Reflections on Text Mining Approaches in Corporate Failure Prediction based on German Financial Statements

Dissertation

for the award of the degree

Doctor rerum politicarum (Dr. rer. pol.)

of the Georg-August-Universität Göttingen

within the doctoral program *Wirtschaftswissenschaften*

of the Göttingen Graduate School of Social Sciences (GGG)

Submitted by

Tobias Nießner

from Eschwege

Göttingen, 2023

Examination Commitee

First supervisor:	Prof. Dr. Matthias Schumann
Second supervisor:	Prof. Dr. Jan Muntermann
Third supervisor:	Prof. Dr. Stefan Dierkes

Date of the disputation 25th of April 2023

„Against it I would say that fair play must be given to the machine. Instead of it sometimes giving no answer we could arrange that it gives occasional wrong answers. But the human mathematician would likewise make blunders when trying out new techniques. It is easy for us to regard these blunders as not counting and give him another chance, but the machine would probably be allowed no mercy. In other words then, if a machine is expected to be infallible, it cannot also be intelligent.“

Turing’s Lecture to the London Mathematical Society on 20 February 1947

Acknowledgments

First, I would like to thank my doctoral supervisor *Prof. Dr. Matthias Schumann* in particular, as he gave me the opportunity to do my doctorate and supported me in every situation throughout my time at the chair, and above all gave me his trust. Furthermore, I would like to thank *Prof. Dr. Jan Muntermann*, the second reviewer of this thesis, as well as *Prof. Dr. Stefan Dierkes* for the role of the third reviewer.

I would also like to thank my colleagues, current and former, *Dr. Julian Busse*, *Christian Finke*, *Dr. Pascal Freier*, *Michael Groth*, *Philipp Hartmann*, *Dr. Sebastian Hobert*, *Christine Jokisch*, *Dr. Raphael Meyer von Wolff*, *Mustafa Pamuk*, *Dr. Henrik Wesseloh* and *Dr. Steffen Zenker* for countless conversations in our hallway, support in daily work and university teaching. I also thank *Nicole Fiedler-Gries* and *Paula Elena Pascaru*, who assisted me through their work in the secretary's office. I am grateful for the experiences, and I appreciate your role in making it a valuable and memorable period of time in my life.

I am deeply grateful to Stefan and Daniel for their valuable contributions to my research. Through numerous discussions, they provided valuable insights that helped shape and refine many of the ideas presented in this dissertation. Their critical thinking and willingness to reconsider various aspects of my work have been instrumental in bringing forth thoughtful approaches. Thanks for working on countless lines of code that underlie this dissertation, even though they remain invisible to the reader.

I express my gratitude to my parents, *Erhard* and *Heike*, as well as my grandparents *Fritz* and *Erna,* for their unwavering support not only throughout my academic journey but also in every aspect of my life. You enabled me to pursue this path, and I am forever grateful.

My heartfelt thank goes to *Liza*, for being a constant source of support throughout my life. Your encouragement and reminders to focus in what truly matters have been invaluable to me. I am happy to have you by my side!

Finally, I would like to thank my friends who have been significant in helping me find distraction and distance from everyday university life.

Göttingen, in winter term 2022/23 Tobias Nießner

Abstract of Contents

Table of Contents

List of Figures

List of Tables

List of Abbreviations

ACM ..... Association for Computing Machinery
AI ..... Artificial Intelligence
AIS ..... Association for Information Systems
AMCIS ..... Americas Conference on Information Systems
BPNN ..... Back Propagation Neural Network
BPW ..... Bannier-Pauls-Walter dictionary
CMNN ..... Cerebellar Model Neural Network
CRISP-DM ..... Cross Industry Standard Process for Data Mining
DAX ..... German share index
EBIT ..... Earnings before interest and taxes
ESMA ..... European Securities and Markets Authority
FNR ..... False-negative rate
FS ..... Financial Statements
FPR ..... False-positive rate
FRES ..... Flesch-Reading-Ease
GridSearchCV ..... GridSearch cross validation
HGB ..... Handelsgesetzbuch
HTML ..... Hypertext Markup Language
IEEE ..... Institute of Electrical and Electronics Engineers
IFRS ..... International Financial Reporting Standards
IT ..... Information Technology
k-NN ..... k-Nearest Neighbor
LASSO ..... Least Absolute Shrinkage and Selection Operator
LIX ..... Läsbarhetsindex
MI ..... Mutual Information
MLP ..... Multilayer perceptron
NACE nomenclature statistique des activités économiques dans la Communauté européenne
NB ..... Naive Bayes
NLP ..... Natural Language Processing
NLTK ..... Natural Language Toolkit
NVM method ..... Nickerson-Varshney-Muntermann method

PACIS ..... Pacific Asia Conference on Information Systems
POS ..... Part-of-Speech
RC ..... Relevance Criteria
ROI ..... Return on Investment
RQ ..... Research Question
SentiWS ..... SentimentWortschatz
SMOTE ..... Synthetic Minority Over-sampling Technique
SOM ..... Self-organizing map
SVM ..... Support Vector Machine
XGBoost ..... eXtreme Gradient Boosting
XML ..... Extensible Markup Language

A Foundations

1 Motivation

"Can we bridge the gap, rather than sever the link, between traditional ratio 'analysis' and the more rigorous statistical techniques which have become popular among academicians in recent years?"

(Altman 1968)

While the introductory quote of this dissertation comes from a publication that, at its time in the late 1960s, initiated the shift towards multivariate statistical models in research concerning the prediction of corporate bankruptcies, its statement is more relevant today than it has been for a long time reflecting the AI winters in history. Current research now has the technical possibilities to look at how more complex artificial intelligence (AI) methods can be used to improve the accuracy of predictive models for an external analysis based on financial metrics extracted from balance sheets (Kirkos 2015; Veganzones and Severin 2021). However, technological progress not only opens up the use of greater computing capacities but also raises the question of whether and to what extent a statistical model must still be limited to a specific data source (Appiah et al. 2015; Jones 2017). If we reflect on the disclosure obligations regarding the preparation of a balance sheet and income statement of companies in relation to their annual financial statements (FS), it quickly becomes clear that these figures are intended to quantitatively represent the past (Nießner et al. 2021). Since our expectations are limited by the existence of past developments, the question arises to what extent we can also evaluate future developments of a company. Concerning this, there seem to be two different approaches to solve such a problem. On the one hand, the manual or artificial generation of data by experts or algorithms and their assessment to simulate unprecedented situations (Zięba et al. 2016) and, on the other hand, the exploration and integration of additional data into such models, which provide additional explanations for the development of companies (Jones 2017). This problem of missing data in relation to entirely new events is

particularly evident in times of crisis due to the dependence of models on financial ratios. There is a lack of valid approaches for integrating the influence of possible external events on the company's business. A large number of researchers agree with regard to the increasing demand of practice for more and more accurate models that a certain potential for bridging this problem lies in the approaches of text mining, which have become popular among academics nowadays (Kloptchenko et al. 2004a; Nassirtoussi et al. 2014; Gupta et al. 2020; Myšková and Hájek 2020). It is assumed that on the one hand, due to the freedom of language, but also legal obligations, assumptions, and expectations flow into the financial statements of companies, which, transformed and adapted to specific statistical models, allow more accurate and reflective bankruptcy forecasts. However, it is important to note that companies have greater incentives to emphasize positive developments and conceal negative ones in their external presentation to stakeholders (Hajek et al. 2014). The goal of the analysis is therefore to overcome the information asymmetry between the company and an external analyst in order to accurately classify its financial situation. AI-based solutions, therefore, present themselves as a way to solve appropriate modeled decision problems based on a variety of factors that are known (see Figure 1).

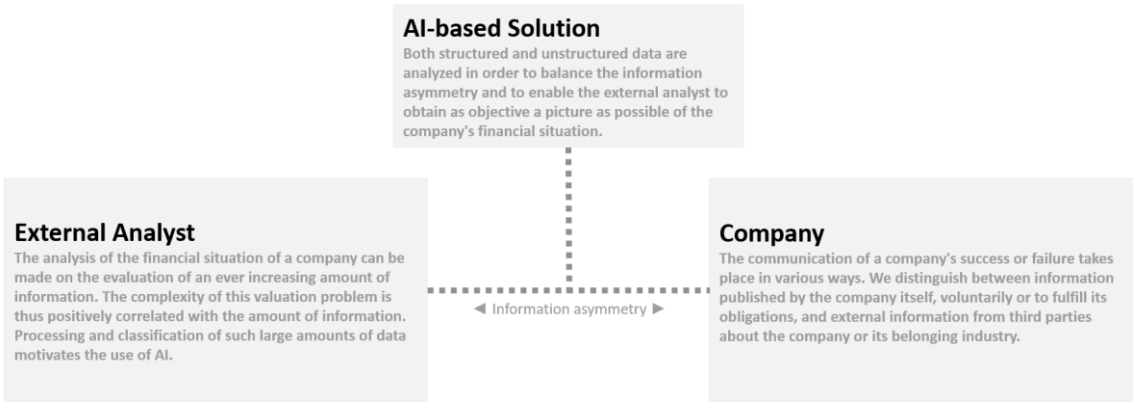


Figure 1. Integration of AI in external financial analysis

While only data that companies are obligated to publish would be considered in this respect, the consideration of externally collected data about companies also plays a role in the scientific discussion (Jones 2017). A distinction is necessary between the information published voluntarily, by duty, and that published by third parties. While both the voluntarily provided information and the mandatory published information should be viewed with a healthy skepticism on the part of

the analyst, the latter information, e.g., recently collected from social media, may open up a more comprehensive picture, as negative developments may be more readily incorporated in equal measure. Nevertheless, it should be noted that information from social media could be also manipulated, e.g., with a wrong emotional connotation of the postings, to reflect a distorted picture of reality (Deng et al. 2018). We see this today in a wide variety of areas, e.g., the U.S. presidential elections of the past years (Miller 2020) or the social discussions within the COVID-19 pandemic (Kupferschmidt 2022). Conscious of this issue, objectively verifiable information concerning the corporate structure and its change appears to be interesting as a research object. Summarizing the research point of view, there is a general interest in the development and evaluation of new approaches to extend the training data basis of statistical models that support an external financial analysis of companies.

The research interest underlying this work can also be explained by looking at developments in practice. While a look at the bankruptcy cases of German companies in recent years initially gives the impression that a downward trend can be observed, a steadily negative correlation between the number of filed bankruptcies and the resulting loss amount has been evident for years, as the latest CreditReform Semi-Annual Report 2022 states (Creditreform 2022). In particular, the influence of possible bankruptcy delays as a result of the suspended bankruptcy obligation in the wake of the COVID-19 pandemic must be critically reflected upon. This explains in particular the sharper fall in the number of bankruptcy applications from 2020, but not the fact that, according to CreditReform data, in the year of suspension 2020, bad debt losses per bankruptcy case more than doubled compared to 2019 (it was EUR 1.2 million in 2019; it was EUR 2.6 million in 2020).

2 Theoretical Background

This chapter provides an introduction to the basic principles of the studies presented in this dissertation.

First, the Cross Industry Standard Process for Data Mining (CRISP-DM) is presented as a model that allows us to summarize the results of this thesis in a structured way and to discuss them in an overall context about the use case of corporate failure prediction. Second, based on a structured literature review, the status quo from a research perspective is considered. The publications identified as relevant will be reflected with the help of the CRISP-DM. Finally, the central research questions are derived and defined from the research status thus ascertained, before a brief overview of the overall structure of the dissertation is given.

Cross Industry Standard Process for Data Mining

Wirth and Hipp (2000) describe data mining as a creative process that requires a variety of skills and knowledge. Since the success of a data mining project depends on the participants of the project, the CRISP-DM offers a framework for structuring the data mining process, which at the same time improves the reproducibility, but also the evaluation of projects. The CRISP-DM subsumes the individual steps of a data mining project into six phases, i.e., business understanding, data understanding, data preparation, modeling, evaluation, and deployment, which are independent of external factors, e.g., industry or the technologies, used. These phases are not bound to strict order, but also influence each other reciprocally and retroactively. All in all, an iterative life cycle of a data mining project is shown, which is limited by a predefined end condition for the evaluation. However, the process can be restarted from scratch at a later point in time, since solutions are not final.

In the following, we address the individual phases, as well as their assigned tasks and corresponding outcomes, in order (Shearer 2000). In this dissertation, we do not consider the deployment phase, since no prototypical implementation has been carried out. This decision is due to the lack of adequate interfaces for automated data acquisition.

Business Understanding. The first step is to create an understanding of the goals and requirements of the data mining project. It is important to create an understanding of the business background and the goals of the project at the enterprise level. Subsequently, the availability of data that can be used to potentially achieve these goals has to be examined. It should also be clarified to what extent the data can be used legally and meaningfully. It is recommended to create a list of potential risks and related solutions in order to perform a cost-benefit analysis for the project. In a next step, the fundamental question of the criterion of success of the project should be answered. For example, a defined degree of precision that a model should have is suitable for this. Based on these considerations, a project plan is created.

Data Understanding. The Data Understanding phase begins by looking at and analyzing the available data. These are compiled from one or more sources and supplemented by a description. The question of whether the data is sufficient in relation to the objectives defined in the previous phase is clarified. In particular, this consideration raises the question of evaluating the quality of the data. A focus should be placed on the detection of missing values, as well as outliers, which can occur due to the data often collected over years.

Data Preparation. Building on the results of the Data Understanding phase, this phase selects data that is suitable for achieving the business objective for further processing. The data is cleansed in terms of data quality according to the results identified in the previous phase before further insights into the data are created through the combination and exploration of data features. For this purpose, features from different data sources can be combined to generate new knowledge. Finally, the data is adjusted according to the planned use of algorithms in the following phase.

Modeling. In this phase, algorithms are selected which are used to develop models for the defined business objective using the available prepared data. In this respect, it is important to define a suitable test design for the different models to enable an evaluation of their suitability to fulfill the business success criteria from the first phase.

Evaluation. In the evaluation phase, the entire process of the first four phases is considered to decide as to which of the developed models is suitable for fulfilling the initially defined business objectives. In the case that no suitable model could be developed, the data mining process can be performed again with the knowledge of the previous iterations.

Methodology

We conducted a structured literature review to identify AI-based solutions for corporate bankruptcy prediction based on financial statements as well as challenges and opportunities for this research topic (Webster and Watson 2002; Levy and Ellis 2006; Vom Brocke et al. 2009). First, following Levy and Ellis (2006), we characterized the focus of our literature search using the taxonomy of Cooper (1988) shortened by vom Brocke et al. (2009) (see Table 1). The review examines research findings and summarizes the state of the art from a neutral perspective of the author. We examined literature in a selected range of databases to which access is available through university licenses. We derived open research questions to address and guide the research within this dissertation. In accordance with the coverage, we aim to give a representative overview of the literature stream.

Characteristic	Categories			
focus	research outcomes	research methods	theories	applications
goal	integration	criticism		central issues
perspective	neutral representation		espousal of position	
coverage	exhaustive	exhaustive and selective	representative	central/pivotal

Table 1. Classification of the literature review

Therefore, we started first using operators such as “AND” and “OR” to form composite search strings of keywords in order to search for publications in an unrestricted way (see Figure 2). In doing so, we defined the search strings in terms of the use case, the data to be considered, and the tools used to map AI in corporate bankruptcy prediction in the search process.

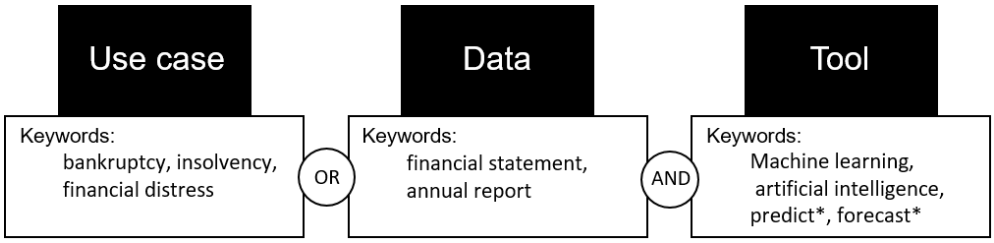


Figure 2. Search strings and combination for the review

To ensure the quality and relevance of a publication we followed predefined criteria (see Table 2) and selected only reviewed publications. These are defined according to both the current use of AI, as well as identifying approaches that have not been automated to date but have the potential to support corporate failure prediction in the future. In the first step, the search hits were checked for duplicates and relevance regarding their titles and abstracts. In the second step, the preselection of publications was reviewed in detail to decide whether they are relevant based on the predefined criteria. Papers that showed a non-binary classification scheme were classified according to the purpose named in the title according to the criteria.

Criteria	Description
RC. 1	Relevant are papers that develop AI-based solutions based on financial statements for corporate bankruptcy prediction.
RC. 2	Relevant are papers that develop AI-based solutions for predicting a financial distress in companies.
RC. 3	Relevant are papers that examine financial statements with respect to corporate bankruptcy prediction and data mining processes.

Table 2. Criteria of relevant (RC) publications

To ensure the completeness of the literature stream, we conducted a backward and forward approach by references and authors search based on Google Scholar (Levy and Ellis 2006).

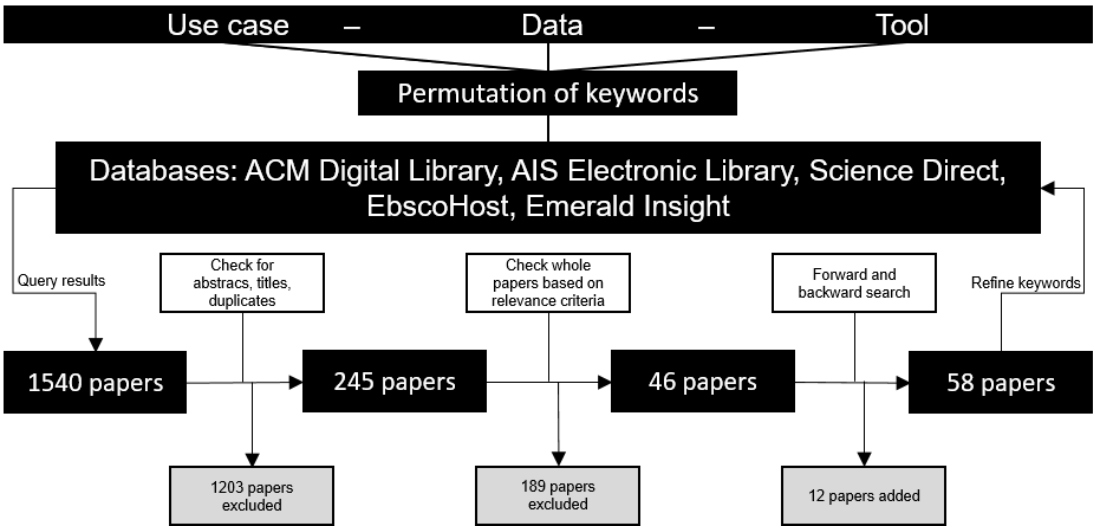


Figure 3. Literature search process

Subsequently, the search strings were refined using the identified literature to iteratively optimize the search in terms of focus (Vom Brocke et al. 2015). A tabular evaluation of the results presented in the following literature review can be seen in the Appendix (see Table 20 – Table 22). An overview of the databases and search results used is shown in Figure 3.

Literature Review

By taking the relevance criteria (see Table 2) for the literature search process into consideration, 58 relevant publications were identified. These can be divided into three different types of approaches based on the predefined relevance criteria: First, the detection of financial distress, second, the prediction of bankruptcy and third, approaches that deal with financial statement analysis with respect to both use cases but do not use AI in their research. First of all, we briefly give two definitions for a distinction between companies in bankruptcy and those in financial distress. Financial distress is defined as a state of a company that is caused by operating decisions or external forces. In contrast, bankruptcy is a possible consequence of financial distress that companies choose to protect their assets from creditors.

First, we consider literature identified according to the first relevance criterion (RC 1) that deals with the use of AI in corporate bankruptcy prediction. Olmeda and Fernandez (1997) demonstrated the development of robust ensemble learning within their study, addressing the problem of class imbalance early on

when datasets of solvent and bankrupt companies are considered. Yang and Harrison (2002) developed an approach based on probabilistic neural networks that consider clusters of companies based on the similarity of their financial situation and showed that it is suitable for detecting bankruptcies based on their dataset from the UK. Brabazon and Keenan (2004) investigated genetic algorithms and used their data from the USA to show that these, as well as neural networks, provide a better result in bankruptcy prediction than a Latent Dirichlet allocation (LDA). Furthermore, a large number of papers were identified that dealt with the comparison of different machine learning algorithms with respect to corporate bankruptcy prediction. What these publications have in common is that the use of ensemble learning provided the best results (Yim and Mitchell 2004, Behr and Weinblat 2017, Zhao et al. 2017, Zhou and Lai 2017, Wyrobek and Kluza 2018). In contrast, Hajek et al. (2014b) showed in a comparison of machine learning algorithms that a Support Vector Machine (SVM) achieves the highest precision. Merkevicius et al. (2006) combined the classical Z-Score bankruptcy prediction approach by Altman (1968) with a self-organizing map (SOM) to predict corporate bankruptcy. Huang et al. (2016) showed that fruit fly optimization can help improve the accuracy of a Z-Score model integrated into a deep learning approach. Another deep learning approach is presented by Ozkan-Gunay and Ozkan (2007) which considers bankruptcy cases of Spanish banks. Ciampi and Gordini (2013) also use a deep learning approach to look at small Italian companies. Camacho-Minano et al. (2015) take up the idea of distinguishing between different company sizes in bankruptcy prediction and consider small and medium-sized companies to identify factors for occurring bankruptcies. Wu et al. (2006) show that processed financial measures lead to an improvement in the predictive accuracy of models in company-level bankruptcy prediction compared to raw measures. An ensemble approach that uses an SVM for classification and a decision tree to derive rules for bankruptcy prediction is presented by Hsu and Pai (2013) using a dataset of Taiwanese companies. Alternatively, Chung et al. (2016) showed an approach using a cerebellar model neural network (CMNN) for the detection of bankruptcies also based on a view of Taiwanese companies. However, it must be reflected here that no ensemble learning was examined.

