European Journal of Operational Research 302 (2022) 309–323

Contents lists available at ScienceDirect

European Journal of Operational Research

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

Innovative Applications of O.R.

Credit default prediction from user-generated text in peer-to-peer lending using deep learning☆,☆☆

Johannes Kriebel*, Lennart Stitz

University of Münster, Universitätsstraße 14-16, 48143 Münster, Germany

ARTICLE INFO

Article history:
Received 15 October 2020
Accepted 16 December 2021
Available online 23 December 2021

Keywords:
OR in banking
Peer-to-peer lending
Deep learning
Textual data
Credit risk

ABSTRACT

Digital technologies produce vast amounts of unstructured data that can be stored and accessed by traditional banks and fintech companies. We employ deep learning and several other techniques to extract credit-relevant information from user-generated text on Lending Club. Our results show that even short pieces of user-generated text can improve credit default predictions significantly. The importance of text is further supported by an information fusion analysis. Compared with other approaches that use text, deep learning outperforms them in almost all cases. However, machine learning models combined with word frequencies or topic models also extract substantial credit-relevant information. A comparison of six deep neural network architectures, including state-of-the-art transformer models, finds that the architectures mostly provide similar performance. This means that simpler methods (such as average embedding neural networks) offer performance comparable to more complex methods (such as the transformer networks BERT and RoBERTa) in this credit scoring setting.

© 2021 Elsevier B.V. All rights reserved.

1. Introduction

Digital technologies produce vast amounts of unstructured data that can be stored and accessed by banks and fintech companies. This includes communication, such as emails or telephone calls, text and image data posted on social media, and records of bank transactions. Further, even before the large-scale dissemination of modern digital technologies, Merrill Lynch is famously quoted as stating that up to 80% of the data in companies are unstructured (Shilakes & Tylman, 1998).

From the perspective of banking operations, a crucial step in accepting and pricing decisions related to credit contracts is default prediction. Today, there is a thorough understanding of which structured data to use for predicting credit risk and which methods best exploit this data (e.g., Baesens, Setiono, Mues, & Vanthienen, 2003, Crook, Edelman, & Thomas, 2007, Kumar & Ravi, 2007, Lessmann, Baesens, Seow, & Thomas, 2015, Gunnarsson, van den Broucke, Baesens, Óskarsdóttir, & Lemahieu, 2021). In contrast, unstructured data currently remains largely unused for this purpose. However, recent literature—such as Iyer, Khwaja, Luttmer, & Shue (2016) (images and text in peer-to-peer lending), Berg, Burg, Gombović, & Puri (2020) (digital footprints), and Liu, Shang, Wu, & Chen (2020) (soft information and social collateral)—has shown that this unstructured data could contain important credit-relevant information. Unfortunately, unstructured data, such as text, cannot be used directly in the standard techniques for credit default prediction without prior processing.

To overcome the difficulties of extracting information from text, prior research has relied primarily on two approaches. The first and more traditional one directly identifies text characteristics that are linked to borrower quality. These include the frequency of identity claims (as in Herzenstein, Sonenshein, & Dholakia, 2011), spelling mistakes, the length of the text, the use of words with social and emotional connotations (as in Dorfleitner, Priberny, Schuster, Stoiber, Weber, de Castro, & Kammler, 2016), the use of punctuation (as in Chen, Huang, & Ye, 2018), and characteristics such as readability, tone, and deception cues (as in Gao, Lin, & Sias, 2021). The advantage of this approach is that it exploits well-established links between textual characteristics and cognitive processes. More recent studies use machine learning combined with for example word frequencies to exploit this information for credit risk predictions. Netzer, Lemaire, & Herzenstein (2019) use ensembles of several classifiers built on the most frequent bigrams in text on Pros-

☆ This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

☆☆ The authors would like to thank Emanuele Borgonovo (the editor), three anonymous referees, Destan Kirimhan, Thomas Langer, Andreas Pfingsten, Doron Reichmann, Judith Schneider, Kevin Wiegratz, the participants of the Financial Management Association Annual Meeting 2020, and the participants of the Banking Research Workshop 2020 for their valuable input and suggestions.

* Corresponding author.
E-mail addresses: johannes.kriebel@wiwi.uni-muenster.de (J. Kriebel), lennart.stitz@wiwi.uni-muenster.de (L. Stitz).

https://doi.org/10.1016/j.ejor.2021.12.024
0377-2217/© 2021 Elsevier B.V. All rights reserved.

per. Xia, He, Li, Liu, & Ding (2020) use cluster methods based on term frequencies.

While these studies draw informative conclusions and their approaches can extract credit-relevant information in several cases, the standard for classifying pieces of text in computer science (e.g., for categorizing customer communication) is a type of machine-learning methods referred to as deep learning. Generally, deep learning refers to fitting complex artificial neural networks to large numbers of cases to extract crucial characteristics without the intervention of a researcher.

In fact, Mai, Tian, Lee, & Ma (2019), Ahmadi, Martens, Koch, Gottron, & Kramer (2018), Matin, Hansen, Hansen, & Mølgaard (2019), and Stevenson, Mues, & Bravo (2021) use deep learning methods in corporate lending to identify good borrowers based on reporting information. Mai et al. (2019) use convolutional neural networks and so-called average embedding networks based on the Management Discussion and Analysis (MD&A) sections of 10-K filings. Matin et al. (2019) use recurrent convolutional neural networks based on management discussions and auditor reports. Ahmadi et al. (2018) use dependency-sensitive convolutional neural networks to filter relevant text passages. Stevenson et al. (2021) use recent state-of-the-art transformer networks to predict the probability of default based on textual assessments from credit experts.

While some studies explore the use of textual information to predict credit defaults, some facets of the field are still not clearly understood. An important question that remains unanswered is how valuable the textual information actually is in improving credit default prediction. For example, Dorfleitner et al. (2016) and Chen et al. (2018) do not find any clear evidence of text characteristics being predictors of credit default. Netzer et al. (2019) find promising results combining machine learning with word frequencies, but the effect of textual information on the quality of predictions in Xia et al. (2020) is rather small. More recently, deep learning approaches such as Mai et al. (2019) and Matin et al. (2019) obtain promising results, while Ahmadi et al. (2018) do not find clear evidence of text being predictive of corporate distress. Stevenson et al. (2021) find text to be a relevant characteristic when assessing text alone, but they find the benefit of text in a situation with other structured information to be unclear. Fitzpatrick & Mues (2021) recently evaluated different prediction methods for profit scoring in peer-to-peer lending. They also include analyses that process text.

Another important aspect to be considered, if text is informative, is which method is best for extracting credit-relevant information from text. This is particularly the case, since many datasets used in the credit scoring literature do not necessarily contain multiple million datapoints as big data is often described (Kraus, Feuerriegel, & Oztekin, 2020). Instead, they might require big data methods because of the complexity of unstructured data. This makes it particularly interesting to study the choice of methods for these settings. In fact, Gunnarsson et al. (2021) recently pointed out the importance of studying deep learning algorithms to make use of unstructured data in credit scoring. Interestingly, there is currently no benchmark study to assess the choice of models for processing text in credit scoring, which is essential for both practitioners and academics.

This study aims to address these issues. We use six deep learning architectures and several other methods to extract credit-relevant information from user-generated text on Lending Club, a major platform for peer-to-peer lending in the United States. The deep learning methods are convolutional neural network, recurrent neural network, convolutional recurrent neural network, average embedding neural network, bidirectional encoder representations from transformers (BERT), and a robustly optimized BERT pretraining approach (RoBERTa). Transformer models are currently considered the state of the art for almost all natural language processing tasks. Our findings show that the textual information extracted by deep learning is a significant predictor of out-of-sample credit defaults, and it can generate substantial additional profit for lenders. The importance is further supported by an information fusion analysis. Average embedding neural networks, convolutional neural networks, convolutional recurrent neural networks, BERT, and RoBERTa achieve similar prediction quality. Simpler deep learning methods such as the average embedding neural network have similar performance when compared to more complex methods, such as BERT or RoBERTa. The deep learning approaches tend to outperform alternative machine learning approaches that use word frequencies, pre-trained word embeddings, or topic models in almost all cases. This also applies to rule-based text characteristics. Among the alternative machine learning methods, those based on word frequencies and a topic model provide considerable prediction performance. Among the text characteristics, spelling mistakes are found to be of potential importance.

We make three important contributions to the literature: Regarding the question whether text is relevant in credit scoring—especially for peer-to-peer lending—we confirm that those texts are important for a wide range of approaches. In a comparison of deep learning techniques, alternative machine learning methods, and rule-based approaches, we find that deep learning tends to outperform the other approaches in almost all cases. Thus, it is preferable to practitioners when performance is the primary issue. This is supported by a comparison using information fusion to fuse deep learning models and alternative models, respectively. Researchers and practitioners might further be interested in choosing from various deep learning architectures. We find that simpler and more complex methods tend to offer similar performance. This is particularly interesting because these methods differ in the effort required by practitioners to build prediction models. In addition, researchers and practitioners find various interesting results on characteristics of interest, such as spelling mistakes.

The rest of this paper is structured as follows. Section 2 outlines important literature relevant to our work. Section 3 describes our data. Section 4 presents the methods we use. Section 5 describes the results regarding the quality of predictions of the deep learning approaches, other machine learning approaches, rule-based text characteristics, information fusion variable importance, and robustness checks. Section 6 evaluates the economic impact of the improvement in prediction quality, Section 7 considers managerial implications, and Section 8 presents conclusions of the study.

2. Literature review

Today, a comprehensive body of literature exists on the prediction of credit defaults from structured data, covering both standard statistical methods and machine learning. The choice of method is reviewed and benchmarked in several studies that analyze default prediction methods (e.g., Baesens et al., 2003, Kumar & Ravi, 2007, Crook et al., 2007, Lessmann et al., 2015, Gunnarsson et al., 2021). Although results differ in their specifics, certain findings are common across studies. First, more complex machine learning methods—such as artificial neural networks and XGBoosting—tend to outperform simpler regression approaches. Second, ensemble methods, which combine the predictions of several models, tend to outperform individual models. This is in line with recent findings of Fitzpatrick & Mues (2016), who compare boosted regression trees, random forests, and various regression approaches in terms of their ability to predict mortgage-loan defaults. They conclude that boosted regression trees lead to the most promising results. Finlay (2010) focus on ensemble methods and compare multiple-classifier systems based on a diverse set of models (i.e., logistic regression, linear discriminant analysis, classification trees, artificial

neural networks, and clustering). They find that bagging and boosting outperform other ensemble methods. Dumitrescu, Hu, Hurlin, & Tokpavi (2022) show that a penalized logistic regression based on rules from short-depth decision trees could outperform various alternatives while providing favorable characteristics in interpretability. The superior performance of ensembles is also supported by Gunnarsson et al. (2021), who compare two deep learning models, logistic regression, a decision tree, and two ensemble methods for credit scoring on structured data. They identify XGBoosting as the best performing method.

