and Companies Agency for details.
[3] Our distress definition implies that firms can enter into distress multiple times. In our sample, 47 of the distresses are such recurrent events.

Type	Input variable
Continuous	Accounts payable*
	Accounts receivable*
	Change in log size
	Corporation tax*
	Current assets*
	Deferred tax*
	Depreciation*
	EBIT*
	Equity/invested capital
	Equity*
	Expected dividends*
	Financial assets*
	Financial income*
	Financing costs*
	Fixed costs*
	Ind. EW avg. net profit*
	Interest coverage ratio
	Inventory*
	Invested capital*
	Land and buildings*
	Liquid assets*
	log(age)
	log(size)
	Long-term bank debt*
	Long-term debt*
	Long-term mortgage debt*
	Net profit*
	Other operating expenses*
	Other receivables*
	Other short debts*
	Personnel costs*
	Prepayments*
	Provisions*
	Quick ratio
	Receivables from related parties*
	Relative debt change
	Retained earnings*
	Return on equity (%)
	Short-term bank debt*
	Short-term mortgage debt*
	Tangible fixed assets*
	Tax expenses*
	Total receivables*
Categorical	Has prior distress
	Is private limited (Danish "Anpartsselskab")
	Large debt change
	Negative equity
	Region
	Sector

Table 1: **Numerical financial variables and their type (continuous or categorical)**. The table lists the 50 numerical financial variables included in the models. An asterisk denotes scaling by the firm size, which is defined as the total debt of the firm when equity is negative and otherwise total assets. We refer to Christoffersen et al. (2018) for the definition of each variable and details regarding the variable selection procedure.

4 Models for Corporate Distress Prediction

In this section we first describe our network architecture for predicting corporate distresses, which incorporates either or both of the two text segments in addition to numerical financial variables, followed by an overview of the competitive baseline models used for comparison in the experimental evaluation.

4.1 Main Model

We first provide an overview of our model in order to improve the understanding of its individual parts:

Word Representation: We use word embeddings to map each word of a text segment into a dense vector in a feature space, where semantically similar words are close to each other. Using this we split the given text segment into half-overlapping blocks of words.

Pattern Extraction: Using the embedded word blocks we utilize a convolutional neural network (CNN) to extract patterns from each block and learn a lower dimensional representation.

Pattern Understanding: The pattern output from the CNN is fed to a recurrent neural network (RNN), and the final text representation is calculated as an attention-weighted sum of the individual RNN outputs.

Feature Extensions and Prediction: We concatenate the attention-weighted sum with the numerical variables listed in Table 1 and feed it through two fully-connected layers to arrive at the final corporate distress probability prediction.

In the following we will explain in detail the individual parts, and we refer to Figure 1 for a visual description of the network architecture.

4.1.1 Word Representation

We choose to represent the semantics of each word through state-of-the-art word embeddings, which is a mapping from a word to a dense vector representation, where semantically similar words are close to each other. We use the *word2vec* model (Mikolov et al. 2013), specifically the skip-gram model. The objective of the skip-gram model is for a word to be able to predict its surrounding words. For a sequence of words $w_1, w_2, ..., w_n$ we maximize the log probability p

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

where c denotes the number of words before and after the current word to consider, which is fixed to 5 in this paper. Negative sampling is used to compute $\log p$ and the words are sub-sampled proportional to their

inverse frequency. In *word2vec* semantically similar words have a high cosine similarity between them and allows for vector calculation of words such that e.g. **king** − **man** + **woman** is very close to **queen**. We do not learn the word embedding from scratch, but rather exploit a model[4] pre-trained on a dump from the Danish Wikipedia, and fine-tune it during network training.

To prepare a given text segment of an annual report for the CNN in the next step, we create half-overlapping blocks of words with a step size of k, such that the first block consists of word $w_1, w_2, ..., w_k$ and the next block of $w_{k/2}, w_{k/2+1}, ..., w_{k/2+k}$. If the word embedding maps to vectors in $\mathbb{R}^v$, then each of these blocks B are a matrix of size $k \times v$, where $v = 300$ in our setup. We consider these embedded word blocks the input to our model, and the CNN in the next step will extract patterns from these.

4.1.2 Pattern Extraction

For each block B we apply a single-layer CNN consisting of a convolution and a max-pooling step. The purpose of the convolution is to extract matching patterns between learned filters and the embedded word block in order to learn a representation that is able to infer which patterns are important for the distress prediction task. We learn m filters from a block B, where each filter generates a new representation $\mathbf{x}$, i.e. we end up with $\mathbf{x}^{(p)}$ representations for $p \in \{1, 2, \ldots, m\}$. The ith entry in $\mathbf{x}^{(p)}$ is given by

$$\mathbf{x}_i^{(p)} = \sum_{s=0}^{\gamma-1} \sum_{j=0}^{v-1} W_{s,j}^{(p)} B_{i+s,j} \tag{2}$$

where γ denotes the number of words considered in the filter (should be less than k), v denotes the size of the word embedding, and W is a learned parameter of size $\gamma \times v$. The filter is only applied when it is not out of bounds, resulting in $\mathbf{x}^{(p)}$ being a vector of size $k - \gamma + 1$.