Kirkos (2015) concludes from a literature review regarding the use of intelligent methods for bankruptcy prediction that a major goal of research in this area is to develop new, high-performance classification methods. Subordinate research strands identified are the investigation of feature selection methods (Rustam et al. 2018, Sankhwar et al. 2020) and alternative data that support the goal of corporate bankruptcy prediction (Jones 2017, Lohmann and Ohliger 2020). Jones (2017) examined a multidimensional framework for predicting bankruptcies. Using 10K reports from the U.S., e.g., market price factors, account-based indicators, ownership concentration ratios, structural variables, analyst recommendations, credit rate changes, and macroeconomic variables were considered. Less classical variables depicting the structure of companies or representing salary distributions were shown to provide the most impact for predicting company-related bankruptcies. This is followed by financial ratios and accounting variables. Macroeconomic variables were found to be negligible. Jones et al. (2017) further recommend the use of so-called "new age" classifiers over old methods, since these firstly achieve higher accuracy, secondly are easier to implement and furthermore require less effort for the preparation of the data, and thirdly can be interpreted more easily. Following this, Tanaka et al. (2019) showed that industry-specific factors have a positive effect on company-level bankruptcy prediction by using a data set of companies that are distributed worldwide. In the examination of several identified papers, it was also noticeable that the problem of an imbalanced data set was often not addressed by oversampling, but instead smaller samples were considered, which corresponds to undersampling (Zhao et al. 2017, Zhou and Lai 2017, Inam et al. 2019). Nevertheless, there is also a recent paper that considers the use of Synthetic Minority Oversampling Technique (SMOTE) for oversampling (Roumani et al. 2020). Others, however, do not specify the distribution of bankrupt and solvent companies at all (Slimene and Mamoghli 2019). Smith and Alvarez (2021) solved the problem by using a robust gradient boosting algorithm and showed that it was able to predict bankruptcies using their 1 in 10 distributed data. Alaka et al. (2020) present a framework for developing classifiers to predict corporate bankruptcies in light of dealing with Big Data, but they disregard the distribution of the classes because they assume an equal distribution. In summary, it can be seen that 38% of the articles classify the use of neural networks and 45 % classify the use of

ensemble learning as promising over other classifiers (see Table 20). Furthermore, only 34% of the identified papers integrated the use of feature selection, only 10% considered distinguishing financial distress, and only 14% investigated the inclusion of other variables besides financial indicators.

In the following, we discuss the literature identified according to the second relevance criterion (RC 2) that deals with the use of AI in financial distress prediction. We first give a brief insight into the content covered by the literature before showing connections. Platt and Platt (2002) investigated to what extent a choice-based sample bias exists with respect to the prediction of company-related late payments. Using an equal grouping approach, they were able to show that this bias exists for a data set from the automotive industry. Bose (2006) showed that the use of rough sets for financial ratios improves classification into financially healthy and distressed companies, as the approach can especially increase the precision of classification of borderline cases. Mora et al. (2008) showed that company-specific data, e.g., the age of the company, company structure data, or events affecting the company, together with financial ratios can be used to predict financial difficulties using evolutionary algorithms and artificial neural networks. Xie et al. (2011) showed in a study that by using an SVM, you could improve the accuracy of predicting financial distress based on financial ratios. Shie et al. (2011) realize that particle swarm optimization is suitable to support feature selection on the one hand, but also parameter optimization for the use of SVM to detect financial distress based on financial ratios. Sun (2012) shows that with regard to the use of SVM, that random sample selection is suitable to reduce computational effort on the one hand and improve the accuracy of classification on the other hand. Pai et al. (2014) present an ensemble learning approach to financial distress detection that combines both SVM and a variety of decision tree algorithms. The random forest is used to reduce the complexity of the data first, whereas the actual classification is performed by an SVM. The decision tree is used afterward by users to adjust rules and simulate possible future events. Goo et al. (2016) investigate the suitability of least absolute shrinkage and selection operator (LASSO) for the feature selection process of data preparation by comparing decision trees, SVM, and neural networks. They

show that their multilayer perceptron (MLP) with a hidden layer benefits the most in terms of prediction accuracy of financial distress. The same conclusion regarding the use of MLPs was also reached by Paule-Vianez et al. (2019) when looking at Spanish banks with regard to the detection of financial difficulties. The superiority of neural networks over k-Nearest Neighbor (k-NN), Naïve Bayes (NB), and SVM was also demonstrated by Salehi et al. (2016), using a study that investigated Iranian companies in terms of predicting financial distress. Building on a LASSO feature selection, Huang et al. (2017) showed, using a dataset of Chinese companies, that you can achieve a better result in predicting financial distress using a random forest, i.e., ensemble learning, than by using a back propagation neural network (BPNN). The use of additional company-related variables in forecasting financial distress of Chinese companies was investigated by Jiang and Jones (2018) using a TreeNet model. In particular, in addition to financial indicators, they also considered, among others, market returns, macroeconomic indicators, valuation multiples, audit quality factors, shareholder ownership/control, executive compensation variables, and corporate social responsibility metrics. Balasubramanian et al. (2019) also use the example of Indian companies to consider other variables in addition to financial indicators. Kim (2018) uses ensemble learning in predicting financial distress to show that factors specific to the U.S. hotel industry exist. Furthermore, they also included additional features, e.g., market returns and stock-price trends. The industry-specific influence on the detection of financial distress has also been demonstrated in other studies (Farooq and Qamar 2019; Clintworth et al. 2021). A deep sentiment text mining approach for detecting financial distress using German companies is shown by Ahmadi et al. (2018), using both the quantitative and qualitative data of financial statements that were available to them. Lohmann and Ohliger (2020) examined German risk statement for distinguishing between solvent, financially distressed, and bankrupt companies. One conclusion from the comparison of the financial statements of companies from different countries is that there are country-specific differences in the presentation of results and that they must therefore be interpreted differently (Pai et al. 2014; Huang et al. 2017; Jiang and Jones 2018). Nevertheless, there is one identified approach so far that tries to develop country-independent solutions (Hsu and Lee 2020). It can be derived that all the identified publications, with the exception of the work by

Farooq and Qamar (2019), made a binary classification of the target. Thus, the distinction of the occurrence of corporate bankruptcy is not distinguished from financial distress and there is a lack of information on whether the discriminatory power of the developed models would also be suitable for distinguishing this case. With regard to the use of machine learning methods, it appears that only two of the identified younger studies involve the use of qualitative data. There is as well a trend toward the use of additional company-related variables for the prediction of financial distress (see Table 21). Ensemble (47%) and Deep Learning (16%) constitute the best classifiers, as was the case with bankruptcy prediction.

Finally, we consider the literature identified according to the third relevance criterion (RC 3), which does not deal with the use of AI, but with the analysis of financial statement data in the context of predicting the financial situation of companies. While the prediction of financial distress, and also bankruptcies, is primarily based on the classical consideration of financial ratios, the third relevance criterion helped to identify mainly papers that consider the combination of quantitative and qualitative data (Kloptchenko 2004; Caserio et al. 2020) or specifically address only the research area of text mining. Text mining of financial statements can be divided into different areas, e.g., sentiment analysis (Loughran and McDonald 2016; Myšková and Hájek 2020), language tone analysis (Magnusson et al. 2005; Luo and Zhou 2020), collocation-based approaches (Shirata et al. 2011) and dictionary-based approaches (Loughran and McDonald 2016; Chou et al. 2018). Furthermore, work has been identified that deals with text mining applications in finance and related to this, the prediction of financial difficulties (Appiah et al. 2015; Chen et al. 2016; Gupta et al. 2020).

Discussion

In the following, we use the CRISP-DM to derive challenges for the use of AI in corporate bankruptcy prediction. For this purpose, following our relevance criteria, we also consider and discuss the literature that addressed the identification of financial distress or data analysis of financial statements without developing a model.

Business understanding. The first step of the CRISP-DM model deals with the business objective that needs to be fulfilled. It is important to not only clearly define the objective but also to define appropriate criteria that are used to evaluate the approach from a business perspective (Wirth and Hipp 2000). A problem is a focus solely on improving existing solutions, as can be seen by the literature review. Since no further company-specific explicit criteria are defined for the success of the project as a consequence the business understanding phase is disregarded in the literature. In particular, also no attention is paid to the extent to which the data availability within a hypothetical information architecture is guaranteed. An example of this is the question of interfaces to external data sources for an automated training process of classifiers. Furthermore, the question of the cost-benefit evaluation of such a project has not yet been raised within the research community.

Data understanding. The phase describes the steps from the data collection to the analysis of these regarding their quality and suitability given the objective (Wirth and Hipp 2000). We have noticed that there is a trend toward using external data in addition to financial statements (Jones 2017; Lohmann and Ohliger 2020; Clintworth et al. 2021). Furthermore, there is potential in the analysis of the qualitative data of a financial statement (Magnusson et al. 2005; Hajek et al. 2014; Lohmann and Ohliger 2020; Gupta et al. 2020). However, there has been little work to date comparing the quantitative and qualitative data with respect to objectives in financial statement analysis (Kloptchenko et al. 2004a; Myšková and Hájek 2020). In terms of data quality, the problem of imbalanced data sets is addressed by researchers (Roumani et al. 2020). We have also found that the dependence of financial statements on industries and countries can have a large impact on data quality (Huang et al. 2017; Kim 2018; Jiang and Jones 2018). However, by analyzing the identified publications, we can confirm that many studies use internationally published data, as these are probably easier and faster to acquire (Tanaka et al. 2019). Overall, data acquisition is a challenge, resulting in an imbalance problem in a large number of studies.

Data preparation. The data preparation step includes all the transformations performed on the data to create a final training dataset for the model development (Wirth and Hipp 2000). In order to reduce the amount of data with regard to the relevance of individual features by means of a feature selection, current studies

exist (Sankhwar et al. 2020). However, these only consider quantitative factors and disregard the qualitative analysis of the financial statements. While the use of expert-assessed data has been examined in terms of risk factors (Lohmann and Ohliger 2020; Myšková and Hájek 2020), there is a lack of consideration of newly constructed features, which may be generated from quantitative data, qualitative data, or a mix of both.

Modeling. Typically, different machine learning algorithms are used in the modeling phase to solve the underlying problem. These are optimized on the basis of their parameters in order to compare them with regard to their suitability (Wirth and Hipp 2000). Within the identified publications, we found a trend towards the use of ensemble learning methods in the last five years' publications (67% of the publications show that ensemble learning methods achieve the best results). Nevertheless, the use of neural networks achieved the best result in 21% of the studies during this period. The selection of ensemble learning methods has the additional advantage that the explainability of the model can be probably simplified compared to a neural network as they often rely on decision tree structures (Jones et al. 2017).

	CRISP-DM					
Criteria	**Business Unterstanding**	**Data Understaing**	**Data Preparation**	**Modeling**	**Evaluation**	**Number of search hits**
RC. 1	-	○	○	●	-	**29**
RC. 2	-	○	○	○	-	**18**
RC. 3	-	○	○	-	-	**11**
Legend: ●: Research exists ○: Research partially exists -: No research exists						

Table 3. Concept matrix with respect to CRISP-DM

Furthermore, these classifiers are usually robust to incomplete or imbalanced data, which is often the case. An evaluation of the model is made in the identified

works exclusively by the improvement in comparison to other approaches. Thus, no criteria are used that were defined ante hoc in terms of business understanding.

Evaluation. The evaluation serves to ensure that a model meets the requirements from both a technical and a business perspective and that the underlying data has been integrated according to possible specifications (Wirth and Hipp 2000). We could not identify any publication that had its model evaluated from a business perspective. Furthermore, it has been shown that there is a possibility that a model will produce reliable results only for a definite period of time (Salehi et al. 2016; Paule-Vianez et al. 2019). In addition, there is also a lack of research regarding data preparation. There are also no results that address factors that foster the acceptance of AI in financial statement analysis. A ranking of research perspectives based on the CRISP-DM can be seen in Table 3. A distinction was made between the absence of research contributions, the existence of partial results, and the existence of more than 10 research contributions to a phase.

3 Organization of the dissertation

It was shown that research in the field of corporate bankruptcy prediction is strongly characterized by the quest to improve existing models. It is crucial to note that a non-technical consideration of the use of AI in this scenario is currently lacking, e.g., consideration of the quality level at which an AI-based model offers practical added value compared to a human being. Based on this premise, the first research question of this dissertation is defined as follows.

Research Area: Assessment of Corporate Failure Approaches	
RQ. 1	How should the current state of research on the use of AI in corporate failure prediction by means of financial statements be assessed?

We further noted that many identified approaches in the literature are based on considering only quantitative financial metrics in the development process. More recent research contributions, on the other hand, are more frequently concerned with the integration of textual data from financial statements or information from external data sources relating to the companies under consideration. Since there is considerable potential to develop new approaches, especially in the area of text mining, and since the question remains open to what extent a variety of possible new data influence each other, but also with regard to the goal of corporate bankruptcy prediction, the following second research question of this dissertation is defined. Since we apply our research on German companies, we limit the research question in this regard.

Research Area: Knowledge Extraction	
RQ. 2	How can quantitative financial statement analysis be supported by using textual data with respect to German companies failure prediction?

In conclusion, this thesis not only wants to comprehensively consider how diverse data sources related to financial statements influence each other with respect to the goal of predicting corporate bankruptcies. In addition, we will take a look at model development based on AI to evaluate to what extent the findings can be aggregated to derive recommendations for action that help in the construction of such text mining approaches. We define the following third research question for this purpose.

Research Area: Artificial Intelligence	
RQ. 3	What is the value of textual data in the development process of an AI-based solution in German companies failure prediction?

The dissertation is divided into three parts (see Figure 4): Part A, where the basics of the dissertation are presented, Part B, the conducted studies and in Part C, the discussion, conclusion and future perspectives. Part A first gives an introduction and motivation to the topic of the dissertation (A.1), before basics of data mining processes, as well as the research background are considered by means of a structured literature review (A.2). Finally, the structure of the thesis is presented in this part (A.3). Part B presents the individual studies that were conducted as part of the dissertation. First, a taxonomy for the use of AI in financial statement analysis is presented in relation to the research area of assessing corporate failure prediction approaches. Second, the research area of knowledge extraction is considered by means of three studies. Another study is at the interface between knowledge extraction and the of AI-based approaches.

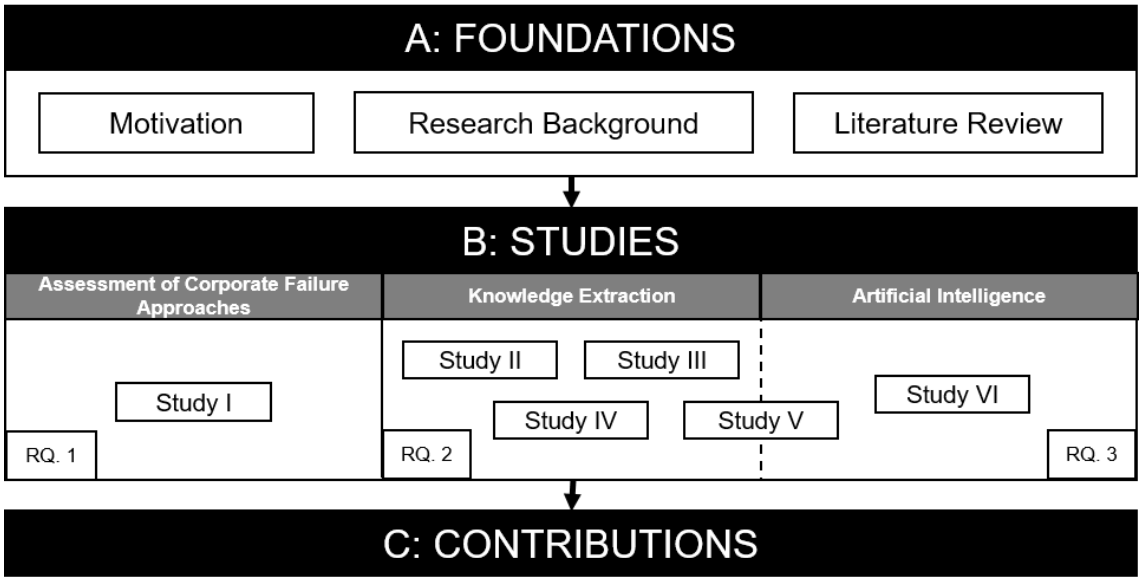


Figure 4. Structure of the dissertation

First, a novel approach to distinguish between solvent, financially distressed, and bankrupt companies based on the theory of evidential strategies is presented (Study II). Subsequently, studies that focus on looking at consecutive financial statements in the context of knowledge extraction are considered. For this purpose, a study that looks at changes in the use of language over a specific period of time with regard to the occurrence of bankruptcies of German companies (Study III) is presented, before the outcomes are reflected by a short publication considering document similarity of reports published by single companies (Study IV). A consideration of the influence of external factors, such as competition and corporate structure on the presentation of a company within a financial statement and therefore, the possibility of making bankruptcy predictions follows (Study V). Part B ends with Study VI which aggregates approaches from the previous studies and in this respect addresses the questions of the importance of text mining approaches in the development of AI-based approaches for predicting financial failure in companies based on financial statements. It also discusses how text mining approaches could be conducted more successfully compared to the status quo. The studies listed within Part B of this dissertation, which have been published at scientific conferences as well as in scientific journals, have been almost entirely adopted in their original format.

Changes were made in wording, the format and embedding of tables and figures in order to adhere a uniform dissertation.

In Figure 5 the life cycle of a data mining project according to the CRISP-DM is presented. With regard to the classification of the studies within this dissertation, the iterative nature of the model should be emphasized here once again, indicating the possibility of non-finite further development and improvement of performance. This shows that solutions for dynamic adaptations, due to changing evaluation criteria or also changing requirements, can be worked out accordingly over further iterations.

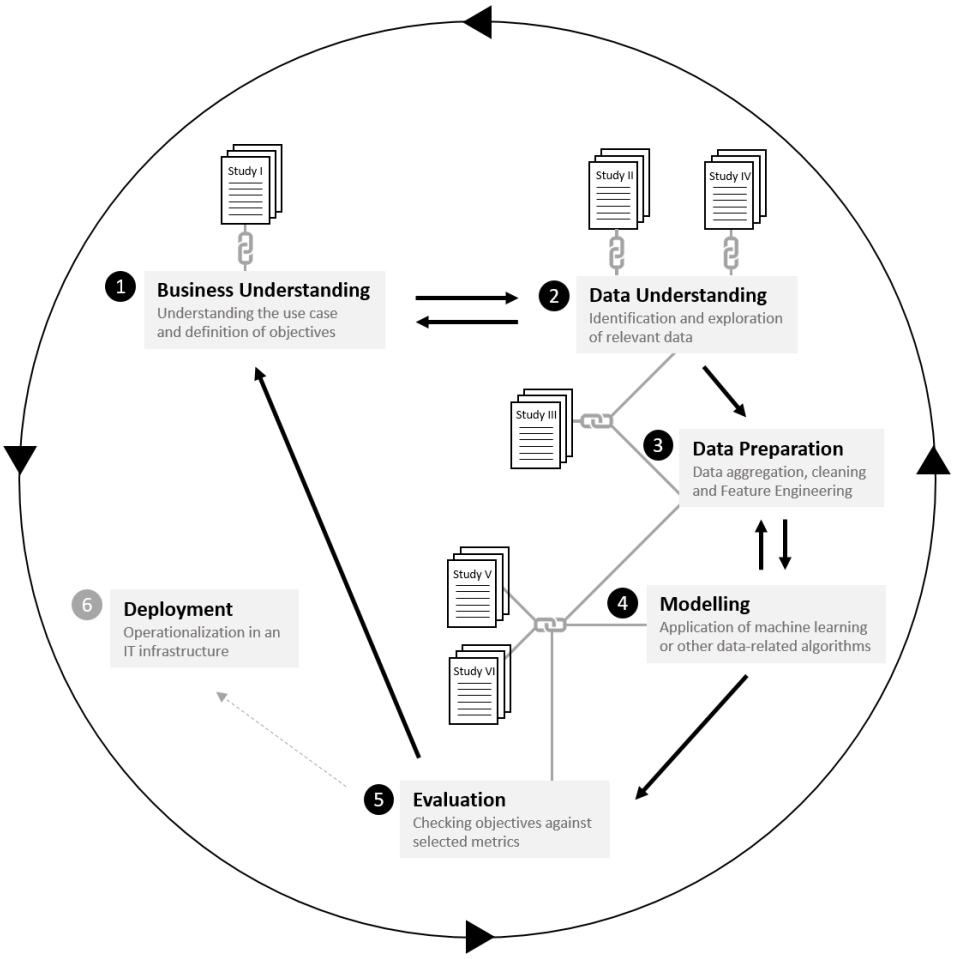


Figure 5. Linkage between conducted Studies and Phases of CRISP-DM (Wirth and Hipp 2000)

Based on the iterative process, the studies conducted can be classified according to the phases supported by their results of themselves. While the developed taxonomy can be understood as a communication tool between practice and research for the classification of approaches in the discussion of AI-based models and thus supports the phase of business understanding further studies II-VI can

be found in the more frequently considered phases of the data mining process. The studies conducted, which are presented below in Part B, can be understood in two ways. On the one hand, results are presented that can be used to improve the prediction accuracy of an AI model to be trained, but on the other hand, they also provide recommendations for the implementation of the different phases for further analysis of texts.

While in Study I a literature review was conducted to identify approaches as data objects, further Studies II-VI are based on the analysis of essentially two differentiable data sets. Study II, V, and VI refer to the analysis of a dataset of financial statements provided in XML format (Bundesanzeiger) by external partners. This dataset was linked with information about the respective companies, corresponding balance sheet ratios from the Amadeus database as well as further information from Zensus and Statista.

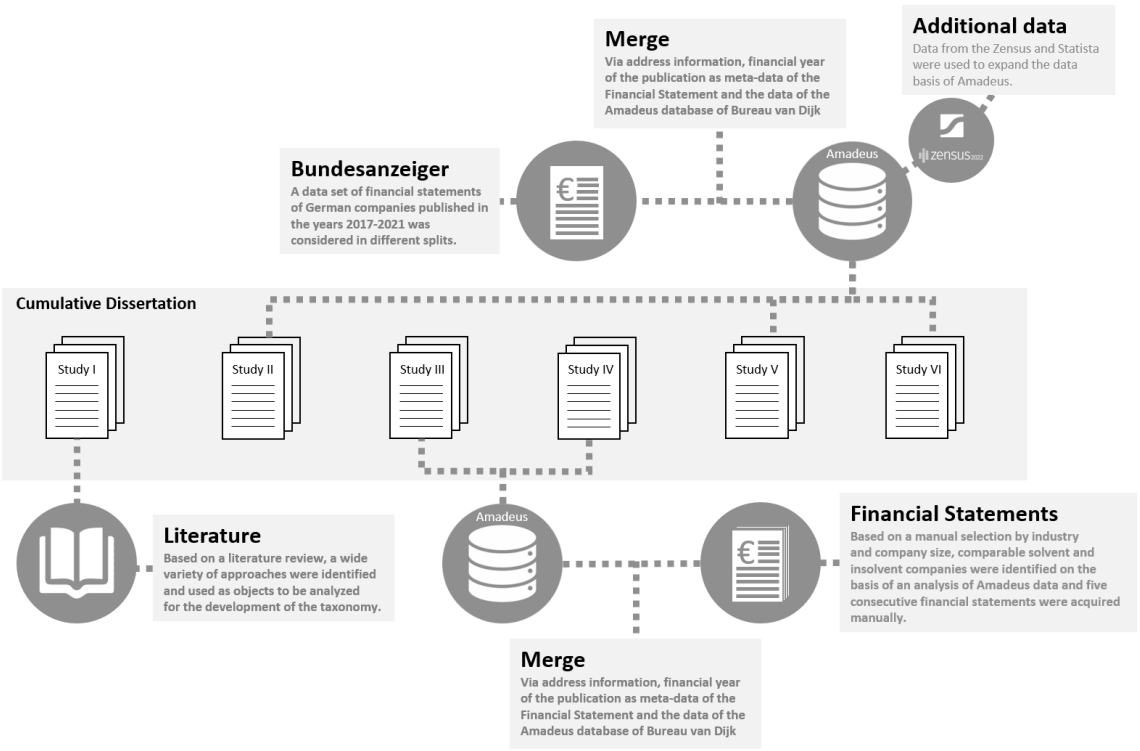


Figure 6. Data use within the conducted Studies

For this purpose, a script was written in python that performed an assignment based on a comparison of company name and address information. While 2,392,184 files were processed in the original data set, the volume was reduced

to 2,163,147 extractable financial statements, which were considered differently within the individual studies with regard to different motives (see Figure 6). It should be noted that financial statements of various sizes of companies were included and therefore corresponding limitations of the data were unavoidable due to redundant or insufficient textual information available. In addition, consecutive financial statements for Studies III and IV were downloaded in HTML format from the Bundesanzeiger website with manual effort and linked analogously using information from the Amadeus database. The selection was balanced between solvent and insolvent companies and with regard to the industry.