While earlier approaches to credit risk prediction are based primarily on structured data, there is strong evidence that including unstructured data—such as text—could significantly increase the quality of prediction. Liu et al. (2020) provide a theoretical model to demonstrate that unstructured information is essential for evaluating risk in a peer-to-peer lending market, making it a highly relevant source of information for default prediction. The value of unstructured information is empirically supported by Iyer et al. (2016), who exploit that early business models of peer-to-peer lending platforms used auctions to determine the interest rate that a borrower is required to pay and the respective debtholders perception of the risk of the exposure. The authors find that the assessment of the peers is more predictive than formal credit scores in terms of R^2 and the area under the curve (AUC).[1] Furthermore, Berg et al. (2020) use digital footprints (such as device type, operating system, or web-form filling behavior) from an e-commerce company to predict payment defaults. The precision of a simple logistic regression based on these digital footprints is similar to credit bureau scores. The model further captures credit-relevant information not present in these scores. Moreover, Lin, Prabhala, & Viswanathan (2013) link the number and type of online friendships on the lending platform Prosper to the loan default rate. Óskarsdóttir, Bravo, Sarraute, Vanthienen, & Baesens (2019) use call-detail records and social network analytics and find significant increases in prediction quality of defaults. Gunnarsson et al. (2021) argue that it is of particular interest to extend the research on the use of deep learning based on unstructured data in credit scoring.

From the perspective of textual data, the business models of several peer-to-peer lending platforms, such as Lending Club and Prosper, previously included an option for prospective borrowers to include free-flow text containing self-descriptions and descriptions of the purpose of the loan in their loan applications. This is an interesting setting to assess how valuable textual information is for predicting credit defaults. Here, the key challenge for researchers and practitioners is to choose methods that facilitate the conversion of text as unstructured information into machine-usable information. One common approach is to extract text characteristics directly in a feature-engineering manner, based on certain pre-defined rules. Herzenstein et al. (2011) study data from Prosper to identify whether borrowers focus on descriptions of their identity or positive personality traits, such as being hardworking, trustworthy, or successful. Dorfleitner et al. (2016) use data from two large European peer-to-peer lending platforms, Auxmoney and Smava, to derive characteristics from user-generated descriptions, such as spelling errors, text length, and mentioning positive emotions. Chen et al. (2018) study data from Renrendai, a Chinese platform regarding the use of punctuation, which is considered to be a measure of impatience, and less formal writing. Gao et al. (2021) investigate readability, positivity, and deception cues in loan descriptions on Prosper. Tsai & Wang (2017) extract words that carry sentiment from financial reports. Similarly, Agarwal, Chen, & Zhang (2016) examine the linguistic tone of credit-rating action reports published by S&P Global Ratings.

Another common approach to dealing with textual data is based on using machine learning methods combined with word frequencies. Netzer et al. (2019) apply ensembles of random forests and regularized logistic regressions on frequently used bigrams. Xia et al. (2020) use loan descriptions from Lending Club to first determine the most characteristic keywords of the loan description in terms of term frequency-inverse document frequency (TF-IDF). These keywords are then transformed by vector-word embeddings and are used to cluster the texts, which are then used in default prediction. Xia et al. (2020) also include the number of words and the number of additions to the text. Their prediction is based on CatBoost, a gradient-boosting algorithm, and several competing prediction methods. Jiang, Wang, Wang, & Ding (2018), Fitzpatrick & Mues (2021), and Stevenson et al. (2021) further use topic modeling to derive information from text and include the outcome in machine learning models.

In addition to the approaches based on text characteristics and those based on machine learning, deep learning could provide a promising approach to using textual information in credit risk prediction. In their overview article, Kraus et al. (2020) highlight the importance of considering deep learning models in business-related decision-making processes. Deep learning in financial applications of operational research has thus been used in several recently published studies. Table 1 provides a summary of these studies. Several among them apply deep learning in the prediction of credit risk based on structured information such as Kvamme, Sellereite, Aas, & Sjursen (2018), Chen, Guo, & Zhao (2021), and Sadhwani, Giesecke, & Sirignano (2021) in mortgage lending, Mahbobi, Kimiagari, & Vasudevan (2021) for credit card data, Fitzpatrick & Mues (2021) for peer-to-peer lending, and Gunnarsson et al. (2021) for multiple datasets. Aside from lending decisions, deep learning has been used in option valuation (Cao, Liu, & Zhai, 2021), stock return predictions (Krauss, Do, & Huck, 2017, Fischer & Krauss, 2018, Huck, 2019, Flori & Regoli, 2021, or Sermpinis, Karathanasopoulos, Rosillo, & de la Fuente, 2021), gold price movement predictions (Jabeur, Mefteh-Wali, & Viviani, 2021), trader classification (Kim, Yang, Lessmann, Ma, Sung, & Johnson, 2020), card fraud prediction (Seera, Lim, Kumar, Dhamotharan, & Tan, 2021), and cryptocurrencies (Seera et al., 2021).

In fact, recent research has also used deep learning techniques to extract credit-relevant information from corporate disclosures for credit risk prediction. Mai et al. (2019) use convolutional neural networks and average embedding neural networks. The empirical results indicate that the average embedding neural networks outperform the convolutional neural networks. Matin et al. (2019) study management statements and auditor reports to derive default predictions using a convolutional recurrent neural network. Ahmadi et al. (2018) process entire annual reports from a large sample of German firms. Extracting information from long pieces of text, such as annual reports, is particularly challenging. Ahmadi et al. (2018) solve this problem by filtering particularly relevant passages of text using dependency-sensitive convolutional neural networks. Stevenson et al. (2021) further suggest the use of transformer neural networks using multilingual BERT based on loan officer reports. We contribute to this literature based on these very interesting results by conducting a benchmark study in the style of Lessmann et al. (2015), Fitzpatrick & Mues (2021), and Gunnarsson et al. (2021) for extracting credit-relevant information from textual data. If the information contained in text is valuable, it is a crucial question how this information can be exploited and which machine learning algorithm to use.

From the authors perspective, there are three relevant research gaps that need to be addressed. First, given that Dorfleitner et al. (2016) and Chen et al. (2018) did not find clear evidence that text

[1] AUC is a common performance metric for assessing the quality of credit-default predictions. AUC is explained in more detail in Section 4.5.

Table 1
Literature review deep learning in financial applications of operational research. This table presents a summary of work on the use of deep learning in financial applications in operational research journals. Abbreviations: AE - average embedding neural network, CONV - convolutional neural network, CONVREC - convolutional recurrent neural network, DBN - deep believe network, DRL - deep reinforcement learning, LSTM - long short-term memory neural network, MLP - multilayer perceptron, NN - neural network, REC - recurrent neural network.

Author (year)	Data	Prediction task	Deep learning methods
	Deep learning on structured information - lending		
Chen et al. (2021)	Mortgage data	Early delinquency	MLP
Fitzpatrick & Mues (2021)	Peer-to-peer lending data	Profit scoring	MLP
Gunnarsson et al. (2021)	Multiple datasets	Credit scoring	MLP, DBN
Kvamme et al. (2018)	Mortgage data	Credit scoring	CONV
Mahbobi et al. (2021)	Credit card data	Credit scoring	MLP
Sadhwani et al. (2021)	Mortgage data	Multiple targets	MLP
	Deep learning and textual data - lending:		
Ahmadi et al. (2018)	Annual reports and financials	Credit scoring	Dependency-sensitive CONV
Mai et al. (2019)	Annual reports and financials	Credit scoring	CONV, AE
Matin et al. (2019)	Annual reports and financials	Credit scoring	CONVREC
Stevenson et al. (2021)	Proprietary corporate lending	Credit scoring	Multilingual BERT
	Deep learning on structured information - other financial applications		
Cao et al. (2021)	S&P 500 option data	Option evaluation	Gated NN
Fischer & Krauss (2018)	S&P 500 stock data	Stock returns	LSTM, MLP
Flori & Regoli (2021)	S&P 500 stock data	Stock returns	LSTM
Huck (2019)	US large cap stock data	Stock returns	DBN
Jabeur et al. (2021)	Gold price data	Gold price movements	MLP
Kim et al. (2020)	UK spread trading data	Trader classification	MLP
Krauss et al. (2017)	S&P 500 stock data	Stock returns	MLP
Schnaubelt (2022)	Cryptocurrency data	Optimal limit orders	DRL
Seera et al. (2021)	Payment card data	Fraud detection	MLP
Sermpinis et al. (2021)	DJIA, NASDAQ 100, and NIKKEI 225 stock data	Stock index returns	MLP, REC
Wu, Chen, Yang, & Tindall (2020)	Hedge fund data	Hedge fund returns	MLP

is informative of default risk in consumer lending, it is imperative to investigate how much important information these texts contain for private customers. Further, Xia et al. (2020) find only a low value for textual information. A better understanding of this issue is important for practitioners to determine whether to use these texts for decision-making. Second, given that the information in text is potentially valuable, there is the additional question of how to exploit this information methodologically. While there is extensive research on exploiting structured data, there are almost no studies that compare methods for unstructured data, such as texts. There are several approaches to extracting textual information that have been suggested: deep learning Ahmadi et al., 2018, Mai et al., 2019, Matin et al., 2019, Stevenson et al., 2021; alternative machine learning approaches Jiang et al., 2018, Xia et al., 2020, and rule-based approaches Herzenstein et al., 2011, Dorfleitner et al., 2016, Gao et al., 2021. However, there is no comprehensive study that compares the performance of these approaches. Third, given that there are several different deep learning architectures available (e.g., convolutional neural networks and average embedding neural networks, Mai et al. (2019), convolutional recurrent neural networks, Matin et al. (2019), BERT, Stevenson et al. (2021), but also other methods such as recurrent neural networks or RoBERTa), which sometimes substantially differ in complexity, one is interested in which methods to use to include text in credit scoring applications.

Our key contributions are addressing these three research gaps. We investigate whether user-generated text from peer-to-peer lending can be used as a predictor of credit defaults. We find that these texts contain substantial credit-relevant information. Based on an information fusion approach, we evaluate that the importance of this information is high compared with other available information. In terms of methods, we compare six deep learning architectures with machine learning approaches based on word frequencies, topic models, and pre-trained word embeddings, along with several rule-based text characteristics. When comparing the deep learning methods, the alternative machine learning approaches, and the rule-based text characteristics, deep learning methods outperform the alternatives in almost all cases. However, when choosing between different approaches, machine learning methods based on word frequencies also provide substantial performance. Considering the choice between deep learning architectures, the average embedding neural networks, convolutional neural networks, convolutional recurrent neural networks, and transformer neural networks (BERT and RoBERTa) achieve similar prediction quality. This is interesting, as some of these differ considerably in complexity. Recurrent neural networks show a lower performance than the other architectures.