Each $\mathbf{x}^{(p)}$ is then max-pooled to provide a smooth signal. We denote a max-pooled vector $\mathbf{x}^{(p)}$ as $\mathbf{z}^{(p)}$, where the ith entry is given by

$$\mathbf{z}_i^{(p)} = \max \mathbf{x}_s^{(p)}, \quad s \in \left[i, i + \tau - 1\right] \tag{3}$$

where τ denotes the pool size. The max-pooling is only applied when it is not out of bounds, resulting in $\mathbf{z}^{(p)}$ having the size $k - \gamma - \tau + 2$. Finally, the results of each filter are concatenated, yielding a final vector representation $\mathbf{z}$ for each block of size $(k - \gamma - \tau + 2) \times m$.

[4] `https://github.com/Kyubyong/wordvectors`

4.1.3 Pattern Understanding

To be able to learn the semantics and sequential nature of the text as a whole, we use an RNN on the block representations we derived in the previous section. Specifically, we employ a Long Short-Term Memory (LSTM) network (Hochreiter and Schmidhuber 1997) on the block representations $\{\mathbf{z}_1, \mathbf{z}_2, ..., \mathbf{z}_T\}$, where T refers to the number of blocks the text segment of the annual report is divided into. At a given step t, an LSTM cell takes three inputs: The tth word block representation $\mathbf{z}_t$, the previous output $\mathbf{h}_{t-1}$, and the previous cell state $\mathbf{c}_{t-1}$. The cell then computes $\mathbf{h}_t$ and $\mathbf{c}_t$ by doing the following

$$\mathbf{f}_t = \sigma\big(W_f \cdot [\mathbf{h}_{t-1}, \mathbf{z}_t] + \mathbf{b}_f\big) \tag{4}$$

$$\mathbf{i}_t = \sigma\big(W_i \cdot [\mathbf{h}_{t-1}, \mathbf{z}_t] + \mathbf{b}_i\big) \tag{5}$$

$$\mathbf{u}_t = \tanh\big(W_u \cdot [\mathbf{h}_{t-1}, \mathbf{z}_t] + \mathbf{b}_u\big) \tag{6}$$

$$\mathbf{o}_t = \sigma\big(W_o \cdot [\mathbf{h}_{t-1}, \mathbf{z}_t] + \mathbf{b}_o\big) \tag{7}$$

$$\mathbf{c}_t = \mathbf{f}_t \odot \mathbf{c}_{t-1} + \mathbf{i}_t \odot \mathbf{u}_t \tag{8}$$

$$\mathbf{h}_t = \mathbf{o}_t \odot \tanh \mathbf{c}_t \tag{9}$$

where σ and tanh are element-wise sigmoid and hyperbolic tangent functions, $\odot$ is element-wise multiplication, all W and $\mathbf{b}$ are learned parameters, and $\mathbf{f}_t$, $\mathbf{i}_t$, $\mathbf{o}_t$ are known as the forget, input and output gates of the LSTM cell.

Instead of using the output at the final step $\mathbf{h}_T$, we use an attention-weighted sum of the step-wise outputs. Specifically, for each $\mathbf{h}_t$ we learn a scalar $score(\mathbf{h}_t)$ that signifies the importance of that specific $\mathbf{h}_t$. The score is computed using a single layer of size 1 with a linear activation. We use the softmax-function to normalize each scalar to derive each attention weight α_t

$$\alpha_t = \frac{\exp(score(\mathbf{h}_t))}{\sum_{i=1}^{T} \exp(score(\mathbf{h}_i))} \tag{10}$$

We can then derive the final attention-weighted textual representation by the weighted sum

$$\mathbf{h}_{\text{final}} = \sum_{t=1}^{T} \alpha_t \mathbf{h}_t \tag{11}$$

The benefit of using attention is to enable the model to focus its attention on fewer, but more important, parts of the text to learn a better descriptive representation (Zhang et al. 2018). Additionally, it enables an improved gradient flow in longer texts, such as the ones we work with in this paper.

4.1.4 Feature Extension and Prediction

We now have a dense textual representation, $\mathbf{h}_{\text{final}}$, which we concatenate with the numerical variables $\mathbf{h}_{\text{num}}$ of Table 1, yielding a vector $\mathbf{h}_{\text{concat}}$ with a length equal to the sum of the number of handcrafted features and LSTM cell size. This concatenated representation is passed through two fully-connected layers of size 200 and 50 with a single neuron layer as the last step with a sigmoid activation. This is to allow the textual representation to interact with the numerical variables before doing the final prediction. The two layers of size 200 and 50 use the rectified linear unit (ReLU) activation function

$$\mathbf{h}_{\text{concat}} = [\mathbf{h}_{\text{final}}, \mathbf{h}_{\text{num}}] \tag{12}$$

$$\mathbf{l}_1 = ReLU\big(W_1 \cdot \mathbf{h}_{\text{final}} + \mathbf{b}_1\big) \tag{13}$$

$$\mathbf{l}_2 = ReLU\big(W_2 \cdot \mathbf{l}_1 + \mathbf{b}_2\big) \tag{14}$$

$$PD = \sigma\big(W_3 \cdot \mathbf{l}_2 + \mathbf{b}_3\big) \tag{15}$$

where PD denotes the predicted distress probability. We train the network using the Adam optimizer (Kingma and Ba 2014) and use the binary cross-entropy as the loss function. We will detail the parameters of the cross-validated network configuration in Section 4.2.