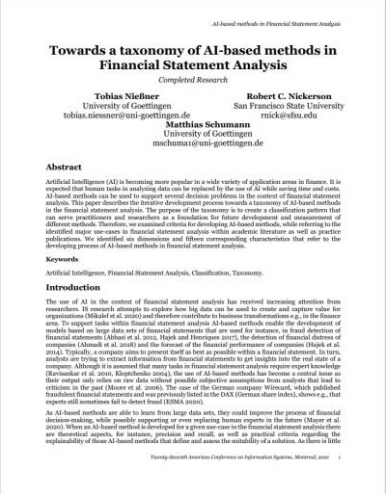
In Part C of the dissertation, the results of the research presented in Part B are discussed (C.1). Afterwards conclusions are drawn (C.2) regarding the research questions posed (A.3). The dissertation ends by pointing out limitations of the work as well as an outlook on further research perspectives (C.3).

B Studies

I Taxonomy of AI-based Methods in Financial Statement Analysis

"Towards a taxonomy of AI-based methods in Financial Statement Analysis"

August 9-13,2021 – AMCIS, Montreal Virtual Conference, Canada



AI-based methods in Financial Statement Analysis

Towards a taxonomy of AI-based methods in Financial Statement Analysis

Completed Research

Tobias Nießner
University of Goettingen
tobias.niessner@uni-goettingen.de

Robert C. Nickerson
San Francisco State University
rnick@sfsu.edu

Matthias Schumann
University of Goettingen
mschuma1@uni-goettingen.de

Abstract

Artificial Intelligence (AI) is becoming more popular in a wide variety of application areas in finance. It is expected that human tasks in analyzing data can be replaced by the use of AI while saving time and costs. AI-based methods can be used to support several decision problems in the context of financial statement analysis. This paper describes the iterative development process towards a taxonomy of AI-based methods in the financial statement analysis. The purpose of the taxonomy is to create a classification pattern that can serve practitioners and researchers as a foundation for future development and measurement of different methods. Therefore, we examined criteria for developing AI-based methods, while referring to the identified major use-cases in financial statement analysis within academic literature as well as practice publications. We identified six dimensions and fifteen corresponding characteristics that refer to the developing process of AI-based methods in financial statement analysis.

Keywords

Artificial Intelligence, Financial Statement Analysis, Classification, Taxonomy.

Introduction

The use of AI in the context of financial statement analysis has received increasing attention from researchers. IS research attempts to explore how big data can be used to create and capture value for organizations (Mikalef et al. 2020) and therefore contribute to business transformations e.g., in the finance area. To support tasks within financial statement analysis AI-based methods enable the development of models based on large data sets of financial statements that are used for instance, in fraud detection of financial statements (Alhaasi et al. 2012, Hajek and Henriques 2017), the detection of financial distress of companies (Ahmadi et al. 2018) and the forecast of the financial performance of companies (Hajek et al. 2014). Typically, a company aims to present itself as best as possible within a financial statement. In turn, analysts are trying to extract information from financial statements to get insights into the real state of a company. Although it is assumed that many tasks in financial statement analysis require expert knowledge (Ravisankar et al. 2010, Kloptchenko 2004), the use of AI-based methods has become a central issue as their output only relies on raw data without possible subjective assumptions from analysts that lead to criticism in the past (Moore et al. 2006). The case of the German company Wirecard, which published fraudulent financial statements and was previously listed in the DAX (German share index), shows e.g., that experts still sometimes fail to detect fraud (KMA 2020).

As AI-based methods are able to learn from large data sets, they could improve the process of financial decision-making, while possibly supporting or even replacing human experts in the future (Mayer et al. 2020). When an AI-based method is developed for a given use-case in the financial statement analysis there are theoretical aspects, for instance, precision and recall, as well as practical criteria regarding the explainability of those AI-based methods that define and assess the suitability of a solution. As there is little

Twenty-Seventh Americas Conference on Information Systems, Montreal, 2021 1

Authors: Tobias Nießner, Robert C. Nickerson and Matthias Schumann

Outlet: *27th Americas Conference on Information Systems (AMCIS). Montreal, Canada, 2021.*

Abstract: Artificial Intelligence (AI) is becoming more popular in a wide variety of application areas in finance. It is expected that human tasks in analyzing data can be replaced by the use of AI while saving time and costs. AI-based methods can be used to support several decision problems in the context of financial statement analysis. This paper describes the iterative development process towards a taxonomy of AI-based methods in the financial statement analysis. The purpose of the taxonomy is to create a classification pattern that can serve practitioners and researchers as a foundation for future development and measurement of different methods. Therefore, we examined criteria for developing AI-based methods, while referring to the

identified major use cases in financial statement analysis within academic literature as well as practice publications. We identified six dimensions and fifteen corresponding characteristics that refer to the developing process of AI-based methods in financial statement analysis.

Keywords: Artificial Intelligence, Financial Statement Analysis, Classification, Taxonomy.

Introduction

The use of AI in the context of financial statement analysis has received increasing attention from researchers. IS research attempts to explore how big data can be used to create and capture value for organizations (Mikalef et al. 2020) and therefore contribute to business transformations e.g., in the finance area. To support tasks within financial statement analysis AI-based methods enable the development of models based on large data sets of financial statements that are used for instance, in fraud detection of financial statements (Abbasi et al. 2012; Hajek and Henriques 2017), the detection of financial distress of companies (Ahmadi et al. 2018) and the forecast of the financial performance of companies (Hajek et al. 2014). Typically, a company aims to present itself as best as possible within a financial statement. In turn, analysts are trying to extract information from financial statements to get insights into the real state of a company. Although it is assumed that many tasks in financial statement analysis require expert knowledge (Kloptchenko 2004; Ravisankar et al. 2011), the use of AI-based methods has become a central issue as their output only relies on raw data without possible subjective assumptions from analysts that lead to criticism in the past (Moore et al. 2006). The case of the German company Wirecard, which published fraudulent financial statements and was previously listed in the DAX (German share index), shows, e.g., that experts still sometimes fail to detect fraud (ESMA 2020).

As AI-based methods are able to learn from large historical data sets, they could improve the process of decision-making, while possibly supporting or even replacing human experts in the future (Mayer et al. 2020). When an AI-based method is developed for a given use case in the financial statement analysis there are theoretical aspects, for instance, precision and recall, as well as practical criteria regarding the explainability of those AI-based methods that define an optimal solution. As there is little research on the communication of researchers and practitioners with respect to the use of AI in financial statement analysis, we argue that there is a need to increase the conceptual understanding of the development process.

We observe that the literature shows a huge diversity of AI-based methods, which are examined on a wide variety of different data (Abbasi et al. 2012). As most of

the current research is driven to develop a model that serves the intended goal better than existing methods, there is no clear focus on criteria to develop those AI-based methods so far. We identified the problem of comparing the different methods with each other in order to address not only technical aspects, e.g., accuracy, but also practical ones, that ensure usefulness in a real-world environment, e.g., explainability. However, so far no way exists that can be used to classify AI-based methods with respect to the development process in the context of financial statement analysis in order to improve the decision-making process of selecting an appropriate method.

One way of classifying AI-based methods is with a taxonomy. The use of such a taxonomy is twofold. First, both researchers and practitioners can structure the current AI methods and make them comparable to each other based on general dimensions and characteristics. Second, an AI-based method can be developed that is feasible in principle and tries to meet the given purpose as much as possible, while respecting criteria from practitioners to deploy it in a real-world environment (Benbya et al. 2020). Researchers profit from a structured view on these methods, because striving to optimize models may lead to a loss of structure in the development process. Therefore, a taxonomy offers a decision support tool for leading the development process. Derived from the motivation of structuring AI-based methods in order to support the development process within financial statement analysis, we define the following research question for this paper as follows:

RQ: How can AI-based methods that are used in financial statement analysis with respect to their development process be classified within useful dimensions and characteristics?

In the following, we first introduce the theoretical background of financial statement analysis and the integration of AI-based methods within it. Subsequently, we present and adapt the methodological approach for taxonomy development proposed by Nickerson et al. (2013) before we show the research process in detail. We then describe the development process leading to the final

taxonomy. Afterward, we discuss our results, mention limitations, and state a conclusion with an outlook on future research.

Theoretical Background

To provide a consistent understanding of what is meant by AI-based methods in financial statement analysis, we start by motivating the analysis and explaining the integration of AI-based methods within it.

Typically, the results of a financial statement analysis are used by stakeholders to evaluate the past, current, and future situation of a company. While companies try to present themselves more positively, it is the goal of a financial statement analysis to identify the true situation. Therefore, it is important to reflect that the published quantitative data within a financial statement is limited to describing the past and current financial situation, whereas qualitative data has the potential to include information that describes future developments (Ravisankar et al. 2011). For the comprehensive evaluation of the financial situation of a company, the assessment of both kinds of data is required (Hajek et al. 2014). The results of such an analysis can be seen from different angles of users. The financial assessment of a company can be, for instance, used to support the decision-making process of potential or existing investors. Furthermore, those results can serve as a base for a benchmark to measure the strengths and weaknesses of a company compared to competitors.

When the number of financial statements that need to be analyzed becomes too large, financial analysts need IT support and data mining as essential tools for support to cope with the analysis promptly (Magnusson et al. 2005; Lin et al. 2015). AI-based methods typically rely on the ability of machines to learn from and simulate human decisions. They can, if possible, use this ability to predict financial parameters or benchmark financial statements in the future.

Research Method

In order to create a taxonomy for conceptualizing knowledge about the development of AI-based methods in financial statement analysis, we decided to follow the taxonomy development process according to Nickerson et al. (2013). We chose this methodology as Oberländer et al. (2019) and Schöbel et al. (2020)

argued that the NVM-Method (Nickerson et al. 2013) is state of the art for taxonomy development as it is the first and only one that is presented in IS research. It has been applied in several other studies, such as Prat et al. (2015) and Janssen et al. (2020). Hence, we use the definition of a taxonomy T as a set of dimensions that are not empty. Each dimension (D_i) consists of any number of mutually exclusive and collectively exhaustive characteristics (C_{ij}) that is greater or equal than two. These definitions can be summarized according to Nickerson et al. (2013) with the following formula:

$$T = \{D_i, i=1,\ldots,n | D_i = \{C_{ij}, j=1,\ldots,k_i, k_i \geq 2\}\}$$

In the first step, the meta-characteristic of the taxonomy needs to be specified as a basis for the choice of characteristics. This choice ensures that the identified characteristics can be summarized as a logical consequence of the meta-characteristics and reflect the purpose of the taxonomy (Nickerson et al. 2013). As the purpose of this taxonomy from the user's perspective is mainly to structure AI-based methods in financial statement analysis in order to develop them, the meta-characteristic of the taxonomy is determined as *criteria for the choice of AI-based methods (Step 1)*. Within this iterative methodology design, empirical-to-conceptual (inductive) as well as conceptual-to-empirical (deductive) approaches can be distinctively used to evolve the taxonomy. The empirical-to-conceptual approach uses data-derived objects (Step 4e) to identify and derive common characteristics and dimensions. Hence if there is domain knowledge from the literature or the authors available (Step 4c), the conceptual-to-empirical approach can be used to create or refine the set of characteristics and dimensions in the taxonomy (see Figure 7).

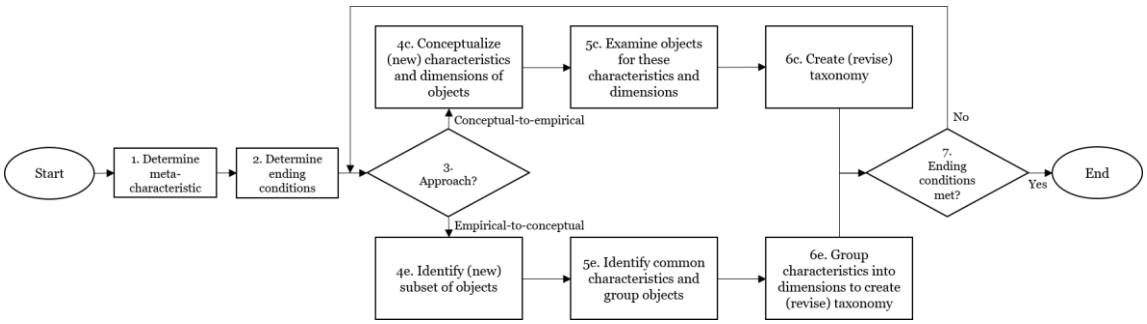


Figure 7. Iterative taxonomy development process by Nickerson et al. (2013)

According to Hevner et al. (2004), the design of this methodology can be seen as a search process. The methodology, therefore, does not claim to develop the optimal taxonomy for a given problem, but rather one that is seen as useful. Consequently, the crucial point in evaluating the developed taxonomy is to discuss its usefulness. Nickerson et al. (2013) ensure the usefulness of the taxonomy by defining objective and subjective ending conditions for the iterative search process. From a subjective point of view, a taxonomy needs to be concise, robust, comprehensive, extendible, and explanatory to be useful, not only in the current state of research and practice but also in future considerations (Nickerson et al. 2013). We only used a subset of the proposed objective ending conditions. In this regard, we decide to apply the condition that, no new dimensions or characteristics were added in the last iteration and also that no existing dimensions or characteristics will be merged or split in the last iteration. We additionally apply the condition that every dimension is unique and not repeated as well as that every characteristic is unique within its dimension (Nickerson et al. 2013). Related to the potential use of the taxonomy as a structuring tool to distinguish AI-based methods, we decided that is not necessary that at least one AI-based method is classified under every characteristic of every dimension (Step 2). This ensures the expandability of our taxonomy for future research and applications.

Research Process

We start by giving a brief summary of arguments that determine the choice of iterative methods within the taxonomy development process (Step 3). As we stated, there is a lot of unspecific knowledge on the properties of AI-based methods in diverse domains. Therefore, we deemed it beneficial to start with a conceptual-to-empirical iteration (1st Iteration) to derive a rough structure. This structure serves as a foundation for the following empirical-to-conceptual approach (2nd Iteration), which reflects use cases in the research literature for AI-based methods in financial statement analysis. In order to refine the taxonomy, we conducted an additional conceptual-to-empirical approach (3rd Iteration), which reflects additional knowledge on AI-based methods with respect to practice-oriented literature. In the last conceptual-to-empirical approach we focus on discussing the actual state of the taxonomy (4th Iteration).

Literature Review

In order to identify empirical data that can be used within the taxonomy development process, we conducted a structured literature review based on the methodological guidelines of Cooper (1988) and Webster and Watson (2002). The scope was narrowed down to papers that describe the use of AI-based approaches in the area of financial statement analysis, hence the focus was to identify central practices or applications. The search was based on established scientific databases such as ACM Digital Library, AIS Electronic Library, IEEE Xplore Digital Library, and Science Direct. The review is limited to publications accessible in the respective database. After analyzing the title and abstract we selected 128 publications of which 26 appeared to be relevant. An additional back- and forward search led to 18 additional relevant publications. We define publications as relevant when only data from financial statements is used to address the development process of AI-based methods. There is a subset of 26 relevant publications that furthermore develop AI-based methods themselves. In summary, the literature reflects a rather exploratory field of research, as it focuses mainly on the technical evaluation of developed AI-based methods. We identified a lack of empirical knowledge on practical requirements for the use of AI. We will use the gathered publications in the 2nd iteration to evolve the taxonomy from a theoretical point of view. In order to identify not only theoretical properties, we decided to expand the literature search process not only to scientific publications but also to those, published by the Big Four audit companies (i.e., EY, Deloitte, KPMG, and PwC) to elaborate the taxonomy in the 3rd iteration from a practical perspective. We identified 3 relevant publications that can be used to examine the use of AI-based methods in financial statement analysis.

1st Iteration

To examine AI-based methods in financial statement analysis, we assume that an aggregated view on use cases, that summarize AI-based methods in classes, (Step 4c) is more beneficial to derive criteria from than a view on every single AI-based method. We assume that a conceptual view on the underlying model of an AI-based method can serve as a basis for further iterations. AI-based methods

can be developed in financial statement analysis for a predefined reason, such as solving a regression, classification, or clustering task to generate a value (Tamm et al. 2020). We can distinguish between generative and discriminative models. Generative models describe the actual distribution of each class, whereas discriminative models describe the decision boundary between different classes (Jebara and Meila 2006). Furthermore, the process of training a model often referred to as “learning”, can be distinguished into supervised learning, unsupervised learning, and semi-supervised learning (Step 5c). The taxonomy after the 1st Iteration is defined as follows (Step 6c):

> T = {Model (generative, discriminative), Learning (supervised, unsupervised, semi-supervised), Task (regressive, classifying, clustering)}

The current state of the taxonomy is neither concise and explanatory nor robust, as the derived dimensions as well as the characteristics are not describing the use of those methods within financial statement analysis explicitly. We can only state that the taxonomy is comprehensive and extendible because the derived dimensions can be used to classify all AI-based methods in the financial statement analysis we know. As we added dimensions and characteristics, we can summarize that neither the subjective nor the objective ending conditions are met after this iteration. Therefore, a 2nd iteration must be conducted (Step 7).

2nd Iteration

In the following, we describe the empirical-to-conceptual approach. Therefore, we derive use cases of AI-based methods in financial statement analysis from the literature review. Regarding the literature, we found that a detailed consideration of the different algorithms does not seem to be useful, as we can assume further development and new approaches in the future that expand the current explorative research. The current literature deals with the use of a variety of algorithms to deliver a proof-of-concept for the use of AI-based methods for a specific purpose (Abbasi et al. 2012, Hajek and Henriques 2017, Ngai et al. 2011). We identified the following sample of use cases for AI-based methods in financial statements (Step 4e) from the literature:

- AI in Fraud identification (Humpherys et al. 2011)
- AI in Bankruptcy prediction (Ahmadi et al. 2018)
- AI in Financial analysis (Kloptchenko et al. 2004)

With regard to this, we identified that the type of data usage from a financial statement is a common characterization that separates AI-based methods. We distinguish between the use of features extracted from quantitative and qualitative data. From the literature, we identified several AI-based methods that use different kinds of data. For instance, financial ratios (Kirkos 2015) or text, including language analysis (Magnusson et al. 2005) are used within financial statement analysis. As Kloptchenko et al. (2004) and Chou et al. (2018) examine the use of financial ratios as well as textual data in combination, we decided for mutual exclusiveness there needs to be a third characteristic for the use of the financial statement as a whole. We also noticed that for instance, Kloptchenko et al. (2004) used an AI-based approach to analyze the performance of companies compared to other competitors. Whereas Hajek et al. (2014) focused on the analysis of one company to support stakeholders' decision-making (Step 5e). In order to deal with this phenomenon, we added the dimension of scope to our taxonomy. We identified that a company-focused and an industry-focused approach are mutually exclusive and can be distinguished. The taxonomy after the 2nd iteration is defined as follows (Step 6c):

> T = {Model (generative, discriminative), Learning (supervised, unsupervised, semi-supervised), Data (ratio-based, text-based, both), Task (regressive, classifying, clustering), Scope (company-focused, industry-focused)}

Since we added three additional dimensions in this iteration, we are following the methodology and decide to conduct a 3rd iteration for developing the taxonomy. We also think that with respect to practical use in a real-world scenario, the current taxonomy is not concise and robust enough, although it becomes more explanatory with respect to the use cases (Step 7).

3rd Iteration

In the previous two iterations, we focused on the use of theoretical knowledge of AI-based methods in general and with respect to financial statement analysis for elaborating the taxonomy. In order to further develop the taxonomy from a practical point of view, we decided that another conceptual-to-empirical approach is beneficial to identify and describe new dimensions and characteristics that are

crucial for deploying AI-based methods (Step 4c). We renamed the dimension “task” to “purpose” and specified the integrated characteristic “classifying” to “identifying”, “regressive” to “predicting”, and “clustering” to “analyzing” according to the identified use cases in the 2nd iteration. We believe that this change elaborates on the taxonomy as it is more explanatory with respect to the mentioned use cases and the functionality of AI-based methods within those. Current studies by companies show that the adoption of AI in financial services is still not common (Deloitte 2018, PwC 2020, KPMG 2019). We identify the opportunity to audit AI as crucial and challenging for the deployment of AI not only in financial services but in general (Deloitte 2018a; KPMG 2019; PwC 2020). In order to address this requirement, we defined the dimension “Transparency”. We, therefore, distinguish between explainable and not explainable AI-based methods in financial statement analysis (Step 5c). The taxonomy after the 3rd iteration is defined as follows (Step 6c):

> T = {Model (generative, discriminative), Learning (supervised, unsupervised, semi-supervised), Data (ratio-based, text-based, both), Purpose (predicting, analyzing, identifying), Scope (company-focused, industry-focused), Transparency (explainable, not explainable)}

Since we refined the taxonomy with respect to requirements from practice and specified the dimension of the task by renaming it, we argue it has become concise and robust for both researchers and practitioners. At this point we assume that the subjective ending conditions are met, but since another dimension has been added this does not apply to the objective ones. Because of that, we conduct a 4th iteration (Step 7).

4th Iteration

To rethink the current state of the taxonomy, we decided to follow an additional conceptual-to-empirical approach to address further requirements from practice (Step 4c). We find that trust and consequently acceptance are crucial aspects when AI should be deployed (PwC 2020). Regarding the financial statement analysis, we assume that trust is defined through the understanding of the different AI-based methods. We conclude that the current taxonomy is robust enough as there is no need to further expand the taxonomy in order to address

the problem of trust in AI, which is mainly influenced by the dimension of "Transparency" (Step 5c).