3. Data description

In this study, we focus on the information contained in user-generated text for credit default predictions. Previous research on text characteristics and credit default predictions often used data from peer-to-peer lending platforms. This is an appropriate research setting, as loan applicants are often asked to provide descriptions of themselves along with the purpose of the loan. We use data from Lending Club, since it is one of the largest peer-to-peer lending platforms in the US. Their data is publicly available through their website.

Lending Clubs business model relies on providing potential lenders and borrowers with a platform for loan applications. During the application process, borrowers fill in generic information related to their loan applications, such as home-ownership status, employment, and income level. For several years, borrowers could further strengthen their case by adding a brief description of themselves and the reason for their loan request. Lending Club then calculates a credit score based on the application information and external credit bureau scores. Then, the loan rate is determined by Lending Clubs credit score. Lenders then check the loan application information, including loan descriptions, credit scores, and loan rates. Based on this information, lenders choose whether to lend money to a certain borrower. Usually, lenders diversify their exposure by lending small amounts to multiple borrowers. In this way, borrowers can obtain funding if the information they provide

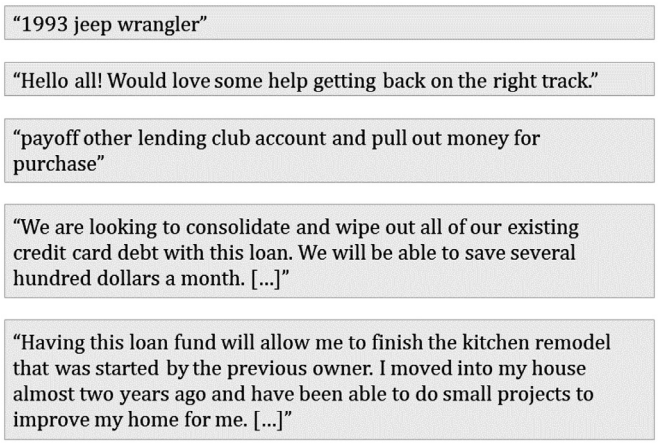

Fig. 1. Examples of loan-application descriptions. This figure displays some user-generated descriptions from loan applications in the Lending Club dataset.

attracts a sufficient number of lenders. The data provided online by Lending Club contains structured and textual loan application information, credit scores, loan rates, and several other variables, including credit status.

Our sample was collected over the period from 2007 to 2014. This period was chosen because the data became available in 2007, and Lending Club removed the option to provide loan descriptions in 2014 owing to privacy concerns. 125,798 of the loans in the database contain loan descriptions. To avoid any issues with incomplete credit relationships, we consider only loans that have either been fully paid or charged-off.[2] In addition, we restrict the sample size to loan applications with descriptions that include at least 40 words. This is done to ensure that loan descriptions actually contain substantial information.[3] The full sample size in our analysis comprises 40,229 funded loans.

As an example, Fig. 1 displays a few user-generated descriptions from loan applications in the Lending Club dataset. The application texts range from very short descriptions of the loans purpose—such as the car model the borrower intends to buy (e.g., 1993 jeep wrangler)—to more extensive descriptions of the purpose of the loan and the borrowers financial situation, as shown in the fourth and fifth examples in the figure.

We perform several preprocessing steps to prepare the text for analysis. The first step is to clean the textual data from automatically created logs and punctuation. Then, numbers and dollar signs ($) are replaced with their corresponding words. Then, all letters are changed to lower case,[4] and the resulting words are mapped to their inflected forms (lemmatization). As part of the deep learning model, the words are further transformed into semantic vector representations. Since the number of input variables must be constant for all observations, we limit the number of words per observation to 115, which is in the 95% quantile of the length of all descriptions in the sample. This ensures that most information is retained even after data preparation. Descriptions with fewer words are zero-padded, i.e., we add zero vectors to the end of texts with fewer than 115 words. This ensures that all observations have the same length without changing their contents.

In one step of the analysis, we are interested in whether the information in the text is relevant when considering the information already contained in credit scores and structured application data. We include commonly accepted determinants of credit defaults in peer-to-peer lending based on related research, and we rely on the comprehensive set of variables used in Fitzpatrick & Mues (2021). These variables are listed in Table 2 and Fig. 2. They show descriptive statistics for numerical and dichotomous variables and relative label frequencies for categorical variables. 85.5% of the loans were paid back in full, while 14.5% were charged-off.

4. Methods

Our methodological goal is to predict credit defaults based on user-generated text. Traditional methods of credit default prediction typically use logistic regressions or other classifiers based on a small set of numerical or categorical variables (see Section 2). Given the nature of textual information, such approaches are not directly applicable to this case. Instead, we benchmark six deep learning architectures with machine learning approaches (based on word frequencies, topic modeling, and word embeddings), and rule-based text characteristics, respectively. The selected deep learning models are partly based on Mai et al. (2019), Matin et al. (2019), and Stevenson et al. (2021), but also include other architectures such as the state-of-the-art transformer model RoBERTa. The machine learning approaches are based on Jiang et al. (2018), Mai et al. (2019), Netzer et al. (2019), Fitzpatrick & Mues (2021), and Stevenson et al. (2021). The rule-based text characteristics are based on Herzenstein et al. (2011), Dorfleitner et al. (2016), Chen et al. (2018), Netzer et al. (2019), Xia et al. (2020), and Gao et al. (2021). This section describes the design of our analysis and it explains the methods in detail.

4.1. Deep learning methods

We employ a set of six different deep learning architectures (convolutional neural networks, recurrent neural networks, convolutional recurrent neural networks, average embedding neural networks, BERT, and RoBERTa) to extract credit-relevant information from text. In general, an artificial neural network comprises several layers of neurons that transmit information from the input layer—via one or several hidden layers—to the output layer.

The deep artificial neural networks applied in this paper have multiple hidden layers that enable the models to aggregate information gradually from multi-dimensional inputs into higher-level features (LeCun, Bengio, & Hinton, 2015). Given this structure, deep learning is especially successful in solving the complex tasks involved in computer vision, audio recognition, and natural language processing.

Basic artificial neural networks called dense neural networks (or multilayer perceptrons when consisting of multiple hidden layers) are characterized by the fact that each neuron in a layer is connected to each neuron in the next layer. For complex classification, this group of fully connected neural networks is computationally inefficient. Therefore, other types of neural networks that do not have fully connected layers of neurons, e.g., convolutional neural networks as used in Mai et al. (2019), have been developed. Convolutional network architectures facilitate the gradual aggregation of information from multi-dimensional inputs into higher-level features (LeCun et al., 2015). First, a special layer creates different convolutions of small sections of the input data. That is, it creates different perspectives on the information. Then, a pooling layer discards less relevant information by focusing on the most extreme outcomes. This interplay of convolution and pooling can be repeated several times in a deep convolutional neural network. According to Goldberg (2016), convolutional neural networks are especially successful when relevant information is available but spread out over different points in the input data, which is the case

[2] According to Lending Club, a loan is charged-off if it is at least 120 days past due and repayment cannot be expected. Earlier charge-offs are possible in case of bankruptcy notifications. We use this definition as it is determined in the data provided by Lending Club.

[3] The results are robust against differences in minimum description lengths. These results are available from the authors upon request.

[4] This step is skipped for RoBERTa since this model is pre-trained on cased text.

Table 2
Descriptive statistics for numerical and dichotomous variables. This table shows the descriptive statistics for numerical and dichotomous variables. The statistics are the arithmetic mean (mean), standard deviation (st.dev.) minimum (min), median, and maximum (max). The number of observations (N) is equal to 40,229. 33 cases of the revolving line utilization are missing. The annual income is winsorized at 0.01 and 0.99.

Variable	Dimension	Mean	St.dev	Min	Median	Max
Adverse public records	Number of records	0.071	0.314	0	0	17
Annual income	US-$ (winsorized)	69,500.280	36,826.810	18,000	60,230	230,000
Debt-to-income ratio	Payments/monthly income	15.72	7.42	0.00	15.57	36.82
Delinquencies (2 years)	Number of delinquencies	0.19	0.61	0.00	0.00	18.00
Employment length unknown	Indicator	0.023	0.150	0	0	1
Employment title unknown	Indicator	0.051	0.220	0	0	1
FICO score	Interval center	707.523	33.653	632	702	848
House price index	Year-to-year change	−5.325	10.838	−26.790	−5.080	17.870
Income verified	Indicator	0.629	0.483	0	1	1
Inquiries within 6 months	Number of inquiries	0.819	1.045	0	0	8
Installment to total income	Installment/annual income	0.007	0.003	0.0001	0.006	0.027
Loan amount	US-$	13883.42	7847.76	700.00	12000.00	35000.00
Loan term	In months	41.629	10.169	36	36	60
Length credit history	In months	171.504	79.646	36	157	684
Revolving balance	US-$	15,400.790	19,328.970	0	11,386	1,746,716
Revolving line utilization rate	Used credit/available credit	50.07	28.17	0.00	53.00	113.00
Unemployment rate	State-wide rate	8.626	1.882	2.400	8.500	14.000
Total open accounts	Number of accounts	10.449	4.633	0	10	53
Loan status	Indicator	0.145	0.352	0	0	1