It is well known that neural networks are susceptible to overfitting (Gu et al. 2018). As a way of regularizing the training process we set aside 10% of the training set as a validation. The validation set is used for early stopping, i.e. we terminate the gradient descent when the network starts to overfit.

4.2 Parameter Tuning in the Main Model

We tune the neural network[5] using cross-validation over the hyperparameter space. For the convolutional neural network we consider block sizes in the set $\{10, 15, 20\}$, number of filters in $\{40, 60\}$, and pool sizes in $\{2, 4, 6\}$. For the recurrent neural network we consider LSTM cell sizes in $\{50, 100, 150\}$. Lastly, we consider learning rates in $\{10^{-3}, 10^{-4}\}$. We run for a maximum of 10 epochs which, however, was never reached due to early stopping and use a batch size of 64 due to memory constraints. We observe that the results[6] are robust across this set of parameters to within one standard error for both text segments. Consequently, we use typical values in our models. For the convolutional neural network this means a block size of $k = 20$ and $m = 40$ filters with a pool size of $\tau = 4$. The recurrent neural network uses an LSTM with a cell size of 100, and we employ a learning rate of 10^{-3}. We set γ (number of words to convolve over) to half the block size, i.e. $\gamma = 10$. The results of the grid search are illustrated in Figure A.1 of the Appendix for both text

[5] We implemented the neural models in TensorFlow (Abadi et al. 2016).
[6] AUC, described in Section 5.1, is used as the performance metric during parameter-tuning.

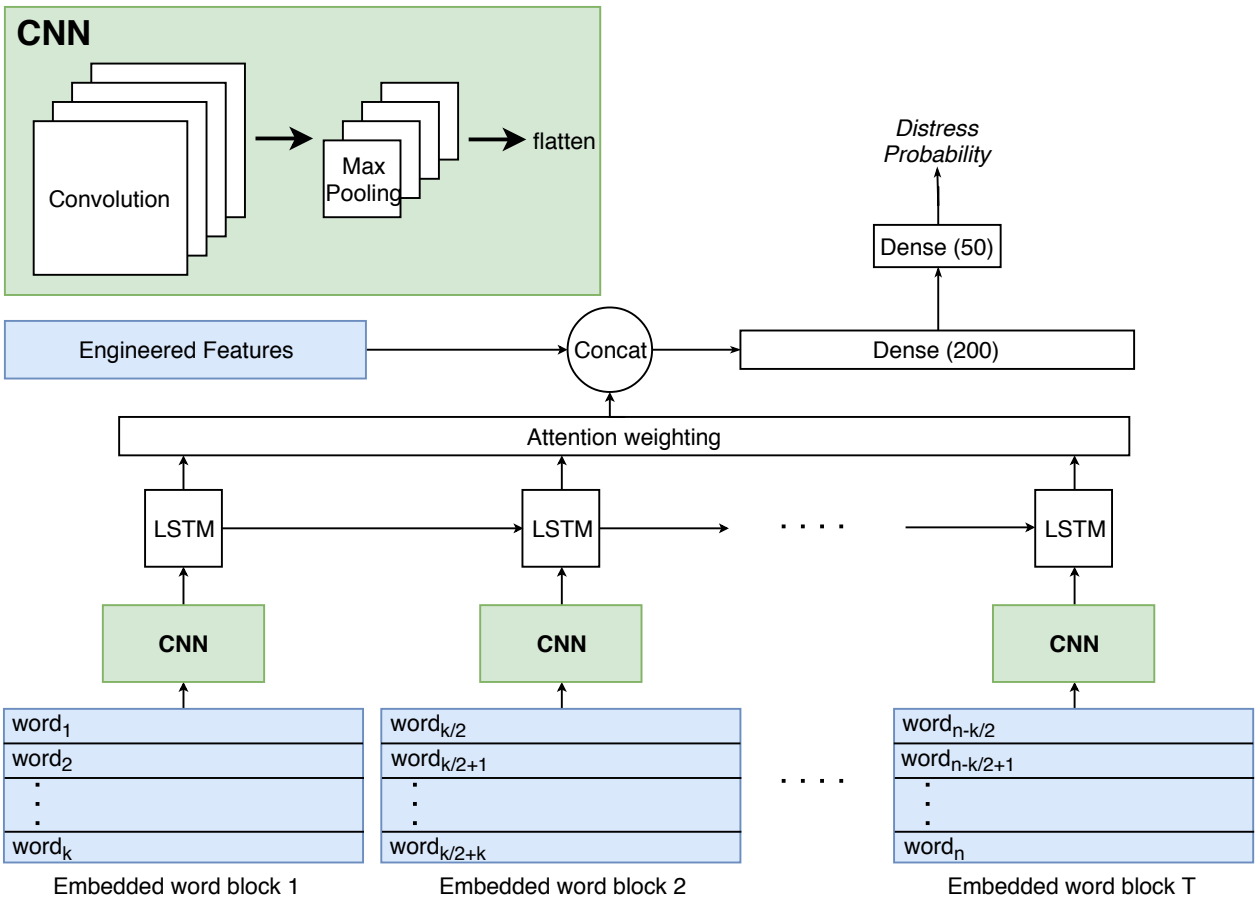


Figure 1: **Network architecture.**

segments.