Final taxonomy

We define the taxonomy after the 4th iteration as follows (Step 6c):

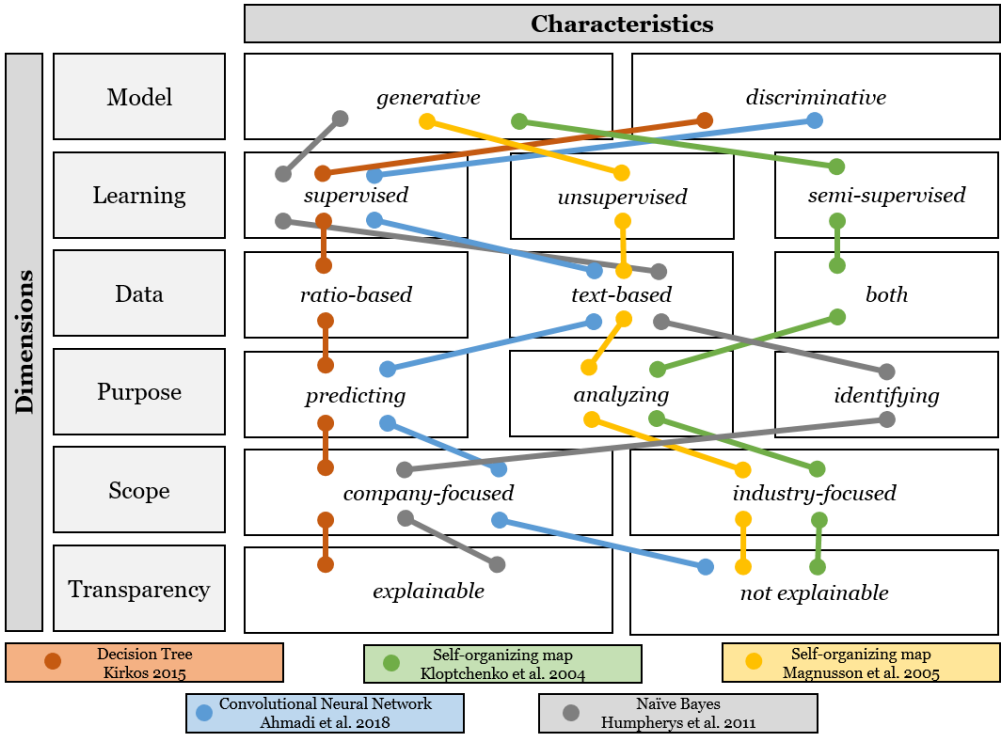


Figure 8. Final taxonomy with application examples of AI-based methods

Since we have not made any further changes to the taxonomy, additionally to the subjective ending conditions, the objective ending conditions are met. We present in Figure 6 the final taxonomy for AI-based methods in financial statement analysis (Step 7). Every characteristic is related to the decision-making process while developing an AI-based method and therefore to the meta-characteristic. The dimensions refer to key components of the AI development process. To demonstrate the application of the taxonomy, we additionally show in Figure 8 how a small subset of representative AI-based methods that are mentioned in the literature for specific use cases can be explicitly classified. Kirkos (2015) and Humpherys et al. (2011) published several methods for comparison. We decided to classify and present, e.g., the decision tree-based approach for bankruptcy prediction (Kirkos 2015) and the Naïve Bayes approach for the detection of

fraudulent financial statements (Humpherys et al. 2011). The colored lines show the respective classification of the characteristics of the AI-based method in the different dimensions.

Furthermore, we classified also the total of n=26 AI-based methods that were identified within the literature review to demonstrate an overview of the usefulness of the taxonomy.

Dimensions	Model		Learning			Data			Purpose			Scope		Transparency	
Characteristics	G	D	S	U	S-S	R	T	Bo	P	A	I	C	Br	E	NE
Used (in total)	9	17	21	4	1	16	8	2	13	7	6	20	6	15	11
Proportion (in %)	35	65	81	15	4	61	31	8	50	27	23	77	23	58	42

Table 4. Summary of the classification of AI-based methods in the literature (n=26)

If there are multiple AI-based methods presented within a publication, we classified the one that fits the given use case in the publication best with respect to the evaluation. A summary of the classification of AI-based methods in all identified publications can be seen by the distribution of the use of the characteristics in Table 4.

Limitations and Discussion

The limitations of this taxonomy development process are mainly related to the subjective nature of selecting and defining suitable dimensions and characteristics as there is always a tradeoff between simplicity and comprehensiveness (Nakatsu et al. 2014). In order to discuss this limitation, we explain the usefulness of the taxonomy for researchers and practitioners. The usefulness serves as an additional indicator for the fulfillment of the subjective

ending conditions. Since the current use of AI in financial statement analysis is more of an explorative nature, there is no structure to create an overview or differentiate between existing AI-based methods in the literature. Therefore, we examine AI-based methods based on use cases in financial statement analysis they are related to in order to identify meaningful dimensions and characteristics they are sharing.

As current academic research is largely based on benchmark datasets, there is a need for new approaches in order to deal with real-world datasets (Zhang et al. 2020). For instance, reporting standards differ from country to country and the language style may evolve. The taxonomy furthermore gives a broader overview to compare new methods with existing ones based on general criteria that are robust to changes in the reporting format. It enables researchers to adapt their development process of AI-based methods to practitioners' needs. We identified that the crucial point in financial statement analysis is to develop more precise and synonymous understandable models. Concerning the rapid development of algorithms in the field of AI, we assume it is beneficial that the taxonomy additionally refers furthermore to requirements derived from deployment in a real-world environment with regulations. Both researchers and practitioners can therefore rely on the taxonomy to weigh the respective benefits of old and new approaches. As there is an expectation that AI has the potential to change our view on how data can be used to structure and design new jobs and how decisions can be made in the future (Benbya et al. 2020), the taxonomy can furthermore serve as a communication platform for researchers and practitioners to enable a structured view on implemented AI-based methods to share ideas in the development of AI in the financial statement analysis. We identified, that there is, e.g., less research based on the use of text-based data to develop AI-based methods compared to the use of ratio-based data. Furthermore, the use of unsupervised or semi-supervised learning is less common compared to supervised learning.

As we are facing an emerging phenomenon due to the nature of AI, we expect new approaches from both researchers and practitioners to expand the taxonomy

in the future. This can be seen as an additional limitation for our research because we only examined AI-based methods that are published and discussed in the literature so far.

Conclusion and Future Research

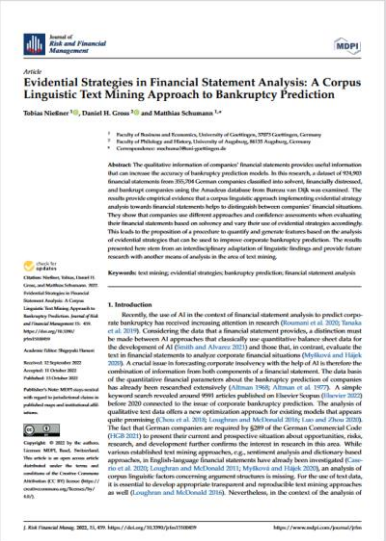
The development of AI-based methods, while fulfilling requirements from practice, plays an important role in financial statement analysis. We have created a taxonomy to increase the conceptual understanding of the development of AI-based methods in financial statement analysis following the iterative method for taxonomy development by Nickerson et al. (2013). Within the iterative process, we followed three conceptual-to-empirical and one empirical-to-conceptual approach. Therefore, we contribute to the existing literature on AI-based methods in financial statement analysis by creating and presenting a taxonomy. We conducted a brief structured literature review to identify AI-based methods and knowledge about the development of AI in financial statement analysis. We used publications from practice to complete our view with requirements for use in companies. In order to position the taxonomy within Gregor's (2006) nature of theory in IS, we classify it as a theory for analysis, because as a tool to structure, the taxonomy states "what is". We observed that the academic literature about AI-based methods in financial statement analysis is not suitably connected to the requirements of practitioners. We identified that the model and the use of data are crucial in the development of AI-based methods. While discussing the use of the taxonomy, we showed that the taxonomy and therefore the dimensions and characteristics contribute to the identified communication problem, between researchers and practitioners. Future research regarding the use of AI in financial statement analysis could focus not only on the communication itself but additionally on providing guidance for the development and deployment of AI. Since we discussed the function of the taxonomy as a communication platform, it is motivating for further empirical research on the use of AI in financial statement analysis. It would be also beneficial to create and evaluate AI-based prototypes that support decision-making in financial statement analysis in order to explicitly define and examine further requirements. Derived from the analysis conducted by Oberländer et al. (2019), the identification of archetypes as well as a lack of

those could be motivational for future research in the adaption of new AI-based methods in financial statement analysis.

II Evidential Strategies in Corporate Bankruptcy Prediction

"Evidential Strategies in Financial Statement Analysis: A Corpus Linguistic Text Mining Approach to Bankruptcy Prediction of Germany Companies"

October 2022, Journal of Risk and Financial Management



Journal of Risk and Financial Management

MDPI

Article

Evidential Strategies in Financial Statement Analysis: A Corpus Linguistic Text Mining Approach to Bankruptcy Prediction

Tobias Nießner, Daniel H. Gross and Matthias Schumann

Abstract: The qualitative information of companies' financial statements provides useful information that can increase the accuracy of bankruptcy prediction models. In this research, a dataset of 924,903 financial statements from 355,704 German companies classified into solvent, financially distressed, and bankrupt companies using the Amadeus database from Bureau van Dijk was examined. The results provide empirical evidence that a corpus linguistic approach implementing evidential strategy analysis towards financial statements helps to distinguish between companies' financial situations. They show that companies use different approaches and confidence assessments when evaluating their financial statements based on solvency and vary their use of evidential strategies accordingly. This leads to the proposition of a procedure to quantify and generate features based on the analysis of evidential strategies that can be used to improve corporate bankruptcy prediction. The results presented here stem from an interdisciplinary adaptation of linguistic findings and provide future research with another means of analysis in the area of text mining.

Keywords: text mining; evidential strategies; bankruptcy prediction; financial statement analysis

1. Introduction

Recently, the use of AI in the context of financial statement analysis to predict corporate bankruptcy has received increasing attention in research (Roumani et al. 2020; Tanaka et al. 2019). Considering the data that a financial statement provides, a distinction must be made between AI approaches that classically use quantitative balance sheet data for the development of AI (Smith and Alvarez 2021) and those that, in contrast, evaluate the text in financial statements to analyze corporate financial situations (Mythili and Hajek 2020). A crucial issue in forecasting corporate insolvency with the help of AI is therefore the combination of information from both components of a financial statement. The data basis of the quantitative financial parameters about the bankruptcy prediction of companies has already been researched extensively (Altman 1968; Altman et al. 1977). A simple keyword search revealed around 9981 articles published on Elsevier Scopus (Elsevier 2022) before 2020 connected to the issue of corporate bankruptcy prediction. The analysis of qualitative text data offers a new optimization approach for existing models that appears quite promising (Chou et al. 2018; Loughran and McDonald 2016; Luo and Zhou 2020). The fact that German companies are required by §289 of the German Commercial Code (HGB 2021) to present their current and prospective situation about opportunities, risks, research, and development further confirms the interest in research in this area. While various established text mining approaches, e.g., sentiment analysis and dictionary-based approaches, in English-language financial statements have already been investigated (Cerrito et al. 2020; Loughran and McDonald 2011; Mythili and Hajek 2020), an analysis of corpus linguistic factors concerning argument structures is missing. For the use of text data, it is essential to develop appropriate transparent and reproducible text mining approaches as well (Loughran and McDonald 2016). Nevertheless, in the context of the analysis of

J. Risk Financial Manag. 2022, 15, 459. https://doi.org/10.3390/jrfm15100459 https://www.mdpi.com/journal/jrfm

Authors: Tobias Nießner, Daniel H. Gross and Matthias Schumann

Outlet: *Journal of Risk and Financial Management 15 (2022) 10. https://doi.org/10.3390/jrfm15100459*

Abstract: The qualitative information of companies' financial statements provides useful information that can contribute increase the accuracy of bankruptcy prediction models. In this research, we examine a dataset of 924,903 financial statements from 355,704 German companies classified into solvent, financially distressed, and bankrupt companies using the Amadeus database from Bureau van Dijk. Our results provide empirical evidence that a corpus linguistic approach towards financial statements helps to distinguish between companies' financial situations. Our results show that companies seem to vary their evidential strategies and use different approaches and confidence assessments when evaluating their financial

statements based on solvency. This leads us to propose a procedure to quantify and generate features based on the analysis of evidential strategies that can be used to improve corporate bankruptcy prediction. Our results stem from an interdisciplinary adaptation of linguistic findings and provide future research with another means of analysis in the area of text mining.

Keywords: Text Mining, Evidential Strategies, Bankruptcy Prediction, Financial Statement Analysis

Introduction

Recently, the use of AI in the context of financial statement analysis to predict corporate bankruptcy has received increasing attention in research (Tanaka et al. 2019; Roumani et al. 2020). Considering the data that a financial statement provides, a distinction must be made between AI approaches that classically use quantitative balance sheet data for the development of AI (Smith and Alvarez 2021) and those that, in contrast, evaluate the text in financial statements for the purpose of analyzing corporate financial situations (Myšková and Hájek 2020). A crucial issue in forecasting corporate bankruptcy with the help of AI is therefore the combination of information from both components of a financial statement. Whereas the data basis of the quantitative financial parameters in relation to the bankruptcy prediction of companies has already been researched over a long period of time (Altman 1968; Altman et al. 1977), the analysis of qualitative text data offers a new optimization approach for existing models that appears quite promising (Loughran and McDonald 2016; Chou et al. 2018; Luo and Zhou 2020). The fact that German companies are required by §289 of the German Commercial Code (HGB 2021b) to present their current and prospective situation with regard to opportunities, risks, research and development further confirms the interest in research in this area. While various established text mining approaches, e.g., sentiment analysis and dictionary-based approaches, in English-language financial statements have already been investigated (Loughran and McDonald 2011; Myšková and Hájek 2020; Caserio et al. 2020), an analysis of corpus linguistic factors with respect to argument structures is missing. Concluding from this, for the use of text data, it is essential to develop appropriate transparent and reproducible text mining approaches (Loughran and McDonald 2016). Nevertheless, in the context of the analysis of German-language annual financial statement data, there are initial text-subdividing approaches that show how extracted information from the risk report of a financial statement can be combined with financial ratios to predict corporate bankruptcies (Lohmann and Ohliger 2020). In this paper, building on the idea of studying word collocations (Kloptchenko et al. 2004b), we consider the concept of evidential strategies in financial statements. In linguistic research evidential strategies have received more attention relatively recently. Results showed that specific evidential expressions are linked to specific discourse domains (Marín Arrese 2017) and

differences in stance strategies were observable within scientific discourse and proven to be quantifiable (Hidalgo-Downing 2017). The motivated use of these strategies has been shown to be linked to a speaker's commitment to an evaluation (Besnard 2017). We argue that these features could make the use of evidential strategies an important measurement in the examination of corporate financial performance. Our goal here is to investigate the suitability of evidential strategies as such a measurement, as well as to develop an approach to quantify them with respect to their use in the development of AI for bankruptcy prediction. Concluding from this, we formulate the following research question for our study:

RQ: How can evidential strategies in financial statements be used for corporate bankruptcy prediction?

In developing our approach based on the examination of evidential strategies, we draw from past literature on the use of textual data in financial statements for corporate bankruptcy prediction as well as from literature on the linguistic category of evidentiality to form an understanding of the corpusanalytical approach that originates in linguistics. We then present the dataset of German financial statements on which the paper is based, before presenting the qualitative analysis and our findings. Herein, we consider three classes of companies: those that are solvent, those that are financially distressed and those that are in bankruptcy proceedings. Finally, we provide an outlook for and discussion on implications of our results for theory and practice. Furthermore, we address limitations of our study as well as future research opportunities.

Theoretical Foundations

In the following, we first summarize the literature stream on the use of textual information in corporate bankruptcy prediction using computational methods. We further provide a basis for understanding the linguistic construct of evidential strategies.

Reviewing Textual Data in Corporate Bankruptcy Prediction

In the following, we review existing research approaches on the use of textual data in corporate bankruptcy prediction to situate the approach presented in this study within existing research. We limit our review of the literature with respect to the usefulness of qualitative information from corporate financial statements and, for the purposes of this study, exclude a look at external sources of information. First of all, the goal of using textual data can be defined as the improvement of AI-based model capabilities to predict corporate bankruptcies, the prediction probability of which should be increased as a result (Hajek et al. 2014). Since it is often assumed in the literature that better information can be abstracted from the analysis of the text of a financial statement than from the analysis of the balance sheet (Kirkos 2015; Luo and Zhou 2020; Nießner et al. 2021), studies exist that deal with text mining approaches (Shirata et al. 2011) or those that study combinations of quantitative and qualitative informations from financial statemens in order to predict companies future financial situation (Chou et al. 2018; Balasubramanian et al. 2019). These studies examine the extent to which text features are suitable for predicting bankruptcies, but also the financial situation of a company in general (Kloptchenko et al. 2004a). Therefore, approaches can be identified that examine, e.g., the readability (Bushee et al. 2018; Luo and Zhou 2020), the use of hedging terms (Humpherys 2009) and also the sentiment of financial statement textual data (Mayew et al. 2015; Caserio et al. 2020). Moreover, there are text mining approaches that are created exclusively for qualitative data within financial statements as well as studies that analyse companies environment variables, e.g., size of a company and industry affiliation, for predicting corporate bankruptcy (Jones 2017). Within the scope of a study of German-language annual financial statements, it was shown that the analysis of the risk report in terms of linguistic complexity, length and emotional presentation provides suitable information for optimizing the bankruptcy prognosis of companies (Lohmann and Ohliger 2020). This is also shown by earlier results that used collocational networks on English-language financial statements to consider partial analyses of texts in the context of forecasting future corporate financial situations (Magnusson et al. 2005). Thus, it is of scientific interest to investigate text mining approaches in the field of corporate bankruptcy prediction based on annual financial statements in a language-independent

manner using data from other countries. Consequently, this motivates the development of a procedure for quantifying the results of our discourse-analytical approach within this research paper in order to be able to use them accordingly in a bankruptcy prediction model. Subliminally, another finding in looking at research in this area is that there is a trend toward studies that do not make a binary classification into solvent and bankrupt companies, but instead propose ratios that allow to predict companies financial states more specific in a larger variety of classes (Balasubramanian et al. 2019; Lohmann and Ohliger 2020; Smith and Alvarez 2021).

Reviewing Evidential Strategies

Evidential strategies are a part of evidentiality, which itself is a cross-linguistic category that was was first discussed by (Boas 1938), holistically systematized by (Aikhenvald 2004), and is described as a general information source marking that also categorizes the manner through which information is acquired. In doing so, it has been shown to display a progressive hierarchy in languages, where so-called evidentials, i.e., grammaticalized semantic categories, are mandatory. In this hierarchy, self-performed actions by a speaker are ranked the highest, followed by visual perception, auditory perception, inferentials, and ending with reported information at the lowest rank of the spectrum (Oswalt 1986). However, our dataset language, German, does not have evidentiality as a distinct grammatical category. Instead, it belongs to the category of E_2-languages, where evidentiality is encoded through an open set of linguistic devices that develop evidential extensions with explicit of implicit reference to the source of information as a side effect (Fetzer 2014). These so-called evidential strategies always serve two functions in E_2-languages: providing information about the existence of a source and the mode of how information was acquired, however, without being “a function of truth or falsity” (Hardman 1986 p. 121), and a primary designation encoding the speaker’s confidence towards that evidence. For our data, we will classify the evidential strategies using a modified version of the model introduced by Chafe (1986) that is quasi-analogous to the systematization of evidentials in that it proposes a progressive hierarchy based on knowledge reliability. Whilst

Chafe differentiates four modes of knowing, viz. belief, induction, hearsay, and deduction, stating that “each of them is based on a different source, which for belief is problematic, for induction is evidence, for hearsay is language, and for deduction is hypothesis” (Chafe 1986, p. 263), our model excludes the category of belief, arguing that belief is so deeply rooted within epistemic modality and detached from fact, that it cannot possibly constitute an evidential category. Additionally, we are opting for a supercategory that entails both hearsay and quotation viz. ‘reported’, since grammatical distinctions between quotative and hearsay evidence are often not present in E_2-languages and distinctions between ambiguous hearsay and direct as well as indirect quotations are bound to the same evidential strategies in E_2-languages. The same applies to inference and assumption: On the one hand, there is a clear distinction between both categories in the sense that inference relies on first-hand acquired evidence that directly motivates a deduction, whereas assumptions do not have a direct connection between what is being witnessed and what it being implied, thus relying on second-hand evidence. On the other hand, both are often used with the same evidential strategy and are thus often only differentiable through qualitative analysis. However, since contextual constraints may hinder these determinations in many cases, we follow Chafe’s suggestion to subsume both inference and assumption in the supercategory ‘deduction’. Finally, we will use the supercategory ‘sensory’ to describe any form of evidence directly acquired by a speaker, since E_2-languages do not draw a binary distinction between a visual and a sensory category. This creates the following model denoting the trinomial segmentation of evidentiality in E_2-languages:

Evidence	Sensory	Deduction	Reported
including	Visual	Inference	Hearsay
	Auditory	Assumption	Quotative
	Tactile		
	Olfactory		
	Gustatory		
Acquisition	First-hand /direct	based on direct (+indirect)	Second-hand /indirect

Table 5. Comprehensive evidence acquisition model

Research Methodology and Data Set

To tackle our research question, this study will employ a discourse-pragmatic frame of reference using an integrated framework that combines quantitative and qualitative methodologies of linguistic corpus analysis within critical discourse analysis (Wodak 2013), as has been firmly established practice in many linguistic studies of recent years (Nartey and Mwinlaaru 2019). In our approach, the corpora will be scanned for German that-clause constructions (*dass-Konstruktionen*), which serve as a starting point, being the most common evidential strategy used to express evidentiality in E_2-languages. Using quantitative measures, we will determine the most frequent verbs carrying evidential extension that introduce these constructions using the Part-of-speech-tagset by Schmid and Laws (2008) and collocational analysis, scanning for verb collocates before the node *dass*. The verb collocates are then qualitatively analyzed for evidential extension, using the model introduced in 2.2. The V+that-cl. constructions that carry evidential meaning will then be sorted by frequency to determine the most genre-defining evidential strategies used in financial statements. Using these most frequent evidential constructions as a quantifiable foundation, we are scanning the collocations behind the V+that-cl. construction node for adjectives carrying positive/negative semantic meaning in accordance with the systematization by Partington (2015) through quantitative measures and analyze the emerging patterned co-occurences qualitatively for evaluative meanings that may emerge in the textual environment, drawing from Partington (2004) and Sinclair (2004). This methodology enables us to filter out quantifiable cues in the use and distribution of evidential strategies for the prediction of corporate bankruptcy.