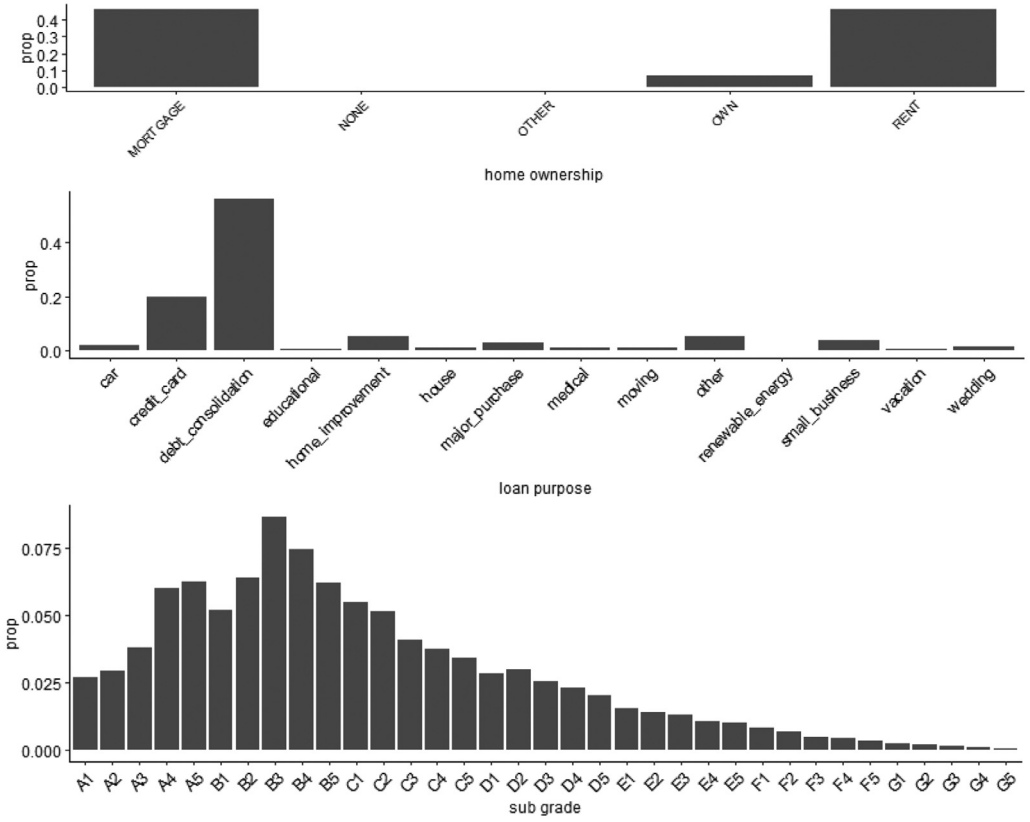

Fig. 2. Barplots of categorical feature value frequencies. This figure displays barplots for the frequency of categorical variables: home ownership, loan purpose, and rating sub grade.

for textual data. The specific network architecture used here for the deep convolutional neural network combines an embedding layer that translates the input (text) data into numerical vectors[5] with several convolutional and pooling layers and subsequent dense layers. This structure allows the model to aggregate the textual information into a set of higher-level features and then freely aggregate this information into a default probability. We also use dropout in our model. It randomly deactivates individual connections between neurons in each iteration of the fitting process as a regularization method. The number of deactivated connections is determined by the dropout rate. This is a common step used to reduce overfitting on the training data. As a second remedy to reduce overfit-

[5] Word embeddings map individual words into a vector space where semantically similar words have similar positions. The deep learning models learn the word embedding while fitting, which enables them to map domain-specific language and spelling errors. Domain-specific word embeddings result in better performance compared to pre-trained word embeddings.

Table 3
AUC of deep learning approaches for different sets of predictors. This table shows the out-of-sample AUC for predicting the default risk of borrowers given different sets of predictors based on the six deep learning approaches: convolutional neural networks, recurrent neural networks, convolutional recurrent neural networks, average embedding neural networks, BERT, and RoBERTa. These sets of predictors include no other predictor, only structured data without credit scores, credit scores without structured data, and structured data combined with credit scores, as listed in the first column. For each method and information set, the table lists AUC values including and not including text and the respective difference. Significance levels are calculated using a DeLong test and adjusted using a Bonferroni correction for multiple testing. They are denoted as $^{*}p_{adj.}<0.1$; $^{**}p_{adj.}<0.05$; $^{***}p_{adj.}<0.01$.

Sample of predictors	Excl. text	Incl. text	Diff. in AUC	Excl. text	Incl. text	Diff. in AUC
		CONV			REC	
Text only	*0.500*	0.612	0.112***	*0.500*	0.572	0.072***
Structured data	0.698	**0.709**	**0.011***	0.698	0.704	0.006***
Credit score	0.675	0.692	0.018***	0.675	0.682	0.008***
Credit score & structured data	0.702	**0.712**	**0.010***	0.702	0.708	0.005***
		CONV REC			AE	
Text only	*0.500*	0.600	0.100***	*0.500*	0.612	0.112***
Structured data	0.698	0.708	0.010***	0.698	0.709	0.010***
Credit score	0.675	0.689	0.014***	0.675	0.693	0.018***
Credit score & structured data	0.702	0.711	0.009***	0.702	**0.712**	**0.010***
		BERT			RoBERTa	
Text only	*0.500*	0.611	0.111***	*0.500*	**0.614**	**0.114***
Structured data	0.698	0.708	0.010***	0.698	0.707	0.009***
Credit score	0.675	0.691	0.016***	0.675	**0.693**	**0.019***
Credit score & structured data	0.702	0.712	0.009***	0.702	0.711	0.009***

ting, we save the fitted model after every epoch of training until there is no further increase in prediction quality on the validation sample. We then use the model with the best validation sample performance to make predictions for unseen data.[6] The convolutional part of the network is followed by one dense layer and one output neuron. The specific hyperparameter sets that were considered and chosen based on the validation sample performance are presented in Table A.1 in the Appendix. We randomly chose 200 combinations from the defined hyperparameter space as potential candidates for the optimal model. The optimization of these hyperparameters, which define the networks structure and training process, is described in Section 4.4.

Both dense and convolutional neural networks have limitations in that one layer can pass information only to the following layers. In contrast to such feed-forward networks, recurrent neural networks are able to pass processed information from one neuron to another within the same layer. Therefore, information such as text can be processed word by word—similar to how humans process text. This allows the recurrent network to identify important word sequences in the input. Gated recurrent unit (GRU) neural networks, developed by Chung, Gulcehre, Cho, & Bengio (2014), are a special type of recurrent network that iterates over information and manages a memory of important text passages that can be passed on to later units.[7] The recurrent architecture used consists of an embedding layer followed by GRU layers and a dense layer. The specific hyperparameter sets that were considered and chosen based on the validation sample performance are presented in Table A.1 in the Appendix.

We also employ a class of networks called convolutional recurrent networks—also used in Matin et al. (2019)—that combine convolutional layers and recurrent layers in one network. The reason for this is to aggregate individual words into relevant pieces of information that are then processed through a set of recurrent layers. This concept closely mimics how humans process text, i.e., not strictly word by word but by combining individual words into groups representing clauses, phrases, concepts, or ideas. For the recurrent layers in this type of network architecture, we also apply GRU layers. The resulting architecture builds an embedding layer followed by convolutional layers, one GRU layer, and one dense layer. The specific hyperparameter sets that were considered and chosen based on the validation sample performance are presented in Table A.1 in the Appendix.

We also use a so-called average embedding neural network that was used by Mai et al. (2019). This neural network also begins by using vector word embeddings. However, the dimensionality of the input data is strongly reduced by taking the average of word vector dimensions across words. This reduction in complexity has the advantage of potentially requiring significantly less data for training. The network architecture thus comprises an embedding layer, an averaging layer, and several dense layers. The specific hyperparameter sets that were considered and chosen based on the validation sample performance are presented in Table A.1 in the Appendix.

We further use two state-of-the-art transformer methods, BERT and RoBERTa. These are deep artificial neural networks based on so-called transformer layers that use a specific attention mechanism (Devlin, Chang, Lee, & Toutanova, 2018, Liu et al., 2019). In general, a transformer model contains several transformer layers that each consist of a self-attention layer followed by a feed-forward layer. The self-attention layer can recognize context across the input rather than relying on straight-forward sequential processing, as in recurrent neural networks for example.

These models have exceptionally high numbers of trainable parameters. Their use, therefore, relies on transfer learning, i.e., a pre-trained model is fine-tuned to be applied to the specific task at hand. Thus, in contrast to the other deep learning approaches used, the implicitly used word embedding in these models is also pre-trained. For BERT, we use the pre-trained BERT base uncased model; for RoBERTa, we use the pre-trained RoBERTa base model. To adjust these models for a classification task, we add a dropout layer and a final dense layer to each model. When BERT was originally published, it achieved state-of-the-art performance for relevant benchmark tasks: the Stanford Question Answering Dataset (SQuAD) (Rajpurkar, Jia, & Liang, 2018; Rajpurkar, Zhang, Lopyrev, & Liang, 2016) and the General Language Understanding Evaluation (GLUE) benchmark (Wang et al., 2018). While BERT and RoBERTa share the same basic structure, the main difference between these two transformer models is the adjusted and more data-intensive pre-training conducted for RoBERTa. As a result, RoBERTas perfor-

[6] These two steps are the same for all deep learning approaches.

[7] So-called long short-term memory (LSTM) neural networks follow a concept similar to GRU neural networks. The results of LSTM are weaker than those for the GRU networks presented in Section 5. These results are available from the authors upon request.

mance across benchmark tasks (SQuAD and GLUE) beats both BERT and other state-of-the-art deep learning models.

In these pre-trained models, the structure is largely fixed from pre-training. Many relevant hyperparameters are, therefore, already set. For BERT and RoBERTa, the optimized hyperparameters in our analysis are the training batch size and the type of optimizer. The specific hyperparameter sets that were considered and chosen based on the validation sample performance are presented in Table A.1 in the Appendix. We apply a grid search to cover this relevant hyperparameter space.

4.2. Machine learning methods

We compare the performance of the deep learning models to machine learning methods based on word frequencies, topic models, and word embeddings. The respective hyperparameter spaces for these models are available in Table A.1 in the Appendix. We sample 200 hyperparameter sets for each model

Focusing on the frequency of individual words and bigrams, our approach is based on Mai et al. (2019) and Netzer et al. (2019), who use TF-IDF. This relates the frequency that a word occurs in a specific document to its occurrences in all documents in the text corpus.

Considering occurrences of two adjacent words (referred to as bigrams), instead of focusing on individual word occurrences, facilitates the inclusion of some dimensions of word order when calculating relatively simple measures, such as TF-IDF. We use the TF-IDF representation of our textual data based on either individual words, also called unigrams, or bigrams. We limit the input variables to the 10,000 most-frequently used unigrams (similar to Mai et al., 2019, who use 20,000) or the 1000 most frequent bigrams (as in Netzer et al., 2019). Specifically, we include the TF-IDF matrices in regularized logistic regressions, random forests, and XGBoosting.

We also use a topic model to extract credit-relevant information from text without the use of deep learning. Topic models, such as the latent Dirichlet allocation (LDA), usually model documents as a combination of topics and the topics have a certain probability of producing a certain set of words. This works well in many applications. However, as Fitzpatrick & Mues (2021) point out, the text in peer-to-peer lending is relatively short. For such tasks, Yan, Guo, Lan, & Cheng (2013) show that typical topic models tend to perform unfavorably as they suffer from sparsity in the documents. We follow Fitzpatrick & Mues (2021) and estimate a biterm topic model (BTM) as suggested in Yan et al. (2013). As in Fitzpatrick & Mues (2021), we use a hyperparameter space of 1 to 20 topics and choose the final model based on the log-likelihood of the model based on the validation sample. This produced a topic model with 19 topics. That is, 19 topic probabilities by loan description that could be further processed in a baseline aggregation model.