4.3 Baseline Models

We implement three baseline models based solely on the numerical financial variables, against which we benchmark our main model.

First, we implement a neural network based on the same architecture as our main model, but where the textual component is not included. That is, the model consists of the two top dense layers in Figure 1. This model serves as a natural benchmark as it will reveal the impact of the text segments on the prediction accuracy.

Secondly, we implement a model based on gradient boosted trees, specifically XGBoost (XGB) (Chen and Guestrin 2016), which typically performs better than neural networks for predicting corporate bankruptcies (Zięba et al. 2016, Jones et al. 2017). It is an ensemble technique which recursively combines multiple relatively simple models, so-called (weak) base learners which consist of regression trees, to produce a highly

accurate prediction rule.

Finally, we implement a logistic regression (logit) which is a relatively simple, yet very common, choice for distress models (see e.g. Shumway (2001), Chava and Jarrow (2004), Beaver et al. (2005), Campbell et al. (2008)).

5 Experimental Evaluation

5.1 Evaluation Measures

We quantify model performance using two metrics, AUC and log score. The AUC (Area Under the receiver operating characteristics Curve) is a commonly used metric in distress prediction models. It measures the probability that a model places a higher risk on a random firm that experiences a distress event in a given year than a random firm that does not experience a distress event in a given year. Hence, 0.5 is random guessing and 1 is a perfect result.

AUC is only a ranking measure; a model may rank the firms well, but perform poorly in terms of the level of the predicted probabilities. Generally, we are interested in well-calibrated probabilities in addition to their ranking. Thus, we look at the log score as well which takes into account the individual predicted probabilities. The log score, $\mathcal{L}$, for a given model is defined as

$$\mathcal{L} = -\frac{1}{N}\sum_{i,t}\left(y_{it}\log(\hat{p}_{it}) + (1-y_{it})\log(1-\hat{p}_{it})\right) \tag{16}$$

where $\hat{p}_{it}$ is the model-predicted distress probability of firm i in year t, y_{it} is a dummy that is equal to 1 if firm i enters into distress in year t and 0 otherwise, and N is the sample size. A smaller log score implies a better model fit.

5.2 Main Results

This section presents the main results of the out-of-sample tests of our models. We use 10-fold cross-validation, where the folds are constructed by sampling firms. Ideally, we would have used an expanding window of data to estimate the models and forecast the probability of the firms entering into distress two years after the estimation window closes, thereby mimicking the true forecasting situation. However, this forecasting scheme is not viable in the current study due to the limited number of years in our data set.

The average AUC and log score across folds with one standard error bands are shown in Figure 2, where the neural network without text is denoted NN and the neural networks with text are denoted $\text{NN}_{\text{aud + man}}$, NN_{aud}, and NN_{man} depending on the text segments included in the model (*aud* refers to auditor's report and

man to the management's statement). This nomenclature will be used for the remainder of the paper. We observe that the neural networks with text have higher AUC and smaller log score than all baseline models. That is, the models with text are both better at ranking firms by their riskiness and provide better model fits in general.

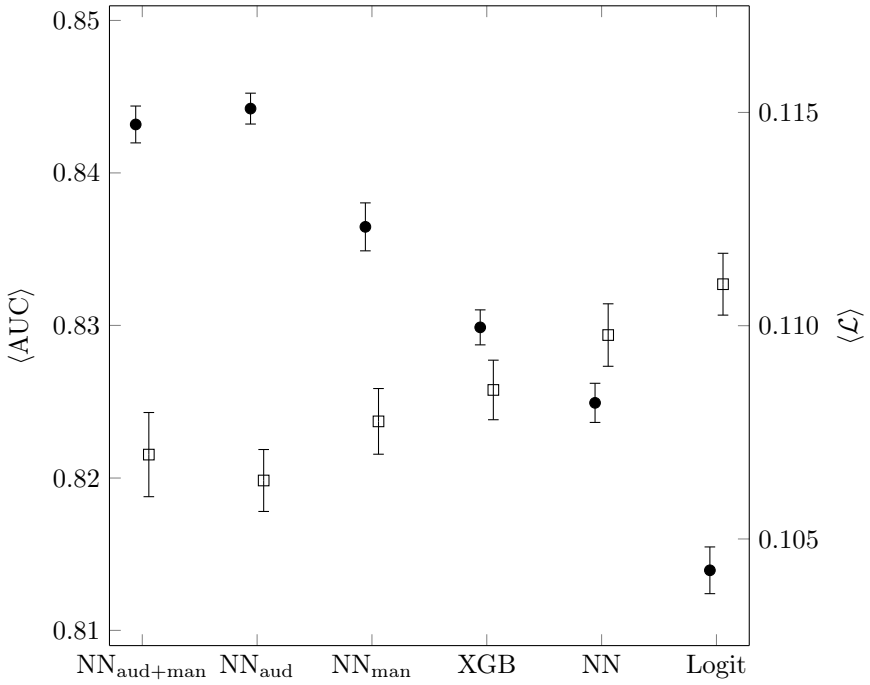


Figure 2: **Average AUC and log score.** The figure shows average AUC (● left axis) and average log score (□ right axis) with one standard error band for each of the six models. Averages and standard errors are calculated based on 10 folds, which are constructed by sampling firms.