The underlying data set of this study consists of 924,903 financial statements taken from the German publication portal *Bundesanzeiger* that were published between the years 2017 and 2021. In total, the data set comprises of 355,704 different German companies, whereby no industry-specific restriction of the companies considered was made. With regard to the bankruptcy prediction, the dataset of financial statements was supplemented by a classification of the

companies based on the *Amadeus* database of *Bureau van Dijk* (Bureau van Dijk 2021). For this purpose, a merge of the data was performed based on the comparison of company names given in the financial statement and the address data of the company using a dedicated Python script. Within the *Amadeus* database, company-specific information is available, which in the following allows the subdivision of the financial statements into the classification scheme of solvent, financially distressed and bankrupt companies. The classification shows the latest financial status for the considered and analyzed companies that is published in *Amadeus*. The distribution of the classes of financial statements identified by the merge results in the consideration of 912,640 financial statements of solvent companies, compared to 8,953 of financially distressed companies and 2,410 bankrupt companies. Therefore, our data set consists of individual financial statements with a minimum text length of 600 characters, a maximum of 1,218,497 characters and an average of 11,277 characters per entry. We therefore consider and analyze textual attachments in terms of their wording. As a result, below we analyze three distinct datasets of financial statements classified according to the financial situation of the publishing companies.

Data Analysis and Results

Using the methodology laid out above, the datasets were analyzed using the text analysis software *SketchEngine*, with which they were initially scanned for verbs carrying evidential extension that introduce German that-clause constructions (*dass-Konstruktionen*). Across all datasets, the verbs carrying evidential extension that occurred most frequently in those positions were *ausgehen von* (assume) and *erwarten* (expect), both primarily expressing epistemic predictions with a second meaning of either assumptive evidentiality, in the case of *ausgehen von*, or, in the case of *erwarten*, assumptive and/or inferential evidentiality, depending on the acquisition of information. This leads to both verbs being positioned in the middle spectrum of the progressive hierarchy of our model, though *erwarten* leans towards a higher hierarchy, whilst *ausgehen von* leans towards a lower one. Scanning for adjectives carrying positive evaluative meaning following the *ausgehen von/erwarten* + *dass* (assume/expect + that) node shows that the most frequent evaluative adjectives are the desirable *gut*

(good) and *positiv* (positive) and the undesirable *negativ* (negative). While this holds true for all three datasets of this study, their frequencies and contextual use varies distinctively.

Corpus of Solvent Companies

Looking at the data collected from solvent companies, *ausgehen von* occurs with a normalized frequency of 75.86 and *erwarten* with 34.22 per million tokens respectively. If not marked differently, all following distributional statistics will refer to the specified normalized frequency. Scanning for adjectives carrying evaluative meaning that occur behind the node specified above, the solvent companies corpus has *positiv* co-occurring with a slightly higher frequency with *erwarten* (1.70) than with *ausgehen von* (1.26), whereas *gut* appears distinctly more frequently with *ausgehen von* (0.94) than with *erwarten* (0.25). However, it needs to be noted that this discrepancy in frequency can largely be attributed to the fixed phrase illustrated in (1) that appears abundantly in this dataset. The same holds true for another positively evaluated adjective, *zufriedenstellend* (satisfactorily), that occurs exclusively in this dataset with a normalized frequency of 0.22. In all but eight instances, it is part of the fixed expression seen in (2), which, together with it being limited to the solvent dataset, leads us to exclude it latter from further analysis.

(1) Hierbei wird **davon ausgegangen, dass** die Marktteilnehmer in ihrem **besten wirtschaftlichen Interesse** handeln. [Herein, it will be **assumed that** all market particpants act in their **best economic self-interest.**]
(2) Insgesamt **erwarten** wir, **dass** sich unsere Geschäfte **zufriedenstellend** entwickeln werden. [Overall, we **expect that** our business dealings will develop **satisfactorily**.]

Much less frequent is the occurrence of *negativ* that co-occurs with *ausgehen von* with the rather low normalized frequency of 0.25 and with an even lower 0.15 with *erwarten*. When co-occuring, *negativ* mostly does so with negation particles such as *kein* (no) in (3), *ohne* (without) in (4) or *nicht* (not), of which the latter is used by the speaker to either raise a negated deduction (*von ausgehen, dass nicht*) as in (5) or negate the deduction itself (*nicht erwartet, dass)* as in (6). Especially noteworthy is the frequent use of intensifiers such as *nennenswert*

(noteworthy), *signifikant* (significant), *wesentlich* (crucial) that, due to being negated, serve a diminutive function here and weaken the use of negative further. Even if no negation particle is used to express a double negative, whatever *negativ* aspect is brought up in the that-clause construction is still considerably weakened: in (7) the speaker assumes that negative and positive effects will single each other out stating the company's *globale Aufstellung* (global positioning) in the causal adverbial phrase introduced by *durch* (due to) as basis for this deduction, whereas the speaker in (8) expects to offset negative effects by means of *innovativen Produkte* (innovative products) that are made explicit via an instrumental adverbial clause introduced by *mit* (with).The speaker in (9) assumes a reduction of *negative EBIT* and bases their deduction on information uttered in the previous utterance through the anaphoric reference *daher* (hence). This high number of explicit evidence mostly introduced through various adverbial phases make the evidential deductions transparent for the audience and raise the credibility of the speaker's deductions.

(3) Wir **gehen** daher **davon aus, dass** das Schadensereignis **keine negativen Auswirkungen** auf die weitere Geschäftstätigkeit von Obermeyer Planen + Beraten haben wird. [We thus **assume that** the damaging event will have **no negative impact** on the further contractual capability of Obermeyer Planen + Beraten.]

(4) Das Unternehmen **geht** jedoch **davon aus, dass** die Rechtsstreitigkeiten **ohne nennenswerte negative Auswirkungen** auf die Finanz- oder Ertragslage des Unternehmens beigelegt werden können. [The company **assumes**, however, **that** the legal disputes will be able to be put aside **without any noteworthy negative impact** on the financial or profit performance of the company.]

(5) Wir **gehen** dabei **davon aus, dass** sich **nicht erneut signifikante negative Sondereinflüsse**, auch nicht aus den bestehenden Pensionsverpflichtungen, ergeben werden. [Herein, we **assume that new significant negative special influences** will **not** ensue, not even from the existing pension obligations.]

(6) Am 31. Dezember 2015 **wird nicht erwartet, dass** diese Angelegenheit **wesentliche negative Auswirkungen** auf die Betriebsergebnisse, Liquidität oder finanzielle Situation haben wird. [As of December 31, 2015, **it is not expected that** this matter will have a **crucial negative effect** on results of operations, liquidity or financial condition.

(7) Durch unsere globale Aufstellung **gehen** wir **davon aus, dass** sich **positive und negative Effekte** weitestgehend ausgleichen und damit beherrschbar sind. [Due to our global positioning we **assume that** the **positive and negative effects** will largely offset each other and thus remain manageable.]

(8) Wir **erwarten, dass** wir diese **negativen Einflüsse** mit neuen innovativen Produkten bzw. Zustellungssystemen ausgleichen können. [We **expect that** we can offset these **negative influences** with new innovative products or delivery systems.]

(9) Insbesondere werden die Investitionen die Basis für die Einrichtung der C-SMC-Produktion durch MCHC legen. Daher **gehen** wir **davon aus, dass** wir 2017 das **negative EBIT** deutlich reduzieren können. [In particular, the investment will lay

the foundation for the establishment of C-SMC production by MCHC. Therefore we **assume that** we can reduce the **negative EBIT** drastically in 2017.]

As the contextual analysis of the occurrences of *negativ* in the dataset shows, negatively evaluated assumptions are hardly to be found. On the contrary: evidential statements by the speakers are predominantly positively evaluated. Looking at the occurrences of *positiv* and *gut*, this is reinforced by the large amount of verbs semantically encoding continuity and progression, especially *fortsetzen* (continue) in (10), (11) and (16), but also *ausbauen* (expand) in (13) and *übertreffen* (surpass) in (15). All these verbs occur in a verb phrase with modal auxiliaries belonging to two semantic domains: the modal auxiliary *kann* (can / be able to) highlights the (expected) dynamic ability of progress in (10), (11), (13) and (15), whereas *werden* (will) expresses a strong epistemic prediction and highlights either the speakers' belief in the evidential assessment of the expected progress of the company as in (12), (14), (16), or in the ability of progress as in (11), (13) and (15).

(10) Es **ist zu erwarten, dass** das Unternehmen die **positive Entwicklung** der vergangenen Jahre durch die hoch qualifizierten Mitarbeiter, die sehr gute Ausstattung des Maschinen- und Anlagenparks sowie dem breiten Know-how des Unternehmens weiter **fortsetzen kann**. [It **is to be expected that** the company **is able to continue the positive progress** of the past years due to the highly-qualified employees, the very good equipment of the machine and facility park, as well as the broad know-how of the company.]

(11) Wir **erwarten, dass** die insgesamt **positive Entwicklung** der vergangenen Jahre auch in den kommenden Jahren **fortgesetzt werden kann**. [We **expect that** the overall **positive development** of the past years **will be able to be continued** the coming years.]

(12) Wir **erwarten, dass** sich diese Maßnahme 2016 **positiv** auf das Kapitalanlageergebnis auswirken **wird**. [We **expect that** this measure **will** have a **positive effect** on the capital investment outcome in 2016.]

(13) Insbesondere **ist zu erwarten, dass** die **gute Marktposition** des Unternehmens in Lateinamerika und Afrika weiter **ausgebaut werden kann**. [Especially, **it is to be expected, that** the **good market position** of the company in Latin America and Africa **can be expanded** further.]

(14) Wir **erwarten, dass** sich der **gute Ausbildungsstand** und die hohe Leistungsbereitschaft unserer Mitarbeiter auch künftig **positiv** auf das Verhältnis unserer Mitglieder und Kunden zu ihrer Bank auswirken **werden**. [We **expect that** the **high level of training and commitment** of our employees **will** also have a **positive impact** on the relationship between our members and customers and their bank in the future.]

(15) Zusammenfassend **ist** daher **davon auszugehen, dass** die **gute Auftragsentwicklung** des Berichtsjahres im Jahr 2016 noch einmal **übertroffen werden kann**. [Summarizing this, it **is** thus **assumed that** the **good** ETO of the 2016 report **will be able to be surpassed** once again.]

(16) Es **ist davon auszugehen, dass** sich dieser **positive Trend** auch in 2016 **fortsetzen wird**. [It **is assumed that** this **positive trend will** also **continue** in 2016.]

Corpus of Bankrupt Companies

Scanning for verbs with encoded evidential meaning introducing that-clause constructions of bankrupt companies produces the highest frequency with the same two verbs previously introduced. While *ausgehen von* occurs with a slightly raised normalized frequency of 77.93 here, *erwarten* occurs almost twice as often as in the solvent dataset, displaying a normalized frequency of 76.27 that is on par with *ausgehen von*. Looking at adjectives carrying evaluative meaning co-occuring with the *erwarten/ausgehen von* + *dass*-constructions produces very different results as well: *ausgehen von* has the adjectival collocates *positiv* and *negativ* with normalized frequencies of 0.63 and 1.25, but displays a very low frequency of 0.21 for *gut*, whereas *erwarten* has *gut* co-occurring with a comparatively high frequency of 1.04. However, looking at the actual data, it is revealed that these co-occurences all stem from the exact same sentence, illustrated in (22), that occurs in various financial statements of the same company throughout the years, tainting the significance of the heightened frequency here. The exact same issue can be observed for the co-occurrence of *positiv* with *erwarten* + *dass* in this dataset, where the repeating sentence is (23). The frequency of *negativ* following the *erwarten* + *dass* constructions displays no such issues and is 0.83 per million words. Whereas the vast majority of *negativ*'s patterned co-occurrences turned out to be double negatives in the solvent dataset, the bankrupt dataset is different: though a small number is still comprised of negative deductions weakened by negated intensifiers (17) or negating the deductions itself (18) and thus render the deduced outcome desirable, (20) and (21) are explicitly negatively evaluated and others carry temporality markers such as *derzeit* (currently), implying the speaker's prediction may change if contradictory evidence occurs in the future. Other markers that weaken the deduction can be found in (19) and (20): in (19) the premodification of *ausgeglichen* (offset) with the approximator *annähernd* (approximately) implies

uncertainty of the speaker's prediction as does the modal adverbial clause in (20): although the speaker expresses a high degree of certainty (*hohe Wahrscheinlichkeit*) that their deduction is true, it is still a comparatively weaker claim than the deduction illustrated in (21) that is devoid of it. What stands out in all these patterned co-occurences is the absence of explicit evidence realized by adverbial phrases that was prominently featured co-occuring with the use of evidential strategies in the dataset of solvent companies.

(17) Die Geschäftsführung schätzt die Risiken als überschaubar ein und **geht derzeit davon aus, dass** sie keinen **nennenswerten negativen Einfluss** auf die Entwicklung der Gesellschaft haben werden. [The management evaluates the risk as manageable and **currently assumes that** they will **have no noteworthy negative impact** on the development of the company.]

(18) Es **ist nicht zu erwarten, dass** dieser **negative Sondereffekt** sich in den Folgejahren wiederholen wird. [It **is not to be expected** that this **negative special effect** will repeat in the years to follow.]

(19) Die Gesellschaft **geht davon aus, dass** diese **negativen Einflüsse** auf Umsatz und Ergebnis im 2. Halbjahr **annähernd ausgeglichen** werden können. [The company **assumes that** the **negative influences** on sales and earnings can be **approximately offset** again.]

(20) Mit **hoher Wahrscheinlichkeit ist zu erwarten, dass** die **negativen Folgen** für die Bank umso stärker sind, je länger die Pandemie anhält. [It **is to be expected with a high degree of certainty that** the **negative implications** for the bank will be the greater the longer the pandemic lasts.]

(21) Es **ist zu erwarten, dass** die **negativen Folgen** für die Wirtschaftsleistung unserer Bank umso stärker sind, je länger die Pandemie anhält. [It **is to be expected that** the **negative implications** for the economic performance of our bank will be the greater the longer the pandemic lasts.]

Whereas the solvent dataset showed an abundance of verbs denoting progress occurring within the that-clause constructions with *positiv* and *gut*, in this dataset positive growth or progress of the companies, if occurring at all, is dependent on external factors that are made explicit via causal adverbial phrases often led in by *aufgrund* (due to) and serve as evidential justifications for the speakers' deductions. The *Umsatzwachstum* (turnover growth) in (22), for example, can only occur due to a good industry position (*Branchenlage*) and further expansion (*zusätzliche Expansion*), and the continuation of the company in (24) is inferred to be predominantly likely (*überwiegend wahrscheinlich*) only due to the positive continuation prognosis (*positive Fortführungsprognose*). While there are also some patterned co-occurences, in which deductions with a positive outlook are

devoid of external factors, and reference to evidence for that matter as in (22), most refer to some form of external factors on which a positive, or in the case of (25) negative, development of the company depends, rendering this an inherent feature of the dataset.

(22) Die Geschäftsleitung **erwartet, dass** sie aufgrund der **guten Branchenlage** im E-Commerce Umfeld sowie durch **zusätzliche Expansion** in andere Produktsegmente, im Jahr 2019 ein **Umsatzwachstum** von 8-10 % im Vergleich zum Vorjahr erreichen kann. [The management **expects that** in 2019 it can achieve a **turnover growth** of 8-10 % in comparison to the previous year due to the **good industry position** in the e-commerce environment and **further expansion** in other product segments.]

(23) Wir **erwarten, dass** sich diese **positive Tendenz** auch in den folgenden Jahren als stabilisierender Faktor und signifikanter Wettbewerbsvorteil für das Unternehmen erweist. [We **expect that** this **positive tendency** will prove to be a stabilizing factor and significant competitive advantage for the company in the years to come.]

(24) Bilanzierung und Bewertung erfolgten trotz bilanzieller Überschuldung der Gesellschaft zu Fortführungswerten, weil die Geschäftsführung **davon ausgeht, dass** aufgrund der vorliegenden **positiven Fortführungsprognose** und der damit von der finanzierenden Bank genehmigten neuen Finanzierungsstruktur die Fortführung der Unternehmenstätigkeit **überwiegend wahrscheinlich** ist. [Despite the company's sheet over-indebtness, balancing and appraisal were done at going concern values because the management **assumes that** the continuation of the company's operation is **predominantly likely** due to the **positive going concern forecast** and the concomitant approval of the new financing structure by the financing bank.]

(25) Nach allen Informationen die uns vorliegen, müssen wir auch für das Jahr 2017 **davon ausgehen, dass** unsere Branche von den **positiven Rahmenbedingungen** speziell in Deutschland **nicht** profitieren konnte und erhebliche Einbrüche im Umsatz mit Gardinen und Dekostoffen (speziell im Mittelpreissegment) zu verzeichnen waren. [Based on all the information available to us, we **must** also **assume** for 2017 **that** our industry could **not** benefit from the **positive underlying conditions**, especially in Germany, and that there were significant downturns in sales of curtains and decorative fabrics (especially in the mid-price segment).

Corpus of Financially Distressed Companies

The financially distressed dataset bridges the gap between bankrupt companies and solvent ones. As has been shown to be a universal in the other two datasets, the verbs with an evidential extension introducing a that-clause construction are once again *ausgehen von* with a normalized frequency of 47.31 and *erwarten*, with 48.52. On the one hand, this continues the trend of both verbs occurring with a roughly equal frequency, but on the other hand is almost exclusively lower than the frequency that was displayed in the other datasets. What is unique to this

dataset, however, is that although the total dataset size is twice as large as the bankrupt one, all adjectives carrying evaluative meaning that occur as collocates to *ausgehen von* + *dass*-constructions occur with the same, albeit low frequency of 0.19. *Erwarten*, however, displays evaluative adjectives following a that-clause construction with distinctly higer frequencies, with *positiv* displaying a normalized frequency of 0.65, *negativ* displaying 0.74, and *gut* being entirely absent. Looking at the data in context reveals a similar marking of certainty that was also observable in the bankrupt dataset. In (26) this goes as far as having parts of the deductions of an undesirable development seen (20) and (21) reoccurring verbatim. What stands out as well, is the more complex, hypotactical syntax that emerges when *negativ* follows an evidential that-clause construction: here, the speakers either state multiple evidence on which their deductions are based as in (28) or justify the omission of evidence, even making explicit doing so out of self-interest as in (27), where the speaker appeals to their audience that it would be only logical (*vernünftigerweise*) to not inform the general public due to the negative repercussions (*negative Folgen*) this may have on the company.

(26) **Mit hoher Wahrscheinlichkeit lässt sich** jedoch bereits jetzt **erwarten, dass** die **negative Folgen** für die Bank umso stärker sind, je länger die Pandemie anhält. [However, it **is** already **to be expected** with a **high degree of certainty that** the **negative impact** will be the greater the longer the pandemic lasts.]

(27) Wir beschreiben diese Sachverhalte in unserem Bestätigungsvermerk, es sei denn, Gesetze oder andere Rechtsvorschriften schließen die öffentliche Angabe des Sachverhalts aus oder wir bestimmen in äußerst seltenen Fällen, dass ein Sachverhalt nicht in unserem Bestätigungsvermerk mitgeteilt werden sollte, weil **vernünftigerweise erwartet wird**, dass die **negativen Folgen** einer solchen Mitteilung deren Vorteile für das öffentliche Interesse übersteigen würden. [We describe these matters in our auditor's report unless law or regulation precludes public disclosure of the matter or, in extremely rare circumstances, we determine that a matter should not be communicated in our auditor's report because the **negative repercussions** of such communication would **logically be expected to** outweigh its public interest benefits.]

(28) Die Geschäftsleitung der flatex Bank AG beobachtet die politischen Entwicklungen kritisch, **erwartet jedoch, dass eventuell negative Auswirkungen** durch den weiteren Ausbau der Aktivitäten mit den bestehenden Partnern sowie neuen Geschäftspartnern im Mandantengeschäft als auch durch neue Handelsprodukte **abgemildert werden können**. [The management of flatex Bank AG is keeping a critical eye on political developments, but **expects that** any **negative effects can be mitigated** by further expanding activities with existing partners and new business partners in the client business, as well as through new trading products.]

This increased need for justification occurring within negatively evaluated evidential strategies is in contrast to the contexts, in which the collocate *positiv* occurs in this dataset: on the one hand, the assessments of the speakers are overwhelmingly positive, as in (29), where a product (*Produkt*) is praised by the speaker to have exceeded expectations and on the basis of this, the deduction that the product will continue to do so, is made. On the other hand, the deductions with a desirable outcome that are missing information source are comparable to the ones observed in the solvent dataset, as in (30).

(29) Das Produkt Logistics übertrifft in den ersten acht Monaten die budgetierten Erwartungen und wir **erwarten, dass** dieser **positive Trend** bis Ende 2017 **anhält.** [The Logistics product exceeded budgeted expectations in the first eight months and we **expect that** this **positive trend to continue** until the end of 2017.]

(30) Es wird **erwartet, dass** sich die **positive gesamtwirtschaftliche Entwicklung** insgesamt **fortsetzt**. [It is **expected that** the **positive macroeconomic development** will **continue** overall.]

Discussion and Implications

This study examines the usefulness of evidential strategies to predict corporate financial distress and bankruptcy based on qualitative data from German companies financial statements. It shows that evidential strategies can be used to distinguish between the classes of companies based on the analysis of their textual financial statement data. In terms of frequency, we found that *ausgehen von* and *erwarten* were the most dominant verbs used in evidential that-clause constructions in German financial statements and thus served as our tertium comparationis.

The strikingly similar distribution of the two verbs under investigation in the financially distressed and bankruptcy dataset suggests that evidential strategies are infrequently used by financially distressed companies and possibly deliberately so. One possible reason for this could be that their financial statements are aimed at diverting from problematic financial situations and especially their causes. That may also be why the linguistic analysis of the few evidential strategies found with this class shows that deductions with positively evualted adjectives are entirely devoid of evidence, whereas the ones with negatively evaluated adjectives not only display little evidence as well, but are also comprised of the most complex hypotactical sentence structures of all three

classes and, accordingly, have the longest word count of the identified sentences in these negatively evaluated that-clause constructions. This differs drastically from the other two classes, the evidence for the various deductions is frequently made explicit through various forms of adverbial phrases and the only difference is the displayed level of certainty of the speakers assesments, which is higher with the solvent class due to a number of strong modals displaying the semantic notions of ability and prediction, and lower with the bankrupt class due to a number of uncertainty markers and formulated preconditions introduced through causal adverbials.