In addition, we include a model based on a pre-trained Global Vectors for Word Representation (GloVe) word embedding that maps words to 50-dimensional vectors (Pennington, Socher, & Manning, 2014). This embedding is based on word co-occurrences, and it is extracted using an unsupervised learning algorithm. We average the embedding vector across words to create a method that is closely related to the average embedding neural network but does not rely on deep learning. The average embedding vectors serve as inputs in a random forest model.

4.3. Rule-based text characteristics

We further compare the performance of the previously described methods with manually engineered features based on rules that are potentially related to the specific personality traits of the loan applicant. The implementation of these features and the reasoning for using them are outlined in this section.

Following Herzenstein et al. (2011), we construct a factor measuring identity claims. Furthermore, Dorfleitner et al. (2016) propose the use of text length as a factor. A longer text can result from several circumstances. This is also used by Netzer et al., 2019 and Xia et al., 2020. Based on this argument, we also measure the length of the title, as proposed by Netzer et al. (2019). We also consider spelling mistakes, as done by Dorfleitner et al. (2016), Gao et al. (2021), and Netzer et al. (2019). We follow the procedure as in Dorfleitner et al. (2016). Regarding complexity, we count the proportion of words with more than six letters, as done by Netzer et al. (2019), and we consider the simple measure of Gobbledygook (SMOG) and the Gunning FOG index, as used by Gao et al. (2021) and Netzer et al. (2019). Gao et al. (2021) propose that the tone of a text is a relevant characteristic. To capture this, we apply a sentiment dictionary based on Loughran & McDonald (2011). According to Chen et al. (2018), we include the proportion of punctuation in the text.[8] Following the approach in Gao et al. (2021), we further construct five measures based on lists of words that are known to be associated with deception cues. Finally, as in Xia et al. (2020), we include the number of subsequent additions to the application description.

4.4. Sample division

Since complex models, such as artificial neural networks, possess many degrees of freedom, in-sample performance is inflated by overfitting the training data. Furthermore, a large set of relevant hyperparameters must be optimized to ensure appropriate ones are chosen. Thus, datasets are usually divided into three distinct sets: training, validation, and test data. The training set is used to train the model, the validation set is used to compare model performance when choosing the optimum model specifications, and the test set is used to estimate out-of-sample performance.

Therefore, we randomly divide our credit default data into three subsamples: 20,000 observations are grouped into a training set and used to train the models with different hyperparameter combinations. Another 10,000 observations are grouped into a validation set, which is used to evaluate the performance of the trained models. The model configuration with the best performance on the unseen validation data is chosen as the optimal model for its corresponding type. The remaining 10,229 observations are grouped into a test set that is used to evaluate the final out-of-sample performance of the chosen optimal models.

4.5. Controls and performance metrics

To determine the value of information in user-generated text, we analyze two different setups. In the first, we assess how valuable the text information is for predicting credit defaults alone. This is particularly relevant for a situation with sparse structured information (as discussed in Berg et al., 2020 and Stevenson et al., 2021). In the second setup, we are interested in how much information is contained in user-generated text that is not already present in the structured data and credit scores. Therefore, for this second assessment, we control for information already contained in those variables. To combine the structured information and the unstructured information, we use a logistic regression model as a baseline that includes the text-based default prediction derived from deep learning, the available structured data, and the internal

[8] Except for this step, the punctuation is removed from the loan descriptions as described in Section 3.

credit rating determined by Lending Club as independent variables. This approach is similar to that in Berg et al. (2020), who add digital-footprint variables to their structured data and to Stevenson et al. (2021) who combine text-based features derived from a topic model with structured information in regularized logistic regressions and random forests in their comparator models. In line with Berg et al. (2020), we use this rather simple aggregation model, which is still often used in the industry, to ensure a level playing field for all text-processing approaches. Section 5.5 presents respective results for more complex methods of aggregation (random forests, XGBoosting, deep neural networks). The logistic regression model is trained using data only from the validation set. This is necessary since the predictions of the text-processing models on the training data are naturally very accurate, and this would make the logistic regression rely too heavily on the in-sample default prediction from complex models. This could lead to a lower out-of-sample performance. We evaluate the relevance of the information in user-generated text by analyzing whether including the textual model prediction in the logistic regression increases the models prediction quality for unseen data.

As a metric for the prediction quality, we use the AUC. This metric is based on the so-called receiver operating characteristic (ROC) curve. It plots true-positive rates against false-positive rates for all possible decision thresholds of a given classifier. A good classifier will only slowly incur more false positives for more true positives. Therefore, the ROC curve of a good classifier deviates significantly from the diagonal line of a random prediction curve. The AUC measures the area under the ROC curve. It usually assumes values between 0.5 (random prediction) and 1 (perfect prediction).

5. Empirical results

5.1. Deep learning methods

In this section, we present the central insights from our empirical analysis.

Table 3 shows the increase in AUC values for an out-of-sample prediction of a model with textual information compared to a model without said information for four different information sets (i.e., text only, structured data, credit score, and credit score & structured data).[9] The table presents the results for the six different deep learning architectures in the individual sections.

Compared to not including any other predictors, the deep learning predictions improve the AUC in Table 3 by between 7.2% and 11.4%, which is both economically and statistically significant ($p<0.01$). The significant improvement in AUC is consistent when credit scores, structured data, or both are included in the models. The increases in AUC with the addition of deep learning prediction are between 0.8% and 1.9% compared to predictions based on the credit score, and between 0.6% and 1.1% compared to predictions based on structured data alone. Even when compared to a model that includes both structured characteristics and credit scores, the increase is substantial and significant with values between 0.5% and 1.0%. These results show that all six deep learning architectures extract substantial credit-relevant information that can further supplement credit scores and structured information.

When comparing the performance of the deep learning architectures, RoBERTa outperforms the others for settings with no other predictors and with only the credit score. This is well in line with RoBERTa being the most advanced of these methods. The average embedding neural network and the convolutional neural network have the next best performance. For the settings with only structured information, the convolutional neural network has the best performance. For structured information and credit score, the average embedding network and the convolutional neural network show the overall best performance. Except for the recurrent neural network, the results are mostly on a similar level.

5.2. Machine learning methods

In this section, we compare the performance of the deep learning models with those of alternative approaches. First, we use a set of benchmark machine learning models based on word frequencies. This is inspired by both Netzer et al. (2019) and Mai et al. (2019). We then use topic models and word embeddings, following Fitzpatrick & Mues (2021) and Stevenson et al. (2021).

Table 4 shows the results for the six models based on word frequencies. We fit a random forest, an XGBoosting model, and a regularized logistic regression, each based on the TF-IDF of the most frequent unigrams, as in Mai et al. (2019), and the most frequent bigrams, as in Netzer et al. (2019). The table then shows results for a biterm topic model, following Fitzpatrick & Mues (2021). We further fit a random forest on average pre-trained embeddings.

When assessing the prediction performance, all eight models have a significant and relevant AUC for the model including text alone. When comparing the models with the unigrams as used in Mai et al. (2019) with the bigrams as used in Netzer et al. (2019), the models fitted on the unigrams yield a stronger performance. Overall, the random forest using the unigrams has the highest performance for the model including text alone. The regularized logistic regression with unigrams and the biterm topic model achieve a similar performance. Some of these benchmark models are able to outperform the recurrent neural networks, but they do not outperform the other deep neural network architectures. This is similar when including the credit score and the structured data as additional information. However, the increase in AUC is insignificant in some cases.

To have a more aggregated assessment of the performance of these models in comparison with the deep learning models, we further use the concept of information fusion. This is a relatively new approach that is used, for example, in Oztekin, Delen, Turkyilmaz, & Zaim (2013), Sevim, Oztekin, Bali, Gumus, & Guresen (2014), Oztekin, Kizilaslan, Freund, & Iseri (2016), Oztekin (2018), and Kim et al. (2020). This approach enables us to fuse the information that is extracted by the individual deep learning architectures on the one hand and the alternative machine learning methods on the other hand in a brief and concise fashion. The procedure is described in Kim et al. (2020). More specifically, different model results are fused by setting weights to the individual model outputs. In the following, this is done using the validation sample performance as weights.

These results are presented in Table 5. Interestingly, the information fusion model for deep learning performs better than the individual models. This indicates that these models capture different information from the text, at least to some extent. The model that includes text has an AUC that is 12.3% higher than the model with no other information. The model including the textual information in addition to the credit score and structured information increases the AUC by 1.2%. The information fusion alternative machine learning model also performs better than the individual models. The information fusion results for machine learning methods, overall, do not outperform the information fusion of deep learning results but show a good performance in this setting as well.

5.3. Information fusion variable importance

Information fusion further allows us to rank the usefulness of the text as extracted from the deep learning models compared to

[9] The value of 0.500 as a comparison for the model with text only is the baseline random AUC performance.

Table 4
AUC of machine learning approaches based on word frequencies for different sets of predictors. This table reports the out-of-sample AUC for predicting the default risk of borrowers, with and without textual information, given different sets of predictors individually using several machine learning approaches. The sets of predictors are: no other predictors and structured data and credit scores. Significance levels are calculated using a DeLong test and adjusted using a Bonferroni correction for multiple testing. They are denoted as $^{*}p_{adj.}<0.1$; $^{**}p_{adj.}<0.05$; $^{***}p_{adj.}<0.01$.

Sample of predictors	Excl. text	Incl. text	Diff. in AUC	Excl. text	Incl. text	Diff. in AUC
Word frequencies - Log. Regr.:		Unigrams			Bigrams	
Text only	*0.500*	0.593	0.093***	*0.500*	0.589	0.089***
Credit score & structured data	0.702	**0.709**	**0.007****	0.702	0.706	0.004
Word frequencies - Random Forest:		Unigrams			Bigrams	
Text only	*0.500*	**0.598**	**0.098*****	*0.500*	0.588	0.088***
Credit score & structured data	0.702	**0.709**	**0.007** **	0.702	0.706	0.004
Word frequencies - XGBoosting:		Unigrams			Bigrams	
Text only	*0.500*	0.587	0.087***	*0.500*	0.578	0.078***
Credit score & structured data	0.706	0.706	0.004	0.705	0.705	0.003
Topic Model - Biterm Topic Model:						
Text only	*0.500*	0.595	0.095***			
Credit score & structured data	0.702	0.708	0.006*			
Word embedding - Average Embedding:						
Text only	*0.500*	0.573	0.073***			
Credit score & structured data	0.702	0.707	0.005**			

Table 5
AUC of information fusion models. This table reports the out-of-sample AUC for predicting the default risk of borrowers, given different sets of predictors individually, using an information fusion model based on deep learning and an information fusion model based on other machine learning approaches to extract textual information. For each method and information set, the table lists AUC values with and without text and the respective difference. Significance levels are calculated using a DeLong test and adjusted using a Bonferroni correction for multiple testing. They are denoted as $^{*}p_{adj.}<0.1$; $^{**}p_{adj.}<0.05$; $^{***}p_{adj.}<0.01$.