The results of Figure 2 are furthermore summarized in Table 2 alongside p-values from a paired two-tailed t-test comparing results of each baseline model to $\text{NN}_{\text{aud + man}}$, NN_{aud}, and NN_{man}, respectively. A statistically significant improvement is observed in the models with text relative to any of the baseline models, both when it comes to AUC and log score. Specifically, we find that including auditors' reports, managements' statements, and both in the neural network increases the AUC by 1.9, 1.1, and 1.8 percentage points, respectively. That is, both the auditors' reports and the managements' statements have significant predictive power beyond what is captured by the numerical financial variables themselves.

The AUC and log score of NN_{aud} is significantly better than that of NN_{man}, i.e. the auditors' reports contain more valuable information than the managements' statements. There can be several explanations for that. First, the auditors' reports are longer, enabling the neural network to learn a better representation

Model	$\langle \text{AUC} \rangle$	$p_{\text{aud+man}}$	p_{aud}	p_{man}
$\text{NN}_{\text{aud+man}}$	0.843	–	–	–
NN_{aud}	0.844	0.233	–	–
NN_{man}	0.836	0.000	0.000	–
XGB	0.830	0.000	0.000	0.003
NN	0.825	0.000	0.000	0.000
Logit	0.814	0.000	0.000	0.000

(a) AUC

Model	$\langle \mathcal{L} \rangle$	$p_{\text{aud+man}}$	p_{aud}	p_{man}
$\text{NN}_{\text{aud+man}}$	0.1070	–	–	–
NN_{aud}	0.1064	0.4263	–	–
NN_{man}	0.1078	0.1471	0.0032	–
XGB	0.1085	0.0643	0.0001	0.0372
NN	0.1098	0.0005	0.0000	0.0001
Logit	0.1110	0.0001	0.0000	0.0000

(b) Log score $\mathcal{L}$

Table 2: **Average AUC and log score.** The table shows (a) average AUC and (b) average log score, where $p_{\text{aud+man}}$, p_{aud}, and p_{man} denote p-values from a paired two-tailed t-test between the scores of the current model and the three models including text. Averages and standard errors are calculated based on 10 folds, which are constructed by sampling firms.

of the text. Secondly, and more importantly, the management's statement about its own business is likely to be less objective and biased towards a brighter outlook on the future, whereas the independent auditor's report contains the auditor's unbiased professional opinion. Interestingly, there is no significant difference between NN_{aud} and $\text{NN}_{\text{aud+man}}$. If anything, there is a small tendency for NN_{aud} to perform better than $\text{NN}_{\text{aud+man}}$. This finding implies that, though there is information in the managements' statements which is not captured by the financial variables, all information in the managements' statements is captured by the auditors' reports. Hence, it might be preferable to focus only on the auditors' reports and leave out the managements' statements in future work.

5.3 Results for Large Firms

We repeat the above test, but only include firms of a size[7] greater than 5 million DKK. These firms correspond to only 35.4% of the sample size, but 95.4% of the total debt. It is of greater interest to quantify the performance among these dominating firms as they hold the majority of the total assets and debt in the economy. Model estimation is still done on the full sample.

The results are summarized in Table 3, and we observe that all models yield better AUC and log score compared to the previous experiment. This is not surprising as large firms likely provide more accurate

[7]Cf. Table 1 we define firm size as the total debt of the firm when equity is negative and otherwise total assets.

annual reports which lead to more accurate model predictions.[8] Interestingly, the AUC now increases by 2.6 percentage points when adding auditors' reports to the neural network, where the increase was 1.9 percentage points in the previous experiment. We speculate that this is due to the auditors' reports of the larger firms being of a higher quality and more informative, implying that the neural network can extract more information from them. On the contrary, we do not see an increased enhancement in AUC when it comes to the managements' statements, and the difference in AUC between XGB and NN_{man} is now insignificant. This highlights that there is information to be extracted from the auditors' reports, in particular when it comes to large firms, whereas the managements' statements are less informative. The loss of significance is possibly caused by the smaller sample size, resulting in more extreme values of the individual folds.

Model	$\langle \text{AUC} \rangle$	$p_{\text{aud+man}}$	p_{aud}	p_{man}
$\text{NN}_{\text{aud+man}}$	0.877	–	–	–
NN_{aud}	0.879	0.562	–	–
NN_{man}	0.864	0.013	0.004	–
XGB	0.860	0.000	0.000	0.290
NN	0.853	0.000	0.000	0.002
Logit	0.834	0.000	0.000	0.000

(a) AUC

Model	$\langle \mathcal{L} \rangle$	$p_{\text{aud+man}}$	p_{aud}	p_{man}
$\text{NN}_{\text{aud+man}}$	0.0611	–	–	–
NN_{aud}	0.0611	0.9815	–	–
NN_{man}	0.0627	0.0551	0.0095	–
XGB	0.0629	0.0085	0.0127	0.6588
NN	0.0640	0.0036	0.0001	0.0046
Logit	0.0657	0.0000	0.0001	0.0002

(b) Log score $\mathcal{L}$

Table 3: **Average AUC and log score of large firms.** The table shows (a) average AUC and (b) average log score, where $p_{\text{aud+man}}$, p_{aud}, and p_{man} denote p-values from a paired two-tailed t-test between the scores of the current model and the three models including text. Averages and standard errors are calculated based on 10 folds, which are constructed by sampling firms larger than 5 million DKK.