Finally, looking at the mere change in frequency of the positively (*gut, positiv*) and negatively (*negativ*) evaluated adjectives illustrated in Figure 9 shows two things: first, apart from *negativ* following after *erwarten,* all evaluative adjectives show a (Chafe 1986) pronounced infrequency of use in the financially distressed

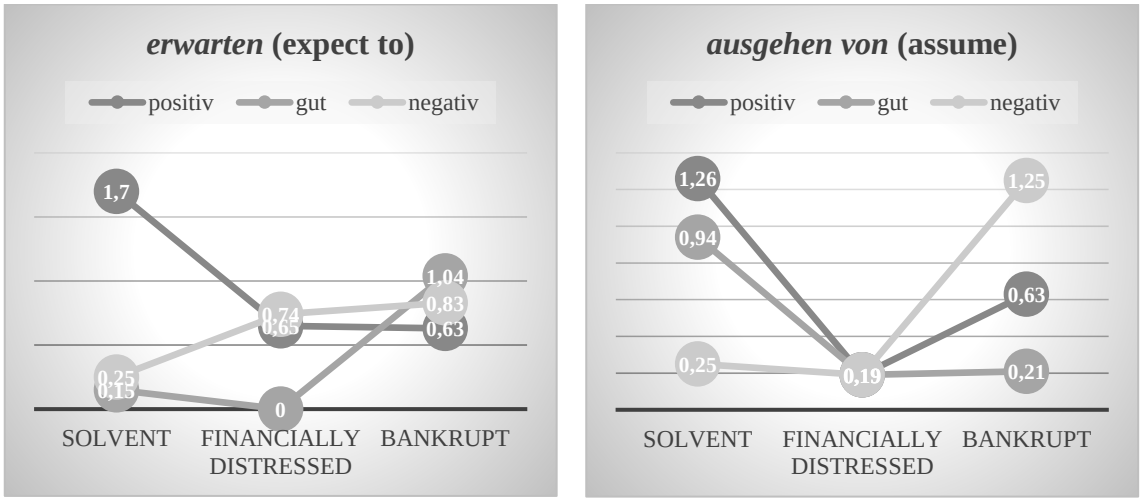


Figure 9. Distribution of evaluative adjectives following ES in FS

class that was also visible with evidential strategies, corroborating our suggestion made above. Second, the use of *positiv* sees a considerable drop in frequency in the bankruptcy class in comparison to the solvent one that seems to be reciprocal to the behaviour of *negativ*, which occurs frequently in the bankrupt class, but displays a much lower frequency in the solvent one. Considering the fact that a closer look showed that, in the case of solvent companies, negatively evaluated adjectives usually occured negated, rendering the reports of solvent companies

by and large desirable even if negatively evaluated adjectives such as *negativ* occur, this distributional shift may serve as the most important cue for a bankruptcy prediction model based on financial statements. In 5.2, we present how a financial statement can be preprocessed according to our results and thus features can be generated for a future improvement of AI-based corporate bankruptcy prediction.

Contributions to Literature

The analysis of the presented data using a corpus linguistic approach undertaken for this study complements the field of corporate bankruptcy prediction with another option for evaluating textual data from financial statements. We were able to show that the concept of evidentiality in financial statements contributes to the distinction between solvent, financially distressed and bankrupt companies. In terms of current approaches, we thus provide a method that can be used to improve future AI-based predictive models for corporate bankruptcy. Based on the idea of the use of collocational networks (Magnusson et al. 2005) and the partial consideration of single components of a text in financial statements (Wei et al. 2019; Lohmann and Ohliger 2020), we have developed a new approach for text mining. The data features that can be developed from that allow, on the one hand, the identification of argumentation structures within a financial statement and, on the other hand, the evaluation of the confidence with which they were made. By successfully adapting the concept of evidential strategies for financial statement analysis and presenting a procedure to quantify our findings, we continue to open the way for future researchers to explore how the results from our corpus linguistic analysis can be used within various other studies.

Feature Engineering Process and Practical Implications

In addition to the theoretical contribution, this study also provides practical implications: We present an approach resulting from this research that allows us to quantify the usage of evidential strategies and preprocess text from financial statements for using those in corporate bankruptcy prediction. Our approach can be divided into four phases, which is illustrated in Figure 10 below.

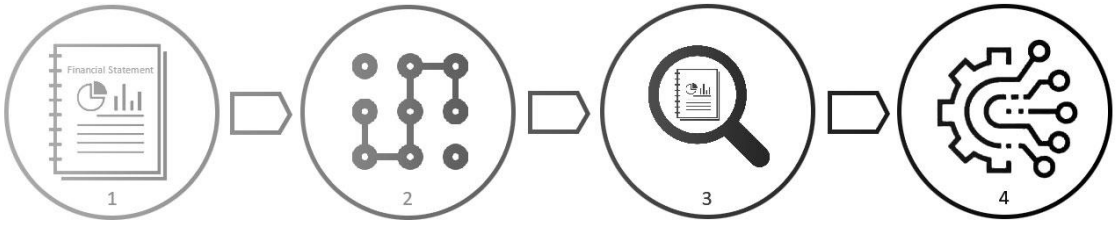


Figure 10. Discourse-analytical text mining process for feature engineering

In the first phase, the pure textual components of a financial statement must be extracted and separated from the other components, depending on the file format. In the second phase, a pattern is defined which makes the lemmas of verbs used as evidential strategies, such as "assume" and "expect", that follow a that-clause construction recognizable. In the third phase, based on these predefined patterns, the corresponding argumentation structures within the text are identified. Finally, these identified structures are evaluated in the fourth phase as follows: On the one hand, the relative frequency of the use of such constructions is calculated and on the other hand, adjectives that are following a that-clause construction will be evaluated by means of a sentiment analysis approach. In this regard, programming with Python enables one way for the recognition and differentiation of adjectives in sentences through the use of part-of-speech tagging (spaCy v3.2 2021) as well as the sentiment-based evaluation, e.g., with SentiWS v2.0 (Remus et al. 2010). However, it should be noted that so far no sentiment dictionary based on German financial statements exists that was tailored to this use case. Consequently, this procedure allows us to extract features for the model development of corporate bankruptcy predictions, which in return allows us to extract and use not only the frequency, but also the structure of statements made within the financial statements.

Furthermore, the possibility of improving predictive models for corporate bankruptcies also enables a risk minimization in regard to debt accumulation and enables stakeholders to classify the financial situations of companies better. Moreover, as has been illustrated, the approach demonstrated in this research paper also allows for a non-binary classification of companies' financial situations.

Limitations and Future Research Opportunities

As with any research, our study has some limitations that need to be considered when interpreting the results and providing future research opportunities. First, we use a dataset that is comparatively large to analyze the argumentation structures, but our results cannot be transferred arbitrarily to financial statements in other languages and even for the German language is due to be further investigated. Because we are in the early stages of implementing corpus linguistic approaches for bankruptcy prediction models, assessing the influence of evidential strategies on bankruptcy prediction on the basis of future financial statements is no straightforward task, especially since we only examined financial statements that were published within the limited time period of 2017 to 2021. Second, our dataset and particularly its status assessment into solvent, financially distressed and bankrupt companies, is, as other studies in the field have shown (Roumani et al. 2020), subject to class imbalance. Considering the use of factors based on evidential strategies in the development of models to predict corporate bankruptcy, this imbalance issue and the question of how to balance the classes represents another opportunity for future research. Finally, the status labeling of the financial statements is subject to the assumption that the classification using the Amadeus database from Bureau van Dijk, which obtains its information from Creditreform, is appropriate. For future research on financial statements of different countries and different languages, it would therefore be conceivable to apply the results to a financial factors based classification of the companies in order to investigate the suitability of evidential strategies in other environments.

Conclusion

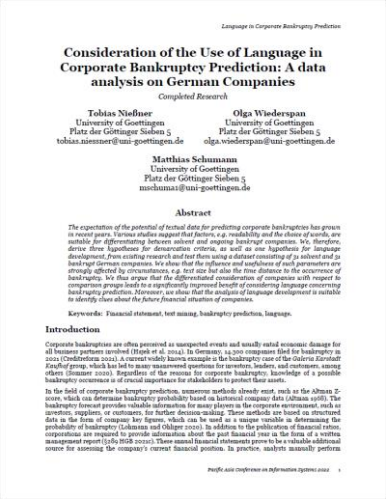
This study has been motivated by the surge of interest in the analysis of qualitative textual data from financial statements for the prediction of corporate bankruptcy. Although initial approaches have been conducted in the area of text analysis to optimize bankruptcy prediction, we have identified a research gap in the study of presented decisions in financial statements, which we have explored with a corpus linguistic methodology in regard to the use of evidential strategies. In this study, we took insight from linguistic theory and recently published study

on the use of textual data within corporate bankruptcy prediction based on financial statements to answer the predefined research question. Considering the use of statistical models to present a solution to this question that would improve corporate bankruptcy predictions, we have developed a new procedure to generate features based on our findings that can enhance the prediction accuracy. We therefore have shown that feature engineering based on the concept of evidential strategies is suitable to distinguish between the classes of solvent, financially distressed and bankrupt companies. Finally, we have to mention that this study presents results based on German financial statements, but also suggests an adaption to incorporate textual data into the development of statistical models for the analysis of financial statements from other countries. The next step would be to further explore to what extent quantitative financial statement analysis can be complemented with insights from qualitative analysis through the use of AI to predict corporate bankruptcies. Beyond the consideration of evidential strategies in the context of corporate bankruptcy prediction, the utility in fraud detection would also be an interesting research topic.

III Language in Corporate Bankruptcy Prediction

"Consideration of the Use of Language in Corporate Bankruptcy Prediction: A data analysis on German Companies"

July 5-9,2022 – PACIS, Taipei/Sydney Virtual Conference, Asia



Language in Corporate Bankruptcy Prediction

Consideration of the Use of Language in Corporate Bankruptcy Prediction: A data analysis on German Companies

Completed Research

Tobias Nießner
University of Goettingen
Platz der Göttinger Sieben 5
tobias.niessner@uni-goettingen.de

Olga Wiederspan
University of Goettingen
Platz der Göttinger Sieben 5
olga.wiederspan@uni-goettingen.de

Matthias Schumann
University of Goettingen
Platz der Göttinger Sieben 5
mschuma1@uni-goettingen.de

Abstract

The expectation of the potential of textual data for predicting corporate bankruptcies has grown in recent years. Various studies suggest that factors, e.g. readability and the choice of words, are suitable for differentiating between solvent and ongoing bankrupt companies. We, therefore, derive three hypotheses for demarcation criteria, as well as one hypothesis for language development, from existing research and test them using a dataset consisting of 31 solvent and 31 bankrupt German companies. We show that the influence and usefulness of such parameters are strongly affected by circumstances, e.g. text size but also the time distance to the occurrence of bankruptcy. We thus argue that the differentiated consideration of companies with respect to comparison groups leads to a significantly improved benefit of considering language concerning bankruptcy prediction. Moreover, we show that the analysis of language development is suitable to identify clues about the future financial situation of companies.

Keywords: Financial statement, text mining, bankruptcy prediction, language.

Introduction

Corporate bankruptcies are often perceived as unexpected events and usually entail economic damage for all business partners involved (Tirpak et al. 2014). In Germany, 14,300 companies filed for bankruptcy in 2021 (Creditreform 2021). A current widely known example is the bankruptcy case of the *Galeria Karstadt Kaufhof* group, which has led to many unanswered questions for investors, lenders, and customers, among others (Sommer 2020). Regardless of the reasons for corporate bankruptcy, knowledge of a possible bankruptcy occurrence is of crucial importance for stakeholders to protect their assets.

In the field of corporate bankruptcy prediction, numerous methods already exist, such as the Altman Z-score, which can determine bankruptcy probability based on historical company data (Altman 1968). The bankruptcy forecast provides valuable information for many players in the corporate environment, such as investors, suppliers, or customers, for further decision-making. These methods are based on structured data in the form of company key figures, which can be used as a unique variable in determining the probability of bankruptcy (Lukmani and Ohliger 2020). In addition to the publication of financial ratios, corporations are required to provide information about the past financial year in the form of a written management report (§289 HGB 2020). These annual financial statements prove to be a valuable additional source for assessing the company's current financial position. In practice, analysts manually perform

Authors: Tobias Nießner, Olga Wiederspan and Matthias Schumann

Outlet: *Pacific Asia Conference on Information Systems (PACIS). Taipei-Sydney, Asia, 2022.*

Abstract: The influence of textual data on corporate bankruptcy prediction has become increasingly influential in recent years. Various studies suggest that factors such as readability and the choice of words are suitable for differentiating between solvent and bankrupt companies. We therefore derive three hypotheses for demarcation criteria, as well as one hypothesis for language development, from existing research and test them using a dataset consisting of 31 solvent and 31 bankrupt German companies. We show that the influence and usefulness of such parameters is strongly affected by circumstances such as text size, but also the time distance to the occurrence of bankruptcy. We thus argue that the differentiated consideration of companies with respect to comparison groups leads to a significantly improved benefit of considering language with

respect to bankruptcy prediction. Moreover, we show that the analysis of language development is suitable to identify clues about the future financial situation of companies.

Keywords: Financial statement, Text Mining, Corporate Bankruptcy Prediction, Language.

Introduction

Corporate bankruptcies are often perceived as unexpected events and usually entail economic damage for all business partners involved (Hajek et al. 2014). In Germany, 14.300 companies filed for bankruptcy in 2021 (Creditreform 2021). A current widely known example is the bankruptcy case of the *Galeria Karstadt Kaufhof* group, which has led to many unanswered questions for investors, lenders, and customers, among others (Sommer 2020). Regardless of the reasons for corporate bankruptcy, knowledge of a possible bankruptcy occurrence is of crucial importance for stakeholders to protect their assets.

In the field of corporate bankruptcy prediction, numerous methods already exist, such as the Altman Z-score, which can determine a bankruptcy probability on the basis of historical company data (Altman 1968). The bankruptcy forecast provides valuable information for many players in the corporate environment, such as investors, suppliers, or customers, for further decision-making. In Germany, SCHUFA provides a score on the probability of corporate bankruptcy (Ahmadi et al. 2018). These methods are based on structured data in the form of company key figures, which can be used as a unique variable in determining the probability of bankruptcy (Lohmann and Ohliger 2020). In addition to the publication of financial ratios, corporations are required to provide information about the past financial year in the form of a written management report (HGB 2021b). These annual financial statements prove to be a valuable additional source for assessing the company's current financial position. In practice, analysts manually perform qualitative annual financial statement report analyses to supplement the quantitative information provided in the balance sheet. The consequence of the mandatory publication of management reports leads to an increasing amount of unstructured data, which offers new potential for qualitative financial statement analysis. The high number of annual financial statement reports offers application possibilities for text mining methods that cannot be realized by manual analysis (Kloptchenko 2004). Text mining extracts the required information as part of a software-based process by modifying previously unstructured data into a structured form. In research, studies have already been able to prove that linguistic features and patterns in annual reports show the possibility to describe future developments of the company (Ravisankar et al.

2011). Moreover, some research has been able to support the management concealment hypothesis, according to which management should have greater incentives to hide bad news about company performance through specific language devices in annual reports (Bloomfield 2002; Li 2008; Lohmann and Ohliger 2020; Le Maux and Smaili 2021).

Text mining is a type of information retrieval that can be effectively used to discover such linguistic patterns in annual reports and analyze them with regard to obfuscation, among other things. In this way, initial indications can be extracted in view of a change in the company's financial situation, which are not yet apparent on the basis of key financial figures (Lohmann and Ohliger 2020). This information can be further used to predict the bankruptcy of a company.

In order to be able to use information for bankruptcy prediction, it is necessary to identify the specific linguistic means that are to be analyzed in order to infer future corporate bankruptcy. For this purpose, this paper examines two groups of companies, which consist of solvent and bankrupt companies. Another consideration relates to the period before a company became bankrupt. Here, it can be investigated whether a certain language development emerges in the annual financial statements over the years, which can be related to future bankruptcy. To this end, the second research question is posed:

RQ: How is the occurrence of bankruptcy reflected in the language within consecutive financial statements?

In order to answer these research questions, this paper is structured as follows. We first introduce the research background of text mining within corporate bankruptcy prediction. Subsequently, we describe, in conjunction with the presentation of the data set used for the case study, data acquisition and data preprocessing. Then, motivated by the literature, we derive hypotheses that will be used to answer our research questions. In addition, we describe our procedure for testing these hypotheses. Afterward, we analyze and discuss our results, state contributions for literature and practice and mention limitations. Finally, we discuss limitations and future research opportunities and give a brief conclusion.

Research Background

In the following, we review published literature on the use of text mining approaches in corporate bankruptcy prediction to show the connecting factors of this study. We consider language-independent studies that demonstrate tools that add value to the prediction of corporate bankruptcies or analysis of financial statements. Furthermore, we exclude studies that continue to use external data sources to achieve this goal. First of all, the use of text mining methods can be motivated by the possibility of improving AI-based models for predicting corporate bankruptcies (Hajek et al. 2014). In this context, it is often assumed that valuable information can be extracted from textual data of a financial statement to predict corporate bankruptcies (Goel and Gangolly 2012; Luo and Zhou 2020). For this reason, there are studies that, in contrast to the classic approach of balance sheet analysis, deal with the qualitative analysis of the financial statement and aim to draw conclusions about the financial situation of companies from the extracted information (Kloptchenko 2004; Balasubramanian et al. 2019; Le Maux and Smaili 2021). Furthermore, they also deal with the forecast of future developments (Li 2008). For this purpose, different approaches are known that consider different linguistic factors, e.g., the readability (Le Maux and Smaili 2021), the analysis of sentiments (Loughran and McDonald 2011, 2016), and the use of hedging dictionaries (Humpherys 2009). In addition, Magnusson et al. (2005) addressed the use of collocational networks to study the analysis of word choice in English-language financial statements. By looking at individual keywords and their neighboring words in the text, it was shown that a targeted analysis is suitable for predicting future developments. An analysis of word collocations reveals language patterns and meanings that are not apparent from frequency lists of individual words or from manual analysis of larger amounts of text (Pollach 2012). A German study made use of this idea and looked at risk reports of German companies in the following (Lohmann and Ohliger 2020). It was shown that the length of the risk report, as well as the linguistic complexity and the emotional writing style of the authors, allow conclusions to be drawn in order to identify not only solvent but also financial distress or bankrupt companies. Since the analysis of textual data of annual financial statements is also always dependent on the language to be analyzed, it is of interest to investigate to what extent identified features of a text can also be adapted to

corporate bankruptcy prediction of other countries in a language-independent manner.

Data Set and Preprocessing

Within the case study, a total of 62 German companies are considered, whereby an equal distribution of solvent and bankrupt companies was chosen. Bankrupt companies from the years 2017 to 2021 were selected on a selective basis. For this purpose, the official homepage for German bankruptcy notices was checked for each company to determine whether a bankruptcy application had been filed (InsBekV 2021). These companies are of medium size, as the German Commercial Code (HGB) requires them to prepare a management report, in contrast to smaller corporations (§264 HGB 2021b; §267 HGB 2021a). Furthermore, when selecting bankrupt companies, care was taken to ensure that a minimum of three annual financial statements were available. With the help of the Amadeus database from Bureau van Dijk (2021), comparable solvent companies are selected for comparison based on revenue, size, and industry of the company. Thus, deviations that could be due to these parameters should be avoided. The annual financial statements belonging to the companies were then acquired in HTML format from the Bundesanzeiger homepage, which is the official publication portal for German corporate publications. In the following, the data set finally results from 142 annual financial statements of bankrupt companies and 143 annual financial statements of solvent companies. For the tracing analysis, the text of the financial statements was extracted using Python by using the BeautifulSoup package (Richardson 2021). In addition, a decomposition into tokens was generated from the raw text. For this, all special characters as well as stop words were removed (Diaz and Suriyawongkul 2020) and a stemming procedure was applied (Porter et al. 2020). For the observation of language development, the raw texts were divided into individual corpora according to their distance from the last available year-end.

Hypothesis Development

In what follows, we detail the research findings picked up in the research background and derive hypotheses from them regarding readability, the use of vagueness terms, and the use of passive language to distinguish solvent from bankrupt companies. We then address the hypotheses on the change of linguistic representation. In the first section, we discuss the motivation of the hypotheses related to the first research question, before we derive the hypotheses related to the second research question from the literature in the second section.

Research already shows a high proportion of studies on the readability of a company's annual financial statements. The background of this research focus is based on Bloomfield's Incomplete Revelation Hypothesis (2002). This hypothesis is also called the "management concealment hypothesis" and states that managers have a greater incentive to conceal information about poor corporate performance because markets and investors are less fully responsive to information that can be extracted from public disclosures at a higher cost (Bloomfield 2002; Li 2008). Li (2008) and Le Maux and Smaili (2021) were able to support this hypothesis by demonstrating the correlation between a company's financial performance or bankruptcy and the readability of its annual reports. The research base of both studies was annual reports in English. These studies also used different readability indices for the English language to examine the correlation.

In contrast, Lohmann andOhliger (2020) examined the readability index (LIX) of German annual reports but analyzed the risk report and the rest of the management report separately (Lohmann and Ohliger 2020). The study distinguishes between the experimental group of bankrupt companies and the control group of solvent but financially distressed companies (Lohmann and Ohliger 2020). Based on the holistic view of the aforementioned studies, there is a research gap with regard to the German-language annual financial statements of German companies. The question arises whether the correlation between readability and bankruptcy can also be shown after a holistic analysis of German annual financial statements and whether differences arise when comparing bankrupt and solvent companies. Adding the management concealment hypothesis, the following hypothesis is therefore formulated:

H1.1: The readability of the annual financial statements of bankrupt companies is worse than that of solvent companies.

The studies by Humpherys (2009) and Goel and Gangolly (2012) are based on a similar approach. These studies analyze English-language financial statements and distinguish between fraudulent companies and companies that do not commit fraud. Both studies assume that fraudulent companies, due to the need to conceal or falsify something, more often use strategic hedging measures in information management to avoid responsibility (Humpherys 2009; Goel and Gangolly 2012). This hedging can appear as a linguistic device in the form of vagueness expressions, as these terms reduce certainty and create vagueness (Humpherys 2009). Concrete examples of these expressions are words such as "should", "could", "would" or "some" (Humpherys 2009). Transferred to the context of financial statement analysis, it may appear in the following sentence, for example:

"The company might make some profits this year." (Humpherys 2009).

The potential of this vagueness puts the focus on the analysis of vagueness expressions in annual reports. While Humpherys (2009) could not confirm the hypothesis, it could be proven by Goel and Gangolly (2012). The theoretical derivation of the hypothesis can also serve as an approach for the investigation of German annual financial statements. Combined with the management concealment hypothesis, Humpherys' (2009) assumption can be linked to the occurrence of a company's bankruptcy. This leads to the possible derivation that when a company is facing bankruptcy, it is more likely to conceal the information of poor earnings and implement it through the linguistic device of vagueness expressions. Based on this assumption, the following hypothesis is formed:

H1.2: The annual financial statements of bankrupt companies have a higher proportion of vagueness expressions compared to solvent companies.