Sample of predictors	Excl. text	Incl. text	Diff. in AUC	Excl. text	Incl. text	Diff. in AUC
Information fusion		Deep learning			Machine learning	
Text only	*0.500*	**0.623**	**0.123*****	*0.500*	0.617	0.117***
Credit score & structured data	0.702	**0.714**	**0.012*****	0.702	0.711	0.009***

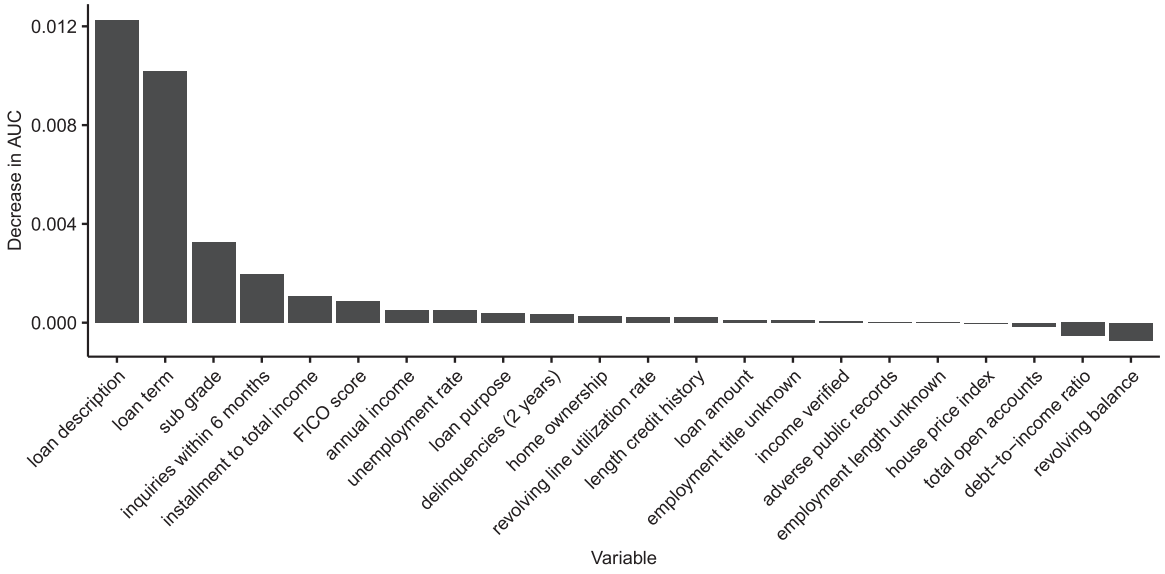

Fig. 3. Barplot of variable importance. This figure displays barplots for the variable importance by measuring the decrease in AUC when individual variables are removed from the deep learning information fusion model.

the other variables in the full model. This is presented in Fig. 3. That figure displays the decreases in AUC when features from the full information fusion model are excluded. The figure shows a decreasing order. One can see from these results that the text ("loan description") is the most influential information based on this variable importance measure. This exceeds the change in AUC that is incurred by removing the term of the loan, which comes next and even the credit score which comes after that as the third bar in the plot.

5.4. Rule-based text characteristics

We further analyze how deep learning and alternative machine learning predictions perform compared to established measures generated from textual information. The choice of these features is outlined in Section 4.3. Table 6 shows the increases in AUC when the rule-based text characteristics are included in a model with no other predictors and a model that includes structured information and the credit score. In the absence of structured data and the

Table 6
AUC for rule-based text characteristics using different sets of predictors. This table reports the improvement in out-of-sample AUC for predicting the default risk when including different rule-based text characteristics, given different sets of predictors. The predictor sets are no other predictors and structured data combined with credit scores. Significance levels are calculated using a DeLong test and adjusted using a Bonferroni correction for multiple testing. They are denoted as $^{*}p_{adj.}<0.1$; $^{**}p_{adj.}<0.05$; $^{***}p_{adj.}<0.01$.

Text characteristic	Identity claims	Word count	Spelling errors	Complexity	Sentiment
Text only	−0.017	0.032***	0.032***	−0.005***	0.005
Credit score & structured data	−0.001	0.000	**0.002**	0.000	0.000
Text characteristic	Punctuation	Deception cues	Subsequent additions		Combined
Text only	0.009	0.001	0.026***		**0.046*****
Credit score & structured data	0.000	−0.002	0.001		**0.002**

Table 7
Increase in AUC of deep learning approaches for different methods of aggregation. This table shows the out-of-sample difference in AUC for predicting the default risk of borrowers with and without textual information as included in three methods for aggregation (random forest, XGBoosting, deep neural network) when structured information and credit score are available. Significance levels are calculated using a DeLong test and adjusted using a Bonferroni correction for multiple testing. They are denoted as $^{*}p_{adj.}<0.1$; $^{**}p_{adj.}<0.05$; $^{***}p_{adj.}<0.01$.

Structured Model	CONV	REC	CONV REC	AE	BERT	RoBERTa
Random Forest: Diff. in AUC	0.009***	0.005***	0.007**	**0.011*****	0.009***	0.010***
XGBoosting: Diff. in AUC	0.009***	0.005***	0.008**	**0.010*****	0.009***	**0.010*****
Neural Network: Diff. in AUC	0.009***	0.005**	0.007**	**0.010*****	−0.001	0.006

credit score, word count, spelling errors, and subsequent additions show a significant increase in the out-of-sample AUC. However, this increase is still lower than that for the deep-learning models in Table 3 and the machine learning models in Table 4. The increase in AUC is considerable for the word count and spelling errors, at 3.2%, and the number of additions to the descriptions, at 2.6%. The increase when including all rule-based text characteristics combined is also considerable, at 4.6%. However, when structured information and the credit score are included, the out-of-sample increase in the AUC is insignificant even for a model that combines all measures.

5.5. Robustness

In our main analysis, we include the credit default prediction from the deep learning models in a logistic regression model. The intuition behind this approach is that we aimed to include the outcome of the deep learning prediction in a relatively simple model to fully focus on the comparison of the text-processing methods. Table 7 presents results for more flexible learners (random forests and deep neural networks as used in Stevenson et al., 2021 and XGBoosting).[10] Deep neural networks as used in Stevenson et al. (2021) provide a convenient approach to directly concatenate text and structured information in one model. When looking at the increase in AUC for adding textual information to the respective models including structured information, the increase is on a similar level or above for the random forest and XGBoosting compared with logistic regression results. The increase is qualitatively similar for most of the textual deep learning architectures, but it is lower for the transformer architectures. This particularly applies to BERT.

Another aspect that we address is to check whether random variations in the performance measurements are relevant to the results. To ensure robustness, we calculate the results as presented in Section 5.1 over five cross-validation runs. This analysis is available in Table 8. The table lists the mean changes in AUC for the five runs for the six deep learning models. The average performance over the five runs is similar to the performance in the main analysis in Section 5.1.

[10] The hyperparameter spaces for the aggregation models are available from the authors on request.

6. Economic value of improved predictions

In addition to the analysis in Section 5, this section aims to evaluate the economic value of the improved predictions. We use the approach recently suggested by Fitzpatrick & Mues (2021). To calculate the economic value of improved prediction as in Fitzpatrick & Mues (2021), we compose portfolios containing 100 loans with the highest expected internal rate of return for different models. The portfolio size of 100 is used because a retail investor could reasonably invest in 100 loans given a minimum investment amount per loan of $25 (Fitzpatrick & Mues, 2021). To calculate the expected internal rate of return, we first predict the probability of default. We do this for a model using the structured information and the credit score and six deep learning models including textual information along the structured information and the credit score. For calculating the expected internal rate of return, we use the mean loss given default from the training and validation samples as loan outcome in the case of default and assume that the loss is incurred equally over the period up to maturity. As is the norm in Lending Club, we use fixed monthly payments covering interest and amortization.

The results are presented in Table 9. They show the realized mean internal rate of return of the resulting portfolios. One can observe that the internal rate of return of the portfolio composed by the baseline model is 16%. All six models that included text achieve a higher return than the model without text. The results for the recurrent and convolutional recurrent neural networks lag a bit behind the other methods. At least for the recurrent neural network, this is similar to the main analysis. In contrast, BERT and the average embedding neural network display a distinctively stronger performance with returns that are higher by 1.7 and 1.6 percentage points, respectively.

Table 8
Increase in AUC of deep learning approaches in cross-validation. This table shows the out-of-sample mean difference in AUC for predicting the default risk of borrowers, given different sets of predictors based on the six deep learning approaches over five independent cross-validation runs: convolutional neural networks (CONV), recurrent neural networks (REC), convolutional recurrent neural networks (CONV REC), average embedding neural networks (AE), BERT, and RoBERTa. These sets of predictors include an empty list and structured data combined with credit scores. The second to the seventh columns list the mean differences in AUC values for the individual models over the individual runs.

Sample of predictors	CONV	REC	CONV REC	AE	BERT	RoBERTa
Text only	0.112	0.076	0.104	**0.12**	0.114	0.117
Credit score & structured data	0.009	0.004	0.009	**0.012**	0.009	0.009

Table 9
Internal rate of return of model portfolios. This table reports the mean internal rate of return of portfolios of 100 loans chosen based on different sets of information and models. The table compares values from the baseline model with no textual information and values, including the deep learning predictions from text. The procedure is based on Fitzpatrick & Mues (2021).