5.4 Robustness: Sampling Across Time

In order to ensure that the observed signal in the text is not merely a result of a particular fold composition where we accidentally gauge a proxy for a temporal effect, we also perform a robustness test where we explicitly construct folds based on the publication year of the annual reports. This gives four folds in total. The results of this experiment are summarized in Table 4, and the scores display the same tendency as in

[8] The large drop in log score can also in part be due to a smaller distress rate among large firms. The change in the composition of the outcome variable will by construction reduce the log score.

Model	$\langle \text{AUC} \rangle$	$p_{\text{aud+man}}$	p_{aud}	p_{man}
$\text{NN}_{\text{aud+man}}$	0.843	–	–	–
NN_{aud}	0.842	0.299	–	–
NN_{man}	0.830	0.003	0.014	–
XGB	0.826	0.001	0.006	0.175
NN	0.822	0.001	0.004	0.054
Logit	0.814	0.000	0.001	0.009

(a) AUC

Model	$\langle \mathcal{L} \rangle$	$p_{\text{aud+man}}$	p_{aud}	p_{man}
$\text{NN}_{\text{aud+man}}$	0.1090	–	–	–
NN_{aud}	0.1095	0.4130	–	–
NN_{man}	0.1114	0.0312	0.1289	–
XGB	0.1109	0.0128	0.0627	0.3484
NN	0.1122	0.0112	0.0166	0.2098
Logit	0.1127	0.0081	0.0005	0.1892

(b) Log score $\mathcal{L}$

Table 4: **Average AUC and log score obtained from sampling years.** The table shows (a) average AUC and (b) average log score, where $p_{\text{aud+man}}$, p_{aud}, and p_{man} denote p-values from a paired two-tailed t-test between the scores of the current model and the three models including text. Averages and standard errors are calculated based on 4 folds, which are constructed by sampling publication years.

Table 2, further validating the results. The slightly larger p-values can be attributed the smaller number of folds in this experiment, resulting in a weaker statistical test.

6 Cases of Blocks with High Attention Weights

The attention weights can be extracted from the individual word blocks to highlight words and phrases, which are important for the prediction. In this section we present examples of individual blocks from five auditors' reports. The attention-color-strength of individual words is relative to the largest weight in that particular block, and stop words are inserted for completeness to make the text more readable. In the five cases below the attention mechanism successfully highlights sections that intuitively should affect the distress prediction, e.g. "*no realistic options for obtaining funding*" and "*significant uncertainty about the firm's ability to continue operations*". Generally, these examples show cases of information that would be very difficult to represent in traditional features, such as those described in Table 1. The texts are originally in Danish, and we note that the translation has required shifting some words, but to the best of our ability we have aimed at a 1:1 comparison. The original texts are in Appendix B for reference.

Example 1

henceforth. It is our assessment that there are no realistic options for obtaining funding and we therefore make the caveat that the statement has been submitted on the basis of continued operations. It is our opinion that the statement as a consequence of the significance

Example 2

the mention in the statement's notes and the management's report where the management explains the significant uncertainty about the firm's ability to continue operations as it is still uncertain if the necessary liquidity can be generated for financing

Example 3

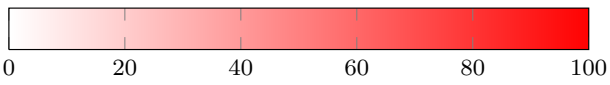
obtaining the liquidity for payment of a significant tax liability. It is uncertain whether the firm will be able to obtain this additional liquidity. We are thus not able to comment on the company's ability to continue operations the coming year, why we have reservations. It should also be noted that there is not

Example 4

has not yet received acknowledgment from the involved bank, and on that basis we can not reach a conclusion regarding the firm's ability to continue operation. Non-Conclusion Due to

Example 5

on note xxnumberxx external accounts, which shows that the company's equity is exhausted. The company's continued operation therefore depends on that the necessary liquidity continues to be provided. The firm's