Staying with the theoretical assumption that fraudulent companies tend to be more strategic in hedging when communicating information, Goel and Gangolly (2012) take this consideration further through another hypothesis. This relates to the more frequent use of passive voice as a linguistic device by companies that

deceive. According to this hypothesis, the use of the passive voice would result in the sentence becoming wordy, unclear, vague as well as misleading, thus creating an unclear meaning in the reader's mind (Goel and Gangolly 2012). This hypothesis could be confirmed after analyzing English annual financial statements (Goel and Gangolly 2012). Analogous to the hypothesis development of H 1.2, this thesis can also be included as an approach in relation to corporate bankruptcy prediction. Since the passive voice is usually used when the speaker wants to obfuscate what is happening so that the sentence becomes unclear and misleading (Goel and Gangolly 2012), this complements the obfuscation assumption in a meaningful way. Accordingly, the aspect of whether companies try to disguise the threat of bankruptcy by using passive language more frequently as a linguistic device in German-language financial statements can be tested. This is reflected in the following hypothesis:

H1.3: The annual financial statements of bankrupt companies show a higher proportion of passive language compared to solvent companies.

In addition to the hypotheses regarding delimitation criteria, this section looks at the language development-specific word collocations within annual financial statements over the years. Magnusson et al. (2005) provide a first attempt in this regard. The aim of their study was to visualize key terms and their connections within a company's quarterly report. For this purpose, the authors used the methodology of collocational networks, which was originally developed by Williams (1998). Collocational networks, or collocations, can be defined as any recurring combination of words (Benson 1989). In the study, it was observed that certain groups of words changed over the course of quarters and became increasingly negative during certain periods (Magnusson et al. 2005). Finally, the study found that these negative changes in collocations around words, such as "turnover," correlated with future deterioration in company performance (Magnusson et al. 2005). This research approach can serve as a useful methodology for developing language in corporate financial statements. An analysis of collocations reveals language patterns and meanings that are not apparent from frequency lists of individual words or from manual analysis of larger amounts of text (Baker et al. 2008; Pollach 2012). An interesting aspect that can be studied in language development is the development of collocations around concrete words describing corporate performance. Derived from this, the

question can be asked whether an actual deterioration in performance leading to corporate bankruptcy is reflected negatively in collocations around keywords such as "sales" or "revenue" accordingly. To investigate this question, the following hypothesis is formed:

H2: The sentiment of collocations around from Magnusson et al. (2005) extracted performance keywords in the annual financial statements of bankrupt companies changes increasingly negatively.

Many studies also take sentiment into account when analyzing English-language annual financial statements. However, there are different results in this research area. Lee et al. (2018) showed in a study that negative and positive sentiments do not correlate with the company's sales performance and thus show no relevance in predicting the company's future financial performance (Lee et al. 2018). Goel and Gangolly (2012) examine sentiment in relation to corporate fraud and were also unable to confirm a correlation when looking at the relative distributions of positive as well as negative sentiment (Goel and Gangolly 2012). Hajek et al. (2014) and Lohmann and Ohliger (2020) were able to provide clearer results on the consideration of sentiment. However, while Hajek et al. (2014) revealed that financially distressed companies tend to have a more negative and less positive tone (Hajek et al. 2014), Lohmann and Ohliger (2020) confirmed a different correlation. After analyzing German annual financial statements, the latter study comes to the conclusion that bankrupt companies have a less negative tone. Based on the hypotheses derived through literature, the following research model is examined (see Figure 11).

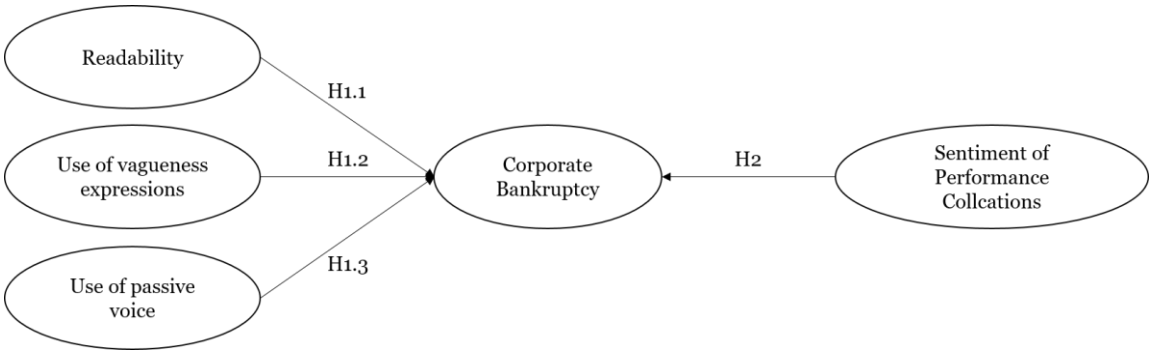


Figure 11. Research model for the analysis of language development

Research Methodology

After deriving the hypotheses from the literature, we consider following our research methodology. First, we describe the procedure for testing the hypotheses with respect to delimitation criteria (H1.1-1.3). We then discuss how the hypothesis concerning language development was tested (H2).

To test the hypothesis (H1.1) regarding the readability of the textual data contained in the financial statement, we will calculate the Readability Index (LIX) and the Flesch Reading Ease Score (FRES) below. These indices use, as an approach, the average sentence length in conjunction with different text features, e.g., the average amount of syllables per word or the number of long words (more than 6 letters), to describe a value that reflects, on a scaled basis, the level of comprehensibility of the text. Since both readability indices were originally developed for other languages, we used the versions adapted to the German language, which can be seen below.

$$(1)\text{FRES} = 180 - \frac{\sum Words}{\sum Sentences} - (58{,}5 \times \frac{\sum Syllables}{\sum Words})$$

$$(2)\text{LIX} = \frac{\sum Words}{\sum Sentences} + \frac{\sum Long\ words}{\sum Words} \times 100$$

For the calculation of the indices, the textacy library, which is based on the well-known *spaCy* Python library (2021), was used to extract the individual text features required. While a small value for the LIX is an indication of an easily readable text, a small value of the FRES is an indication of a very difficult text.

The investigation of hypothesis (H1.2) regarding the use of vagueness expressions is based on the approach of Humphreys (2009). This lexicon was developed following Hyland (1998) considering a financial context and was tested in fraud detection. We use a version of this dictionary translated into German to test whether the use of vagueness terms can also be adapted to the application context of corporate bankruptcy prediction. Furthermore, we used a list of German vagueness terms developed in an economic context to complement the dictionary (Clemen 1998). In this regard, we calculate the absolute number of vagueness terms used using the *Natural Language Toolkit* (NLTK). To test the

hypothesis, we define a hedging score that describes the use of vagueness terms by their ratio to the total number of words.

In order to test the hypothesis (H1.3) regarding the use of passive speech, a pattern must first be defined in order to make it recognizable by machine. The German language distinguishes between the state passive and the process passive (Hinrichs 1999). The process passive is processual and is formed with the conjugated forms of "become" and the past participle II of the full verb. The passive of state, on the other hand, is not processual and is formed with the conjugated form of "to be" and the past participle II of the full verb. Both passive forms described can occur in other tenses besides the present, e.g., preterite or perfect tense. Additionally, the passive can also be formed in combination with modal verbs, e.g., can, may, must, should, and will (Reimann 2021). These considerations serve as a basis for the definition of concrete rules on how to recognize a passive construction. To implement the recognition using text mining, a rule-based procedure is applied. This procedure is similar to the lexicon-based procedure, but instead of using fixed words, this procedure uses a general knowledge base about certain structures. The basis in this case is the structure of passive sentences. The rule-based analysis is implemented using *part-of-speech tagging* (POS tagging) in Python. *POS tagging* assigns certain word classes to words (Kumawat and Jain 2015).

<table>
<tr><td>Pattern 1</td><td rowspan="7">{'TAG':'VAFIN', 'OP':'+'}</td><td rowspan="8">,{'IS_ALPHA':True, 'OP':'*'}</td><td>,{'TAG':'APPR', 'DEP':'spb', 'OP':'+'}</td><td rowspan="2">,{'IS_ALPHA':True, 'OP':'*'}</td><td rowspan="2">,{'TAG':'VVPP','DEP':'oc', 'OP':'+'}</td></tr>
<tr><td>Pattern 2</td><td>,{'TAG':'APPRART', 'DEP':'spb', 'OP':'+'}</td></tr>
<tr><td>Pattern 3</td><td rowspan="6">,{'TAG':'VVPP','DEP':'oc', 'OP':'+'}</td><td rowspan="5">,{'TAG':'VAINF','OP':'+'}</td><td>,{'TAG':'VMINF', 'OP':'+'}</td></tr>
<tr><td>Pattern 4</td><td>,{'TAG':'VAINF','OP':'+'}</td></tr>
<tr><td>Pattern 5</td><td rowspan="3"></td></tr>
<tr><td>Pattern 6</td></tr>
<tr><td>Pattern 7</td></tr>
<tr><td>Pattern 8</td><td>{'TAG':'VMFIN', 'OP':'+'}</td><td>,{'TAG':'VVPP','DEP':'oc', 'OP':'+'}</td><td>,{'TAG':'VAINF','OP':'+'}</td></tr>
</table>

Table 6. Passive pattern construction with spaCy

The implementation of *POS tagging* in Python was again realized with the help of *spaCy* (2021) and the *de_core_news_lg* NLP model. For this, patterns consisting of a certain combination of word classes to be searched for in the text were first created. We use the following pattern to detect passive use in the analyzed financial statements. Analogous to the methodology of testing the use of vagueness expressions, a passive-score is calculated that represents the use of the passive voice in terms of text length, by calculating the quotient of the number of passive constructions used by the number of sentences in the text.

Hypothesis H2 is tested with the help of the *#LancsBox* tool by (Brezina et al. 2021). The tool has been developed for the analysis of speech and can, among other things, extract collocations in text. Following Magnusson et al. (2005) and Williams (1998), the extraction of collocations is performed using the Mutual Information (MI) value. MI is a widely used information-theoretic concept from linguistics (Church and Hanks 1989; Magnusson et al. 2005). The score is calculated by comparing the frequency of the co-occurrence of words and collocations with the frequency of their independent occurrence (Magnusson et al. 2005). Thus, the simple MI score emphasizes the exclusivity of the collocation relationship. A drawback here, however, is that the MI score tends to emphasize unusual combinations, e.g., including nonstandard spellings that occur only once or twice in the corpus (Brezina et al. 2015). (Daille 1995) recognized this problem and revised the calculation of the MI value by giving more weight to the observed frequencies and thus giving a higher value to collocations that occur relatively frequently in the text (Brezina et al. 2015). As a result, unusual word combinations are not emphasized despite their exclusivity in the text. For this reason, the revised MI value, which is called MI3 (Brezina et al. 2015), is used. As a starting point for the extraction, the keywords around which the collocations are to be analyzed must be determined. Hypothesis H2 refers to the words that describe business performance. Following Magnusson et al. (2005), the following performance keywords are defined and analyzed:

Performance Keywords	Sales (Absatz), Revenue (Erlös), Return (Ertrag), Gain (Gewinn), Profit (Profit), ROI (Rendite), Turnover (Umsatz), Growth (Wachstum)

Table 7. Examined Performance Keywords

According to Williams (1998), a limit should be set on the number of words that appear before and after the specified keyword respectively (Brezina et al. 2015). This limitation is set to a number of 5 words each to the left and right of the keyword.

Data Analysis and Results

For the consideration of the first hypothesis (H1.1) the readability indices *FRES* and *LIX* were calculated. Since sentence count and word count were needed for these calculations, these values were additionally determined using the Python library *textacy* by DeWilde (2021). The descriptive statistics were compiled considering the financial statements of all available years and companies. Noticeable here are the different text sizes for solvent and bankrupt companies. While the solvent companies report a minimum number of 61 sentences (429 words), the minimum number for the bankrupt companies only starts at 101 sentences (810 words). The maximum number is also 1471 sentences (18629 words) higher for bankrupt companies. However, these differences in the absolute frequencies of the calculation bases for the readability indices are only slightly reflected in the *FRES* and LIX values. According to the *FRES*, the financial statements of both groups are, on average, classified as very difficult texts. With reference to the *LIX* value, the annual financial statements are also rated as very difficult texts according to Bjornsson (1983). The Kolmogorov-Smirnov-Lilliefors test does not show a clear normal distribution of the data with respect to the FRES and the LIX value. Since this is a violation of the requirements for the parametric mean comparison, this is performed using the Mann-Whitney-U test. When considering the second hypothesis (H1.2), the normal distribution of the data could be confirmed using the Kolmogorov-Smirnov test with a value of $p > 0.05$. In addition, the homoscedasticity of the data set was demonstrated with a value of $p > 0.05$. By these fulfilled conditions the comparison of means is carried out with the help of a t-test. As for the first hypothesis (H1.1), when considering the passive score (H1.3), no normal distribution of the bankrupt data can be assumed and therefore a Mann-Whitney-U test was also chosen for investigation.

		N	Kolmogorov-Smirnov-Liliefors Test sig.	Levene-Test	Median difference	Mann-Whitney-U Test one-sided sig.	t-Test
#Sentences	S	143	<,001		-50	**,021**	
	B	142	<,001				
FRES	S	143	**,200***		1,06	,114	
	B	143	<,001				
LIX	S	143	,015		0,76	,080	
	B	142	**,200***				
Hedging score	S	143	**,200***	**,670***	-,028		0,75
	B	142	**,200***				
Passive score	S	143	**,200***		,743	,345	
	B	142	,003				

Legend: S=Solvent, B=Bankrupt, N=Number of financial statements * p<0,05, ** p<0,01, *** p<0,001

Table 8. Results regarding hypotheses H1.1-H1.3

In summary, none of the hypotheses (H1.1-H1.3) could be confirmed based on our data set (see Table 8). Nevertheless, it was found that both groups differ significantly in terms of the number of sentences with a value of $p < 0.05$ (see Table 8). This again, assuming that publication length correlates with company size, is similar to the findings of Pamuk et al. (2021) which showed this relationship in relation to corporate bankruptcy prediction. For this reason, further analysis is performed with subdivisions in the number of sentences. First, reports of up to 1000 sentences are examined in order to exclude very long annual reports from the analysis. In a further subdivision, reports with a sentence count below 500 are examined. The last subdivision only considers annual financial statements with at least 500 and more sentences. The following Table 9 (long text) and Table 10 (short text) show which one-sided significances result in each case for the subdivisions of the number of sentences. With regard to the readability indices, as well as the use of passive language, no normal distribution could be demonstrated by the Kologomorov-Smirnov-Lilliefors test even within the data samples differentiated by the text length of the financial statements. Therefore, the Mann-Whitney-U test was again chosen to test for significance.

			Mann-Whitney-U Test						t-Test	
#Sentences		N	Median	FRES	Median	LIX	Median	Passive score	Mean	Hedging score
>= 500	S	31	17,53	0,189	76,41	0,401	22,690	**0,016***	0,651	**0,048***
	B	34	15,81		76,43		18,570		0,702	
Legend: S=Solvent, B=Bankrupt, N=Number of financial statements								* p<0,05, ** p<0,01, *** p<0,001		

Table 9. Differentiation of the results regarding long texts

The results show two significant differences between solvent and bankrupt companies exclusively when considering the readability for the LIX, which can be seen in both differentiations of short annual financial statements. No significance for readability criteria can be proven for long annual financial statements.

The analysis of the use of vagueness terms (H1.2) showed that the conditions for performing a t-test are fulfilled by the Kolmogorov-Smirnov-Lilliefors test for normal distribution and by the F-test for variance homogeneity. The considered data in all subdivisions are therefore normally distributed and variance homogeneous. The t-test reveals a significant difference in the vagueness score between solvent and bankrupt companies with a $p = 0.048$ when excluding shorter annual financial statements. The annual financial statements of bankrupt companies have a higher average proportion of vagueness expressions than those of solvent companies.

The two mean comparisons for the use of passive language (H1.3) reveal two significances when analyzing the short and long annual financial statements separately. On the one hand, the short annual financial statements with a sentence count of less than 500 sentences reveal a significant difference in the proportion of passive language between the groups of companies. Accordingly, bankrupt companies with short annual financial statements use more passive language on average than the group of solvent companies that tend to have short annual financial statements. On the other hand, an opposite difference

becomes significant when looking at long annual financial statement reports. It can be seen that bankrupt companies with long annual financial statements use the passive less on average than solvent companies with long reports (see Table 9). For the investigation of the collocations that relate to the company performance within the framework of hypothesis H2, a total of 36 collocations were extracted from the solvent companies using the defined performance keywords, and 40 collocations were extracted from the bankrupt companies. A comparison of the mean values for sentiment based on the MI3 score between solvent (SD=,013; M=,079) and bankrupt (SD=,015; M=,086) companies using the t-test reveals one-sided significance with a value of $p < 0.05$. Accordingly, it can be assumed that there is a fundamentally significant difference in the sentiment of the performance keyword collocations between solvent and bankrupt companies.

			Mann-Whitney-U Test						t-Test	
#Sentences		N	Median	FRES	Median	LIX	Mean	Passive score	Mean	Hedging score
<500	S	112	16,75	0,177	76,76	**0,031***	19,683	**0,040***	0,587	0,200
	B	108	15,84		75,72		20,743		0,607	
<1000	S	134	16,70	0,214	76,87	**0,015***	20,201	0,192	0,602	0,399
	B	121	15,82		75,69		20,689		0,607	
Legend: S=Solvent, B=Bankrupt, N=Number of financial statements								* p<0,05, ** p<0,01, *** p<0,001		

Table 10. Differentiation of the results regarding short texts

A correlation analysis of the collocations of the performance keywords shows a significant correlation between the sentiment of the collocations and the year of publication of the financial statements for all values examined, with a value of $p <= 0.001$. In addition, there is a positive correlation for all values examined. In this case, a positive correlation means that the sentiment becomes more positive the further back the year under consideration is from *year 1*, i.e., the year of the bankruptcy filing. The strength of the respective correlations differs between the solvent and bankrupt companies. Whereas the correlation for bankrupt companies is 0.723, the correlation for solvent companies is 0.482, which is

lower. Thus, it can be observed that the sentiment of performance collocations changes more negatively for bankrupt companies over the years towards bankruptcy than for solvent companies over the same period. The concrete development of the sentiment, taking into account all collocations of the performance keywords, is shown below by a regression curve (see Figure 12). The explained variance of the regression curve with respect to the bankrupt companies has an R-squared value of 0.631 and that of the solvent companies has an R-squared value of 0.712. With regard to the regression curves and taking into account the proven correlations, a negative development of the sentiment of the performance keyword collocations at the bankrupt companies is clearly recognizable and differs from the development of the collocations at solvent companies. Thus, the hypothesis H2 can be confirmed.

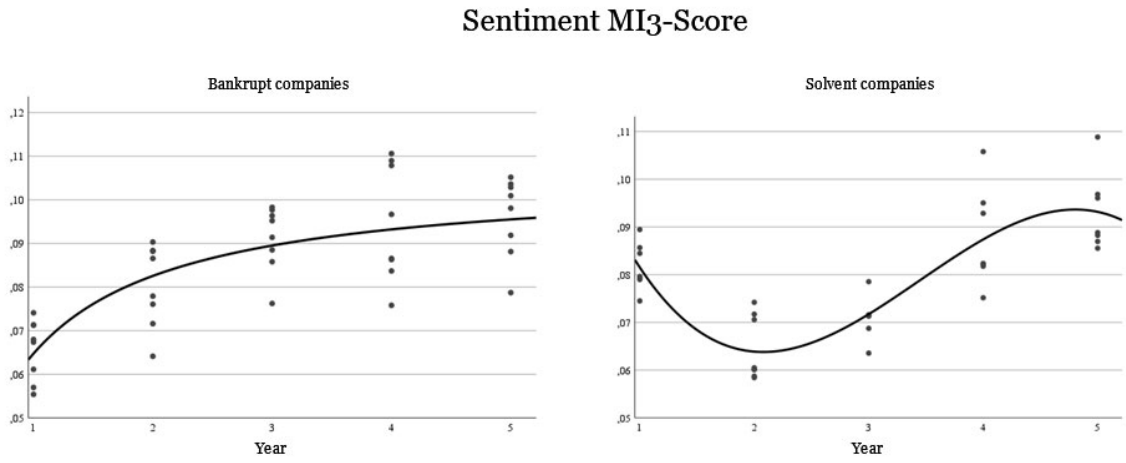


Figure 12. Results of the performance collocation sentiment

Discussion and Implications

This paper examines the extent to which linguistic indicators of textual data in financial statements are useful for forecasting corporate bankruptcy. It examines both statistical correlations and dynamic developments with respect to time series leading up to the occurrence of the bankruptcy filing. They provide an understanding of how qualitative data should be preprocessed to develop an AI-based bankruptcy prediction tool. In summary, narrowing down the data in terms

of text length is an important preprocessing step, as it was shown that significant differences only emerge once similar length reports are compared. This finding for the qualitative analysis is in line with the results in the literature (Ciampi and Gordini 2013; El Kalak and Hudson 2016; Pamuk et al. 2021), which state that subdividing the data in terms of company size also has a positive effect on the bankruptcy prognosis of companies. Although this is a ubiquitous and topical issue with strong practical relevance, the correlations between linguistic indicators and the occurrence of bankruptcy are not well-researched and many studies come to contradictory conclusions based on financial statements of companies from different countries. With respect to the analysis of German financial statements, this paper offers both theoretical and practical implications, as not only static but also dynamic linguistic indicators have been studied.

Contributions to Literature

This study adds to the existing literature not only in terms of examining indicators using another language but also by looking at dynamic aspects of language use. First, the literature on the studies of readability with respect to bankruptcy prediction can be supplemented in several ways (Kloptchenko 2004; Li 2008; Butler and Kešelj 2009; Le Maux and Smaili 2021). It could be shown that limiting the consideration of differentiated text sizes adds value to the information content of established metrics with regard to bankruptcy prognosis and thus can also explain contradictory existing study results, as they have not made any distinction so far. Nevertheless, based on the results, the question must be considered to what extent established readability indices are suitable for specific text forms, such as the financial statement in this case. Dividing financial statements by text length also showed a positive effect in the analysis of the correlation between the use of vagueness expressions and passive constructions.

Second, an adaptation of the Literature streams regarding the use of vagueness expressions and passive language in fraud detection of financial statements to bankruptcy prediction was made. While no significant correlations could be concluded based on the dataset, these results should be tested in further studies as well as for other languages.