Model	Excl. text	Incl. text	Diff.	Model	Excl. text	Incl. text	Diff.
CONV	0.160	0.171	0.011	REC	0.160	0.166	0.006
CONV REC	0.160	0.165	0.005	AE	0.160	0.176	0.016
BERT	0.160	**0.177**	**0.017**	RoBERTa	0.160	0.171	0.011

7. Managerial implications

To sum up the results from an academic point of view, our analysis indicates that even short pieces of user-generated text are valuable for credit risk prediction in peer-to-peer lending, and deep learning is well suited to extract this information compared to alternative approaches. Furthermore, most of the deep learning architectures achieve a similar performance (comparing simpler and complex architectures). This section serves to discuss the managerial implications of these results. Our findings have direct implications for both borrowers and lenders in peer-to-peer lending as well as for the platforms themselves. Regarding the borrowers, text seems to be a way to convey information about credit quality. On the other side, we find that lenders can use this to better distinguish between good and bad borrowers and thus increase their profitability, according to Section 6. These results are important as they provide empirical evidence for arguments such as those by Liu et al. (2020). Liu et al. (2020) argue that peer-to-peer lending platforms have an advantage over traditional lenders in being more able to process unstructured information. The results are in favor of strengthening this aspect of peer-to-peer lending and to think of additional ways to include unstructured information. This is important as including this information could come at a cost to platforms such as the effort to meet data privacy standards, since this is particularly challenging in unstructured data. Considering the results of Dorfleitner et al. (2016), it has not been clear whether short pieces of text contain credit-relevant information. Following the arguments by Stevenson et al. (2021), text could be particularly important when not much other information is available. Iyer et al. (2016) consider a prediction quality of 60% in AUC a decent model in an information-sparse environment. Most deep learning models based only on text in our analysis exceed this threshold. This shows that short text is adequate on its own if no other sources of information are available.

An analysis such as this naturally relies on borrowers maintaining their behavior. This refers to the well-known Lucas critique (Lucas, 1976). What is interesting about texts is that convincing text could be a signal of creditworthiness that is difficult for an individual to imitate. The prevalence and informativeness of spelling errors in the descriptions, in fact, suggests that some customers find it difficult to use text to convince lenders to provide financing. One could think of situations where a borrower receives support in writing text from friends or relatives. Yet, even then, well-written text could still be a sign of good preparation and healthy social ties. As discussed by Fitzpatrick & Mues (2021), the opportunity to increase returns from lending using additional information is also conditional on the decision to include this information in the rating produced by a peer-to-peer platform itself. This would make it more difficult for lenders to benefit from this additional information in the future. However, it would be a benefit for borrowers as they receive loan terms that are fairer and more appropriate for their level of risk.

It is further interesting to discuss the implications of these results outside of peer-to-peer lending, such as private lending in banks. The results show that text produced by private individuals provides additional credit information. One could, therefore, think of banks evaluating text written by their customers, such as customer communication, social media posts, or text in transactions. Such further data sources are outside of the scope of this paper, but it would be very interesting for further research to analyze such data. For banks that plan to make use of these data sources, a particularly relevant outcome of this study are the implications for the choices of methods. Among the methods that we apply, deep learning produces the highest quality of prediction in almost all cases. However, average embedding, as a relatively simple deep learning architecture, achieves similar if not better results than much more complex transformer architectures. In this way, banks that want to make use of this data could begin by using simpler deep learning architectures for modeling and then check whether increased performance is possible by using more complex methods. In addition, our results have implications for model development outside of deep learning. While the alternative machine learning methods generally did not achieve a similar performance as most deep learning models, the results are still noteworthy. A method such as regularized logistic regression paired with word frequencies further has very nice properties in terms of interpretability.

8. Conclusion

There is a large body of literature that addresses the choice of variables and methods for predicting credit defaults. The ability to collect, store, and process unstructured data has recently become increasingly available to banks and fintech companies. This has allowed researchers to begin evaluating how valuable this unstructured data is and how best to exploit it. This paper assesses the value of credit-relevant information in user-generated text on Lending Club. It analyzes six deep learning architectures, alterna-

tive machine learning approaches, and rule-based text characteristics in terms of their ability to predict the default probability from textual information. Gunnarsson et al. (2021) recently pointed out the importance of conducting more research on deep learning methods to include unstructured information in credit scoring. Several recent studies have discussed the value of textual information in credit scoring. To the best of our knowledge, this study is the first account benchmarking various approaches, in particular including various deep learning architectures, for this task.

The results indicate that user-generated text is valuable for predictions. Using an information fusion approach, we find that the textual information is particularly relevant compared with other pieces of information. Regarding the choice of method, deep learning tends to perform better than alternative methods in almost all cases. However, some alternative approaches, such as machine learning based on word frequencies, show very promising results as well. When comparing the different deep learning architectures, most perform comparably well. This is interesting because architectures such as average embedding neural networks are rather simple but seldom used, while BERT or RoBERTa introduce a substantial increase in complexity and computational effort. It is particularly interesting to look at simpler and more complex methods, since many datasets used in credit scoring do not necessarily contain multiple million datapoints as big data is often described (Kraus et al., 2020) and which might be necessary to fit very complex methods; although, they might require big data methods because of the complexity of unstructured data. Because the peer-to-peer lending text tends to be short, simpler methods might be sufficient to capture the content of these texts as the good performance of the average embedding model indicates. As a side-note, spelling mistakes may be a relevant feature to consider. The results are important for academics and practitioners to assess how substantial the improvement in prediction quality and profitability is when unstructured data is included in default prediction models and to decide which methods to choose for this purpose.

Declaration of Competing Interest

None.

Appendix A

Table A.1
Hyperparameter search spaces. This table shows the considered hyperparameter spaces for the methods used in this study. The final hyperparameter combination is chosen based on validation sample performance. The best hyperparameter value is shown in the fourth column. The optimizers are abbreviated as root mean square propagation (rmsprop) and adaptive moment estimation (adam).

Model	Hyperparameter	Parameter space	Best hyperparameter
CONV	Batch size	32, 64, 128, 256, 512, 1024, 2048, 4096	1024
	Conv layers	1,2,3	1
	Dropout rate	0 to 0.5	0.032
	Kernel size	1, 2, 3, ..., 8, 9, 10	4
	Number of filters	5, 10, 15, 20	15
	Optimizer	rmsprop, adam	adam
	Pooling size	1, 2, 3, 4, 5	2
REC	Batch size	32, 64, 128, 256, 512, 1024, 2048, 4096	32
	Dropout rate	0 to 0.5	0.150
	Optimizer	rmsprop, adam	rmsprop
	Rec layers	1,2,3	1
	Rec units	2, 4, 6, ..., 28, 30, 32	8
CONV REC	Batch size	32, 64, 128, 256, 512, 1024, 2048, 4096	256
	Conv layers	1,2	2
	Dropout rate	0 to 0.25	0.205
	Kernel size	1, 2, 3, ..., 8, 9, 10	8
	Number of filters	5, 10, 15, 20	5
	Optimizer	rmsprop, adam	rmsprop
	Pooling size	1, 2, 3, 4, 5	5
	Rec layers	1,2	1
	Rec units	2, 4, 6, ..., 28, 30, 32	30
AE	Batch size	32, 64, 128, 256, 512, 1024, 2048, 4096	512
	Dense layers	1, 2, 3, 4, 5, 6	1
	Dropout rate	0 to 0.5	0.177
	Optimizer	rmsprop, adam	adam
BERT	Batch size	4, 8, 16, 32, 64	16
	optimizer	rmsprop, adam	adam
RoBERTa	Batch size	4, 8, 16, 32, 64	128
	optimizer	rmsprop, adam	adam
Model	**Hyperparameter**	**Parameter space**	**Unigrams|Bigrams**
Reg. logistic regression	Inverse regularization strength	0 to 0.5	0.497|0.497
Random forest	Maximum depth	10 to 100 (or infinite)	87|67
	Maximum features	50 to 400	377|266
	Minimum weight fraction leaf	0.5, 0.1, 0.01, 0.001	0.001|0.01
	Number of estimators	1000 to 10,000	2596|1702
XGBoosting	fraction of features	0.05, 0.1, 0.25, 0.5, 0.75	0.25|0.5
	lambda	2^{-1}, 2^{-2}, 2^{-3}, ..., 2^{-8}, 2^{-9}, 2^{-10}	2^{-5}|2^{-6}
	Learning rate	0.005 to 0.05	0.031|0.009
	Maximum depth	2 to 20	3|2
	Minimum child weight	0.01, 0.1, 0.5, 0.75	0.01|0.5
	Number of estimators	1000 to 10,000	4646|1913
Biterm topic model	Number of topics	1 to 20	19
Average embedding model	Maximum depth	10 to 100 (or infinite)	45
	Maximum features	3 to 37	8
	Minimum weight fraction leaf	0.5, 0.1, 0.01, 0.001	0.01
	Number of estimators	1000 to 10,000	4007

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. https://doi.org/10.1287/mnsc.2015.2243.

Ahmadi, Z., Martens, P., Koch, C., Gottron, T., & Kramer, S. (2018). Towards bankruptcy prediction: Deep sentiment mining to detect financial distress from business management reports. In *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)* (pp. 293–302). https://doi.org/10.1109/DSAA.2018.00040.

Baesens, B., Setiono, R., Mues, C., & Vanthienen, J. (2003). Using neural network rule extraction and decision tables for credit-risk evaluation. *Management Science, 49*(3), 312–329. https://doi.org/10.1287/mnsc.49.3.312.12739.

Berg, T., Burg, V., Gombović, A., & Puri, M. (2020). On the rise of fintechs: Credit scoring using digital footprints. *The Review of Financial Studies, 33*(7), 2845–2897. https://doi.org/10.1093/rfs/hhz099.

Cao, Y., Liu, X., & Zhai, J. (2021). Option valuation under no-arbitrage constraints with neural networks. *European Journal of Operational Research, 293*(1), 361–374. https://doi.org/10.1016/j.ejor.2020.12.003.

Chen, S., Guo, Z., & Zhao, X. (2021). Predicting mortgage early delinquency with machine learning methods. *European Journal of Operational Research, 290*(1), 358–372. https://doi.org/10.1016/j.ejor.2020.07.058.

Chen, X., Huang, B., & Ye, D. (2018). The role of punctuation in P2P lending: Evidence from China. *Economic Modelling, 68*, 634–643. https://doi.org/10.1016/j.econmod.2017.05.007.

Chung, J., Gulcehre, C., Cho, K., & Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint arXiv:1412.3555.

Crook, J. N., Edelman, D. B., & Thomas, L. C. (2007). Recent developments in consumer credit risk assessment. *European Journal of Operational Research, 183*(3), 1447–1465. https://doi.org/10.1016/j.ejor.2006.09.100.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

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 & Finance, 64*, 169–187. https://doi.org/10.1016/j.jbankfin.2015.11.009.

Dumitrescu, E., Hu, S., Hurlin, C., & Tokpavi, S. (2022). Machine learning for credit scoring: Improving logistic regression with non-linear decision-tree effects. *European Journal of Operational Research, 297*(3), 1178–1192. https://doi.org/10.1016/j.ejor.2021.06.053.

Finlay, S. (2010). Credit scoring for profitability objectives. *European Journal of Operational Research, 202*(2), 528–537. https://doi.org/10.1016/j.ejor.2009.05.025.