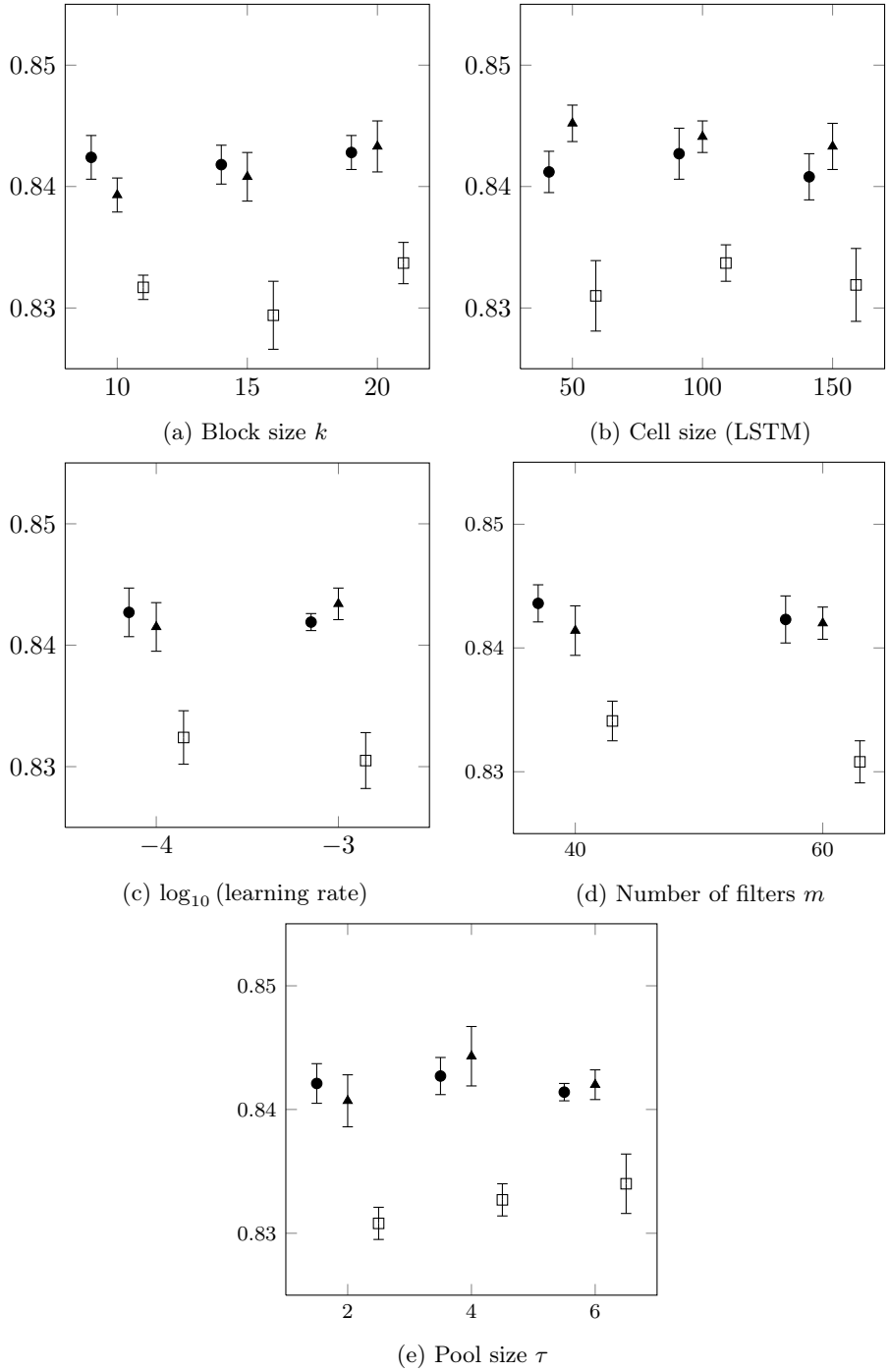
7 Outlook and Conclusion

We have introduced a network architecture consisting of both convolutional and recurrent neural networks for predicting corporate distresses using auditors' reports and managements' statements of annual reports. By concatenating the neural network model with numerical financial variables, we found that the model with

auditors' reports increased the AUC by almost 2 percentage points compared to a neural network without text while the managements' statements only gave an enhancement of roughly 1 percentage point. The enhancement in model performance is statistically significant at the 1% level in both cases, demonstrating that there is useful information to be extracted from text segments besides what is already contained in the numerical financial variables. Statistical tests also revealed that auditors' reports provide significantly more information than managements' statements and that all useful information contained in the managements' statements is contained in the auditors' reports as well. These findings suggest that further analyses should focus merely on auditors' reports. For firms with a size greater than 5 million DKK the auditors' reports enhanced the AUC by more than 2.5 percentage points, while it was still roughly 1 percentage point for managements' statements, showing that textual analysis of auditors' reports is especially useful for large firms from which we benefit the most from accurate distress predictions.

In future work it would be interesting to investigate individual paragraphs within the auditors' reports and managements' statements to see if certain paragraphs in combination are more suited for the distress prediction.

A Results of Parameter Tuning in the Main Model



(a) Block size k

(b) Cell size (LSTM)

(c) $\log_{10}$ (learning rate)

(d) Number of filters m

(e) Pool size τ

Figure A.1: **Parameter-tuning of the neural network.** The figure illustrates AUC (▲ $\text{NN}_{\text{aud+man}}$; ● NN_{aud}; □ NN_{man}) for different parameter choices. Error bars denote one standard error.