Fischer, T., & Krauss, C. (2018). Deep learning with long short-term memory networks for financial market predictions. *European Journal of Operational Research, 270*(2), 654–669. https://doi.org/10.1016/j.ejor.2017.11.054.

Fitzpatrick, T., & Mues, C. (2016). An empirical comparison of classification algorithms for mortgage default prediction: Evidence from a distressed mortgage market. *European Journal of Operational Research, 249*(2), 427–439. https://doi.org/10.1016/j.ejor.2015.09.014.

Fitzpatrick, T., & Mues, C. (2021). How can lenders prosper? Comparing machine learning approaches to identify profitable peer-to-peer loan investments. *European Journal of Operational Research, 294*(2), 711–722. https://doi.org/10.1016/j.ejor.2021.01.047.

Flori, A., & Regoli, D. (2021). Revealing pairs-trading opportunities with long short-term memory networks. *European Journal of Operational Research, 295*(2), 772–791. https://doi.org/10.1016/j.ejor.2021.03.009.

Gao, Q., Lin, M., & Sias, R. W. (2021). Words matter: The role of texts in online credit markets. *Journal of Financial and Quantitative Analysis*. https://doi.org/10.2139/ssrn.2446114. in press

Goldberg, Y. (2016). A primer on neural network models for natural language processing. *Journal of Artificial Intelligence Research, 57*, 345–420. https://doi.org/10.1613/jair.4992.

Gunnarsson, B. R., vanden Broucke, S., Baesens, B., Óskarsdóttir, M., & Lemahieu, W. (2021). Deep learning for credit scoring: Do or dont? *European Journal of Operational Research, 295*(1), 292–305. https://doi.org/10.1016/j.ejor.2021.03.006.

Herzenstein, M., Sonenshein, S., & Dholakia, U. M. (2011). Tell me a good story and I may lend you money: The role of narratives in peer-to-peer lending decisions. *Journal of Marketing Research, 48*, S138–S149. https://doi.org/10.1509/jmkr.48.SPL.S138.

Huck, N. (2019). Large data sets and machine learning: Applications to statistical arbitrage. *European Journal of Operational Research, 278*(1), 330–342. https://doi.org/10.1016/j.ejor.2019.04.013.

Iyer, R., Khwaja, A. I., Luttmer, E. F. P., & Shue, K. (2016). Screening peers softly: Inferring the quality of small borrowers. *Management Science, 62*(6), 1554–1577. https://doi.org/10.1287/mnsc.2015.2181.

Jabeur, S. B., Mefteh-Wali, S., & Viviani, J.-L. (2021). Forecasting gold price with the XGBoost algorithm and SHAP interaction values. *Annals of Operations Research*. https://doi.org/10.1007/s10479-021-04187-w.

Jiang, C., Wang, Z., Wang, R., & Ding, Y. (2018). Loan default prediction by combining soft information extracted from descriptive text in online peer-to-peer lending. *Annals of Operations Research, 266*, 511–529. https://doi.org/10.1007/s10479-017-2668-z.

Kim, A., Yang, Y., Lessmann, S., Ma, T., Sung, M.-C., & Johnson, J. E. V. (2020). Can deep learning predict risky retail investors? A case study in financial risk behavior forecasting. *European Journal of Operational Research, 283*(1), 217–234. https://doi.org/10.1016/j.ejor.2019.11.007.

Kraus, M., Feuerriegel, S., & Oztekin, A. (2020). Deep learning in business analytics and operations research: Models, applications and managerial implications. *European Journal of Operational Research, 281*(3), 628–641. https://doi.org/10.1016/j.ejor.2019.09.018.

Krauss, C., Do, X. A., & Huck, N. (2017). Deep neural networks, gradient-boosted trees, random forests: Statistical arbitrage on the S&P 500. *European Journal of Operational Research, 259*(2), 689–702. https://doi.org/10.1016/j.ejor.2016.10.031.

Kumar, P. R., & 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. https://doi.org/10.1016/j.ejor.2006.08.043.

Kvamme, H., Sellereite, N., Aas, K., & Sjursen, S. (2018). Predicting mortgage default using convolutional neural networks. *Expert Systems with Applications, 102*, 207–217. https://doi.org/10.1016/j.eswa.2018.02.029.

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

Lessmann, S., Baesens, B., Seow, H.-V., & Thomas, L. C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *European Journal of Operational Research, 247*(1), 124–136. https://doi.org/10.1016/j.ejor.2015.05.030.

Lin, M., Prabhala, N. R., & Viswanathan, S. (2013). Judging borrowers by the company they keep: Friendship networks and information asymmetry in online peer-to-peer lending. *Management Science, 59*(1), 17–35. https://doi.org/10.1287/mnsc.1120.1560.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. arXiv preprint arXiv:1907.11692.

Liu, Z., Shang, J., Wu, S.-y., & Chen, P.-y. (2020). Social collateral, soft information and online peer-to-peer lending: A theoretical model. *European Journal of Operational Research, 281*(2), 428–438. https://doi.org/10.1016/j.ejor.2019.08.038.

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

Lucas, R. E. (1976). Econometric policy evaluation: A critique. *Carnegie-Rochester Conference Series on Public Policy, 1*, 19–46. https://doi.org/10.1016/S0167-2231(76)80003-6.

Mahbobi, M., Kimiagari, S., & Vasudevan, M. (2021). Credit risk classification: An integrated predictive accuracy algorithm using artificial and deep neural networks. *Annals of Operations Research*. https://doi.org/10.1007/s10479-021-04114-z.

Mai, F., Tian, S., Lee, C., & Ma, L. (2019). Deep learning models for bankruptcy prediction using textual disclosures. *European Journal of Operational Research, 274*(2), 743–758. https://doi.org/10.1016/j.ejor.2018.10.024.

Matin, R., Hansen, C., Hansen, C., & Mølgaard, P. (2019). Predicting distresses using deep learning of text segments in annual reports. *Expert Systems with Applications, 132*, 199–208. https://doi.org/10.1016/j.eswa.2019.04.071.

Netzer, O., Lemaire, A., & Herzenstein, M. (2019). When words sweat: Identifying signals for loan default in the text of loan applications. *Journal of Marketing Research, 56*(6), 960–980. https://doi.org/10.1177/0022243719852959.

Oztekin, A. (2018). Information fusion-based meta-classification predictive modeling for ETF performance. *Information Systems Frontiers, 20*, 223–238. https://doi.org/10.1007/s10796-016-9704-4.

Oztekin, A., Delen, D., Turkyilmaz, A., & Zaim, S. (2013). A machine learning-based usability evaluation method for elearning systems. *Decision Support Systems, 56*, 63–73. https://doi.org/10.1016/j.dss.2013.05.003.

Oztekin, A., Kizilaslan, R., Freund, S., & Iseri, A. (2016). A data analytic approach to forecasting daily stock returns in an emerging market. *European Journal of Operational Research, 253*(3), 697–710. https://doi.org/10.1016/j.ejor.2016.02.056.

Pennington, J., Socher, R., & Manning, C. (2014). GloVe: Global vectors for word representation. In Q. C. R. I. Alessandro Moschitti, G. Bo Pang, & U. o. A. Walter Daelemans (Eds.), *Proceedings of the 2014 Conference on empirical methods in natural language processing (EMNLP)* (pp. 1532–1543). Stroudsburg, PA, USA: Association for Computational Linguistics. https://doi.org/10.3115/v1/D14-1162.

Rajpurkar, P., Jia, R., & Liang, P. (2018). Know what you don't know: Unanswerable questions for SQuAD. arXiv preprint arXiv:1806.03822.

Rajpurkar, P., Zhang, J., Lopyrev, K., & Liang, P. (2016). SQuAD: 100,000+ questions for machine comprehension of text. arXiv preprint arXiv:1606.05250.

Sadhwani, A., Giesecke, K., & Sirignano, J. (2021). Deep learning for mortgage risk*. *Journal of Financial Econometrics, 19*(2), 313–368. https://doi.org/10.1093/jjfinec/nbaa025.

Schnaubelt, M. (2022). Deep reinforcement learning for the optimal placement of cryptocurrency limit orders. *European Journal of Operational Research, 296*(3), 993–1006. https://doi.org/10.1016/j.ejor.2021.04.050.

Seera, M., Lim, C. P., Kumar, A., Dhamotharan, L., & Tan, K. H. (2021). An intelligent payment card fraud detection system. *Annals of Operations Research*. https://doi.org/10.1007/s10479-021-04149-2.

Sermpinis, G., Karathanasopoulos, A., Rosillo, R., & de la Fuente, D. (2021). Neural networks in financial trading. *Annals of Operations Research, 297*, 293–308. https://doi.org/10.1007/s10479-019-03144-y.

Sevim, C., Oztekin, A., Bali, O., Gumus, S., & Guresen, E. (2014). Developing an early warning system to predict currency crises. *European Journal of Operational Research, 237*(3), 1095–1104. https://doi.org/10.1016/j.ejor.2014.02.047.

Shilakes, C. C., & Tylman, J. (1998). Enterprise information portals.

Stevenson, M., Mues, C., & Bravo, C. (2021). The value of text for small business

default prediction: A deep learning approach. *European Journal of Operational Research, 295*(2). https://doi.org/10.1016/j.ejor.2021.03.008.

Tsai, M.-F., & Wang, C.-J. (2017). On the risk prediction and analysis of soft information in finance reports. *European Journal of Operational Research, 257*(1), 243–250. https://doi.org/10.1016/j.ejor.2016.06.069.

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. R. (2018). GLUE: A multi-task benchmark and analysis platform for natural language understanding. arXiv preprint arXiv:1804.07461.

Wu, W., Chen, J., Yang, Z. B., & Tindall, M. L. (2020). A cross-sectional machine learning approach for hedge fund return prediction and selection. *Management Science, 67*(7), 4577–4601. https://doi.org/10.1287/mnsc.2020.3696.

Xia, Y., He, L., Li, Y., Liu, N., & Ding, Y. (2020). Predicting loan default in peer–to–peer lending using narrative data. *Journal of Forecasting, 39*(2), 260–280. https://doi.org/10.1002/for.2625.

Yan, X., Guo, J., Lan, Y., & Cheng, X. (2013). A biterm topic model for short texts. In *WWW '13: Proceedings of the 22nd international conference on World Wide Web* (pp. 1445–1456). New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/2488388.2488514.

Óskarsdóttir, M., Bravo, C., Sarraute, C., Vanthienen, J., & Baesens, B. (2019). The value of big data for credit scoring: Enhancing financial inclusion using mobile phone data and social network analytics. *Applied Soft Computing, 74*, 26–39. https://doi.org/10.1016/j.asoc.2018.10.004.