B Cases of Blocks with High Attention Weights (Original)

Example 1

fremover. Det er vores vurdering, at der ikke er realistiske muligheder for at fremskaffe finansiering og vi tager derfor forbehold for, at årsregnskabet er aflagt under forudsætning af fortsat drift. Det er vores opfattelse, at årsregnskabet, som følge af betydeligheden

Example 2

til omtale i årsregnskabets noter og ledelsesberetningen, hvori ledelsen redegør for væsentlig usikkerhed om selskabets evne til at fortsætte driften, da det endnu er usikkert, om den nødvendige likviditet kan frembringes til finansiering

Example 3

fremskaffelse af likviditeten til betaling af en væsentlig momsgæld. Det er usikkert hvorvidt selskabet vil være i stand til at fremskaffe denne yderligere likviditet. Vi er således ikke i stand til at udtale os om selskabets evne til at fortsætte driften det kommende år, hvorfor vi tager forbehold herfor. Det skal endvidere bemærkes, at der ikke

Example 4

har endnu ikke modtaget tilkendegivelse fra det involverede pengeinstitut, og på den baggrund kan vi ikke nå frem til en konklusion vedrørende selskabets evne til at fortsætte driften. Manglende konklusion På grund

Example 5

på note xxnumberxx eksternt regnskab, hvoraf det fremgår, at selskabet egenkapital er tabt. Selskabets fortsatte drift er derfor afhængig af at der fortsat stilles den nødvendige likviditet til rådighed. Selskabets

References

Abadi, M., P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, et al. (2016). Tensorflow: A system for large-scale machine learning. *OSDI 16*, 265–283.

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

Atiya, A. F. (2001). Bankruptcy prediction for credit risk using neural networks: A survey and new results. *IEEE Transactions on Neural Networks 12*(4), 929–935.

Beaver, W., M. McNichols, and J. Rhie (2005). Have financial statements become less informative? Evidence from the ability of financial ratios to predict bankruptcy. *Review of Accounting Studies 10*(1), 93–122.

Bird, S., E. Klein, and E. Loper (2009). *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O'Reilly.

Buehlmaier, M. M. M. and T. M. Whited (2018). Are financial constraints priced? Evidence from textual analysis. *The Review of Financial Studies 31*(7), 2693–2728.

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

Caruana, R. and A. Niculescu-Mizil (2006). An empirical comparison of supervised learning algorithms. *Proceedings of the 23rd International Conference on Machine Learning, ICML '06*, 161–168.

Cerchiello, P., G. Nicola, S. Ronnqvist, and P. Sarlin (2017). Deep learning bank distress from news and numerical financial data. *Working paper*.

Chava, S. and R. A. Jarrow (2004). Bankruptcy prediction with industry effects. *Review of Finance 8*(4), 537–569.

Chen, T. and C. Guestrin (2016). Xgboost: A scalable tree boosting system. *Proceedings of the 22nd ACM sigkdd international conference on knowledge discovery and data mining*, 785–794.

Christoffersen, B., R. Matin, and P. Mølgaard (2018). Can machine learning models capture correlations in corporate distresses? *Working paper*, Danmarks Nationalbank.

Gu, J., Z. Wang, J. Kuen, L. Ma, A. Shahroudy, B. Shuai, T. Liu, X. Wang, G. Wang, J. Cai, and T. Chen (2018). Recent advances in convolutional neural networks. *Pattern Recognition 77*, 354–377.

Hájek, P. and V. Olej (2013). Evaluating sentiment in annual reports for financial distress prediction using neural networks and support vector machines. *International Conference on Engineering Applications of Neural Networks*, 1–10.

Hochreiter, S. and J. Schmidhuber (1997). Long short-term memory. *Neural computation 9*(8), 1735–1780.

Honnibal, M. and I. Montani (2017). spaCy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. *To appear*.

Hosaka, T. (2019). Bankruptcy prediction using imaged financial ratios and convolutional neural networks. *Expert Systems with Applications 117*, 287–299.

Jones, S., D. Johnstone, and R. Wilson (2017). Predicting corporate bankruptcy: An evaluation of alternative statistical frameworks. *Journal of Business Finance & Accounting 44*(1-2), 3–34.

Kingma, D. P. and J. Ba (2014). Adam: A method for stochastic optimization. *Working paper*.

Loughran, T. and B. McDonald (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-ks. *The Journal of Finance 66*(1), 35–65.

Mikolov, T., I. Sutskever, K. Chen, G. S. Corrado, and J. Dean (2013). Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 3111–3119.

Ohlson, J. (1980). Financial ratios and the probabilistic prediction of bankruptcy. *Journal of Accounting Research 18*(1), 109–131.

Porter, M. F. (2001). Snowball: A language for stemming algorithms.

Rönnqvist, S. and P. Sarlin (2017). Bank distress in the news: Describing events through deep learning. *Neurocomputing 264*, 57–70.

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

Sun, J., H. Fujita, P. Chen, and H. Li (2017). Dynamic financial distress prediction with concept drift based on time weighting combined with adaboost support vector machine ensemble. *Knowledge-Based Systems 120*, 4–14.

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

Tsai, C.-F. and J.-W. Wu (2008). Using neural network ensembles for bankruptcy prediction and credit scoring. *Expert Systems with Applications 34*, 2639–2649.

Zhang, L., S. Wang, and B. Liu (2018). Deep learning for sentiment analysis: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*.

Zięba, M., S. K. Tomczak, and J. M. Tomczak (2016). Ensemble boosted trees with synthetic features generation in application to bankruptcy prediction. *Expert Systems with Applications 58*, 93–101.

DANMARKS NATIONALBANK
HAVNEGADE 5
DK-1093 COPENHAGEN K
WWW.NATIONALBANKEN.DK

DANMARKS NATIONALBANK