Finally, the study extends the results of Magnusson et al. (2005). While the latter showed on the basis of three selected companies that collocations are suitable for estimating the financial performance of companies, we were able to show that this approach is also suitable for company-related bankruptcy prediction. Furthermore, we complemented the methodology by looking at time series of financial statements and were able to show that precisely an inconsistent development of sentiment around these collocations can be a decisive criterion for bankruptcy prediction. This, in turn, establishes a need for analysis of changes in the language style of financial statements for future research.

Subsuming the results of this study, the question can additionally be asked in the future to what extent classical dictionary-based approaches can, but should, be used to distinguish between financial situations of companies since it is precisely the development of suitable linguistic indicators in successive financial statements of a company that can provide more information about the financial development than the consideration of individual financial statements.

Practical Contributions

In addition to complementing the literature, the identified results help analysts to rethink and optimize their methodology when analyzing qualitative financial statement data. In particular, in looking at word collocations, we have shown one way of looking at the sentiment in which linguistic developments can also be easily illustrated for stakeholders. This allows us to conclude that the steady financial development of a company correlates with the linguistic presentation of its financial statements and does not become apparent only when bankruptcy occurs. Consequently, this insight can be used to complement qualitative balance sheet analysis and optimize the accuracy of corporate bankruptcy forecasting. A more accurate bankruptcy prediction has a positive effect overall for all the stakeholders involved, as any resulting damage can be reduced by knowing about it.

Limitations and Future Research

As with any research, our study has some limitations that need to be considered when interpreting the results. First, the most obvious limitation of the case study is that the dataset of 285 financial statements, split across 62 companies considered, only provides insight into the impact of using and changing language in financial statements toward corporate bankruptcy. The findings of this study should be tested in the future using larger German language data sets. Furthermore, with respect to the goal of predicting bankruptcy, we looked at companies that have filed for bankruptcy. In the context of future research, it would also be interesting to investigate the extent to which indicators exist that correlate with the course after filing for bankruptcy. This could be helpful in order to make an assessment of the company's situation beyond the binary bankruptcy forecast. Furthermore, the use of the dictionary-based approaches, as well as the underlying NLP model within the POS tagging, is to be mentioned as a limitation, since these can also only recognize a limited number of expressions or passive constructions due to their finite data structure. Regarding hypothesis H2, it should be noted that the collocations considered were taken from the literature. They are therefore limited by their number and for future research, it would be desirable to explore which collocations allow valid information about the financial situation of companies. Furthermore, the question arises whether sentiment-based approaches alone can be sufficient to interpret collocations. We have shown that a dictionary-based approach for assessing the sentiment of collocations is suitable to obtain important information with respect to the occurrence of bankruptcy, but requires a classification in a time series of financial statements. This finding offers a new perspective on the existing literature on the use of sentiment analysis for bankruptcy prediction based on financial statements, as this literature in particular has focused only on the analysis of individual reports (Nopp and Hanbury 2015; Myšková and Hájek 2020). Further research is needed to investigate the extent to which, rather, changes in linguistic indicators may also be indicative of the occurrence of corporate bankruptcy. Although we have only established this on the basis of German text data, an investigation with regard to qualitative financial statement analysis for the prediction of corporate bankruptcies is certainly relevant for other languages as well.

Conclusion

This study has been motivated by the surge of interest in the analysis of qualitative textual data from financial statements for the prediction of corporate bankruptcy. Although there are existing approaches in the area of analyzing linguistic indicators based on English-language financial statements for bankruptcy prediction, we identified a research gap in looking at German-language reports, and we added a time-series view of collocations to the existing literature. In this case study, we first identified linguistic indicators based on known results on financial statements in other languages, which were then examined using our dataset. Considering the use of AI-based models to predict corporate bankruptcies, our results help to assess the suitability of linguistic features, highlighting in particular structural problems when looking at individual reports statically, as none of the hypotheses H1.1-H1.3 could be confirmed using the data. However, with respect to the research question posed, we have shown that a differentiated view of financial statements of different lengths leads to the possibility of distinguishing between solvent and bankrupt companies, even with the help of factors such as readability and the use of passive language. Furthermore, it could be shown that the use of collocational networks and especially the consideration of the sentiment of these allows conclusions to be drawn about a company going bankrupt. Finally, it should be mentioned that this study presents results based on German annual financial statements. Nevertheless, when examining financial statements in other languages, a time series view of language indicators in terms of a textual analysis should also be considered. A next step would be to compare the change in these indicators with the financial situation of a company based on the financial ratios reported in the financial statements.

IV Consecutive Analysis of Financial Statements

“Analysis of consecutive financial statements concerning bankruptcy prediction”

October 20-21,2022 – RISK2022, Barcelona, Spain

Analysis of consecutive financial statements concerning bankruptcy prediction

Tobias Nießner[1], Matthias Schumann[2]
University of Goettingen[1], Chair of Application Systems and E-Business[2], Germany[1], tobias.niessner@uni-goettingen.de[1], mschuma1@uni-goettingen.de[2]

ABSTRACT

While classical bankruptcy prediction models focus on the analysis of financial ratios extracted from a balance sheet, there is a consensus in research that harnessing textual data from annual reports can optimize these models. Motivated by the expectations of researchers and analysts, we addressed the informative value of unstructured data by looking closely at period of a company's evolving bankruptcy over time. To this end, we used term frequency-inverse document frequency (TF-IDF) to analyze the cosine similarity of consecutive financial reports of 23 solvent and insolvent German companies over 5 years. Our results suggest that optimization of bankruptcy prediction models should not only depend on extracted key figures of individual annual reports text but should also consider the development of those within past years to arrive at a more accurate result. The changes are more revealing than the commonalities, especially concerning the analysis of textual data.

1. INTRODUCTION

While various studies deal with the prediction of corporate insolvencies, it is easy to see that the analysis of textual components of annual financial statements often focuses on static characteristics, e.g. sentiment (Myšková and Hájek, 2020) or readability (Le Maux and Smaili, 2021), of individual documents. In practice, however, an analyst may also be interested in the extent to which planned investments, but also the presentation of risks for a company, are depicted beyond these textual key figures (Lohmann and Ohliger, 2020). In this context, research findings show that the communication of companies to stakeholders can be subject to a wide variety of motives that condition the presentation of developments (Bloomfield, 2002).

Authors: Tobias Nießner and Matthias Schumann

Outlet: *8th Workshop on Risk Management and Insurance Research. Barcelona, Spain, 2022.*

Abstract: While classical bankruptcy prediction models focus on the analysis of financial ratios extracted from a balance sheet, there is a consensus in research that harnessing textual data from annual reports can optimize these models. Motivated by the expectations of researchers and analysts, we addressed the informative value of unstructured data by looking closely at period of a company's evolving bankruptcy over time. To this end, we used term frequency-inverse document frequency (TF-IDF) to analyze the cosine similarity of consecutive financial reports of 23 solvent and bankrupt German companies over 5 years. Our results suggest that optimization of bankruptcy prediction models should not only depend on extracted key figures of individual annual reports text but should also consider

the development of those within past years to arrive at a more accurate result. The changes are more revealing than the commonalities, especially concerning the analysis of textual data.

Introduction

While various studies deal with the prediction of corporate bankruptcies, it is easy to see that the analysis of textual components of annual financial statements often focuses on static characteristics, e.g., sentiment (Myšková and Hájek 2020) or readability (Le Maux and Smaili 2021), of individual documents. In practice, however, an analyst may also be interested in the extent to which planned investments, but also the presentation of risks for a company, are depicted beyond these textual key figures (Lohmann and Ohliger 2020). In this context, research findings show that the communication of companies to stakeholders can be subject to a wide variety of motives that condition the presentation of developments (Bloomfield 2002).

In the following, the question arises to what extent an impending bankruptcy of a company can also be anticipated by stakeholders through an analysis of the development of textual components of the annual financial statements. Of particular interest is whether new information can be extracted from looking at textual data compared to previous years.

RQ: How does the information sharing within financial statements change about developing bankruptcy?

We start with a presentation of our data acquisition and selection process. Then, we describe our methodology, present our findings and discuss them concerning our research question. In the end, we unfold limitations and complete our paper with a conclusion.

Data Collection

We refer in this study to a data set of published annual financial statements by the german central platform *Bundesanzeiger* for official announcements as well as for legally relevant company news. Therefore, within the company selection process, we identified bankrupt medium-sized companies in Germany using the *Amadeus database of Bureau van Dijk*. Based on the criterion of company size, we were able to ensure that the companies we selected are required by German law to publish a management report within their annual financial statements (HGB 2021b, 2021a) and thus provide a suitable amount of textual data for the

company's self-presentation to external stakeholders. Furthermore, it was determined that only bankrupt companies that published their last annual financial statements after 2017 were considered. However, to avoid any bias due to temporal selection, the financial statements of the solvent peer companies from the same years were chosen. To be able to compare a suitable and comparable set of consecutive annual financial statements, it was further ensured by the *Bundesanzeiger* platform that five consecutive annual financial statements are available for the consideration of the change until the occurrence of a company's bankruptcy. To additionally ensure that the information in the database is up to date, we checked with the German portal for bankruptcy announcements (InsBekV 2021) and whether official bankruptcy applications had been filed. To obtain a corresponding control group for comparison purposes, the *Amadeus database* was used to identify solvent companies in terms of company size and industry affiliation. A total of 23 solvent and bankrupt companies were identified for the analysis (see Table 11).

Bankrupt companies	Solvent companies
Company name	Company Name
Alma-Küchen GmbH & Co. KG	3B GmbH
AWG GmbH	Aerologic GmbH
Bachtrup GmbH	Ascent AG
Böhm AG	BCD Travel GmbH
Curasan AG	Beckermann Küchen GmbH
Clean Garant GmbH	Christophorus Trägergesellschaft mbH
Deutsche R + S GmbH	DO & CO Holding GmbH
Envirotherm GmbH	Dr. Rehfeld Fashion AG
Esprit Retail B.V & Co. KG	Gamestop Deutschland GmbH
Galeria Karstadt Kaufhof GmbH & Co. KG	Hilfiger Stores GbmH
Germania Fluggesellschaft mbH	Holmer Maschinenbau GmbH
GerryWeber AG	IBR Gebäudem. GmbH & Co. KG
Greensill GmbH	Kaco GmbH & Co. KG
Hallhuber GmbH	Kaufland Stiftung & Co. KG
Katharina Kasper ViaSalus GmbH	Logaer Maschinenbau GmbH
Kath. Klinikum Oberhausen GmbH	NKD Deutschland GmbH
Klier Holding GmbH	RENAFAN Holding GmbH

Thomas Cook GmbH	Stadtwerke Groß-Gerau GmbH
Spiele Max GmbH	TAKKO Holding GmbH
Vapiano SE	Vinzenz Murr GmbH
Veritas AG	Webac Holding AG
Vidrea Deutschland GmbH	Wortmann GmbH
Wilke Waldecker GmbH & Co. KG	Zara Deutschland B.V. & Co. KG

Table 11. Overview of companies

Methodology

To analyze the linguistic change of texts, various levels of granularity, i.e., character, word, sentence, and document, can be distinguished according to the taxonomy of Fromm et al. (2019). In this study, we consider a syntactic level of the linguistic analysis of financial statements and subsequently use the term frequency-inverse document frequency (TF-IDF) to obtain a representation of the individual documents in the form of a vector. To assess the change of the individual financial statements, which are considered in time series of five years each, we use the cosine similarity of the vectors calculated in this way among each other as a metric of change. This methodology is widely used in IS research (Bankamp et al. 2021) and has achieved good results in uncovering new and reducing redundant information (Zhang et al. 2002). Nevertheless, we are aware that there is also a strong focus in recent research on other document representation methodologies, e.g., latent dirichlet allocation, word embeddings, and document embeddings, in finance and accounting (Bankamp and Muntermann 2022). However, since we have to start from the classical approach of documents that contain relatively rigid content over years, we consider our chosen approach as experimental in the context of this study to fundamentally show the tendency of change concerning occurring bankruptcy.

The cosine similarity describes the cosine of the angle, i.e., α, between two vectors, i.e., x and y, and can thus be represented by the scalar product of both vectors and the euclic norm for the distance in the following formula:

$$\textit{Cosine Similarity}(x,y) = \cos(\alpha) = \frac{x \cdot y}{||x|| \cdot ||y||}$$

Since we consider frequency vectors we get a range of values from 0 to 1. It is directly obvious from the formula that a smaller angle, i.e., a greater commonality

of both documents, is associated with a higher value for the cosine similarity. For the individual calculations of cosine similarity and the term-frequency inverse document frequency, it is recommended to use the scikit-learn Python package, for example (scikit-learn 2022).

Results and Discussion

In the following, we use *Y* to denote the year in which an bankruptcy petition was filed and, consequently, *Y-n* to denote the *n*-th year before that. Since distance matrices are symmetric by definition, Table 12 and Table 13 are each upper triangular matrices. We see directly that the cosine similarities among financial statements regardless of whether bankrupt or solvent companies are considered have a very high lower bound of about 0.94.

Cosine Similarity				
Y	Y-1	Y-2	Y-3	Y-4
1	0.9688	0.9445	0.9409	0.9327
	1	0.9648	0.9561	0.9453
		1	0.9729	0.9563
			1	0.9710
				1

Table 12. Distance matrix of bankrupt companies

It is directly interesting to see that just before bankruptcy occurs, stronger changes are evident than in solvent companies and a one-time larger difference in Y-2 to Y is also visible, while the observation of the solvent companies suggests that a constant change is present.

Cosine Similarity				
Y	Y-1	Y-2	Y-3	Y-4
1	0.9788	0.9631	0.9500	0.9430
	1	0.9750	0.9570	0.9484
		1	0.9723	0.9570
			1	0.9749

1

Table 13. Distance matrix of solvent companies

If we also look at the upper secondary diagonal of both matrices, it also becomes clear that this difference essentially becomes more apparent when we look at a longer time series, since according to our calculations, successive annual financial statements remain more or less constant for both bankrupt and solvent companies. Nonetheless, the similarity score of consecutive financial statements for bankrupt companies averages 0.969, slightly lower than that of solvent ones at 0.975.

Our study is, of course, subject to limitations. It should be noted that Lang and Stice-Lawrence (2015) were able to show in a study using an analysis of the cosine similarity of pairwise consecutive annual financial statements of various companies from different countries that the adaptation of IFRS led to a significant increase in new information. An interesting consideration is the impact of market-changing situations, such as the COVID-19 pandemic. In *Delivery Hero's* annual financial statement, it is clear that this situation had a major impact on reporting, i.e., 65 mentions in 2020 (Delivery Hero 2020) and 56 mentions in 2021 (Delivery Hero 2021). It must therefore be assumed that a change in reporting can occur for a variety of reasons and cannot be regarded solely as a decision criterion for the occurrence of bankruptcy. Nevertheless, our analysis shows that a supporting function is given and future research could further consider whether a separate consideration and classification of the new information identified in this way can provide further information on a change in the financial situation of a company. Furthermore, it has to be considered that our research is also limited in terms of the amount and scope of the data and should therefore be validated using a larger and balanced sample. Our methodology is suitable for this study to map informal changes and make them measurable, but it is subject to the limitation that no attention is paid to the order of content. Nevertheless, it is also of interest in which order information is disclosed in a financial statement. Considering the sheer length of such a report, the question arises for future research to what extent the placement of information may be motivated by a company.

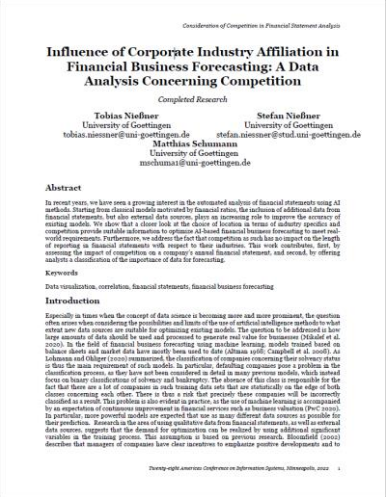
Conclusion

We were able to show that differences between solvent and bankrupt companies can be made visible in the presentation of information in annual financial statements, which offer an approach to support the bankruptcy prediction of companies. Nevertheless, it must be critically reflected that the financial situation and development alone are not necessarily decisive for a change in reporting. Nevertheless, our results concerning the selected data set also indicate that we should not expect extreme changes in consecutive financial statements. Concerning current research on the analysis of textual components of annual financial statements, we thus raise the question of whether an analysis of entire financial statements is instrumental in the assessment of the financial situation of companies or whether an analysis of the quantity difference is of more interest.

V Industry Affiliation in Financial Business Forecasting

“Influence of Corporate Industry Affiliation in Financial Business Forecasting: A Data Analysis Concerning Competition”

August 10-14,2022 – AMCIS, Minneapolis, USA



Consideration of Competition in Financial Statement Analysis

Influence of Corporate Industry Affiliation in Financial Business Forecasting: A Data Analysis Concerning Competition

Completed Research

Tobias Nießner
University of Goettingen
tobias.niessner@uni-goettingen.de

Stefan Nießner
University of Goettingen
stefan.niessner@stud.uni-goettingen.de

Matthias Schumann
University of Goettingen
mschuma1@uni-goettingen.de

Abstract

In recent years, we have seen a growing interest in the automated analysis of financial statements using AI methods. Starting from classical models motivated by financial ratios, the inclusion of additional data from financial statements, but also external data sources, plays an increasing role to improve the accuracy of existing models. We show that a closer look at the choice of location in terms of industry specifics and competition provide suitable information to optimize AI-based financial business forecasting to meet real-world requirements. Furthermore, we address the fact that competition as such has no impact on the length of reporting in financial statements with respect to their industries. This work contributes, first, by assessing the impact of competition on a company's annual financial statement, and second, by offering analysts a classification of the importance of data for forecasting.

Keywords

Data visualization, correlation, financial statements, financial business forecasting

Introduction

Especially in times when the concept of data science is becoming more and more prominent, the question often arises when considering the possibilities and limits of the use of artificial intelligence methods to what extent new data sources are suitable for optimizing existing models. The question to be addressed is how large amounts of data should be used and processed to generate real value for businesses (Nikolaï et al. 2020). In the field of financial business forecasting using machine learning, models trained based on balance sheets and market data have mostly been used to date (Altman 1968; Campbell et al. 2008). As Lohmann and Ohliger (2020) enumerated, the classification of companies concerning their solvency status is thus the main requirement of such models. In particular, delimiting companies pose a problem in the classification process, as they have not been considered in detail in many previous models, which instead focus on binary classifications of solvency and bankruptcy. The absence of this class is responsible for the fact that there are a lot of companies in such training data sets that are statistically on the edge of both classes concerning each other. There is thus a risk that precisely these companies will be incorrectly classified as a result. This problem is also evident in practice, as the use of machine learning is accompanied by an expectation of continuous improvement in financial services such as business valuation (PwC 2020). In particular, more powerful models are expected that use as many different data sources as possible for their prediction. Research in the area of using qualitative data from financial statements, as well as external data sources, suggests that the demand for optimization can be realized by using additional significant variables in the training process. This assumption is based on previous research. Bloomfield (2002) describes that managers of companies have clear incentives to emphasize positive developments and to

Authors: Tobias Nießner, Stefan Nießner and Matthias Schumann

Outlet: *28th Americas Conference on Information Systems (AMCIS), Minneapolis, Minnesota, USA, 2022.*

Abstract: In recent years, we have seen a growing interest in the automated analysis of financial statements using AI methods. Starting from classical models motivated by financial ratios, the inclusion of additional data from financial statements, but also external data sources, plays an increasing role to improve the accuracy of existing models. We show that a closer look at companies’ location selection in terms of industry specifics and competition provides suitable information to optimize AI-based financial business forecasting to meet real-world requirements. Furthermore, we show that the local competition with companies of the same industry has no impact on the length of reporting in financial statements. This work contributes, first, by assessing the impact of competition on a company's annual financial

statement, and second, by offering analysts a classification of the importance of data for forecasting.

Keywords: Data Visualization, Correlation, Financial Statements, Financial Business Forecasting

Introduction

Especially in times when the concept of data science is becoming more and more prominent, the question often arises when considering the possibilities and limits of the use of artificial intelligence methods to what extent new data sources are suitable for optimizing existing models. The question to be addressed is how large amounts of data should be used and processed to generate real value for businesses (Mikalef et al. 2020). In the field of financial business forecasting using machine learning, models trained based on balance sheets and market data have mostly been used to date (Altman 1968; Campbell et al. 2008). As Lohmann and Ohliger (2020) summarized, the classification of companies concerning their solvency status is thus the main requirement of such models. In particular, defaulting companies pose a problem in the classification process, as they have not been considered in detail in many previous models, which instead focus on binary classifications of solvency and bankruptcy. The absence of this class is responsible for the fact that there are a lot of companies in such training datasets that are statistically on the edge of both classes concerning each other. There is thus a risk that precisely these companies will be incorrectly classified as a result. This problem is also evident in practice, as the use of machine learning is accompanied by an expectation of continuous improvement in financial services such as business valuation (PwC 2020). In particular, more powerful models are expected that use as many different data sources as possible for their prediction. Research in the area of using qualitative data from financial statements, as well as external data sources, suggests that the demand for optimization can be realized by using additional significant variables in the training process. This assumption is based on previous research. Bloomfield (2002) describes that managers of companies have clear incentives to emphasize positive developments and to conceal negative ones when preparing financial statements. In our paper, we, therefore, analyze financial statements according to their content based on topics that are typically included, e.g., research, chances, and risk reports. Furthermore, research results for U.S. companies show that company structure-describing variables can be used to improve bankruptcy prediction (Jones 2017). In addition, however, a company's choice of location also has an impact on its success (Hack 1999; Barovick and Steele 2001). To investigate the suitability of variables from these described

directions of data acquisition in the following, we formulate the following research question for this paper:

RQ: What influence does corporate industry affiliation have on the preparation of financial statements and companies' financial success?

To answer this research question, this paper is structured as follows. We first introduce the research background being used as well as the use of qualitative and external data within financial statement analysis for financial business forecasting. Subsequently, we describe the process of data acquisition and preprocessing. Then, motivated by the literature, we derive hypotheses that will help to answer the research question. Afterward, we analyze and discuss our findings. In addition, we mention limitations and show future research opportunities as well as theoretical and practical implications. Finally, we give a brief conclusion.

Related work

In this section, we review the published literature on the use of disparate data to develop AI-based models for financial business forecasting. We consider language-independent studies to show what types of data have been considered so far to optimize the quality of such models. While simple classical models mostly consider the balance sheet of an annual financial statement as the main source of their information, novel machine learning techniques also allow to process of larger amounts of data and accordingly exploit additional data that might have been considered useless so far. In recent years, textual data from annual financial statements was considered to have great potential when it comes to forecasting future financial developments (Ravisankar et al. 2011; Hajek et al. 2014). A separate stream of literature has developed that deals with the extent to which information can be extracted and used from texts in a suitable manner (Kloptchenko 2004). The calculation of sentiments (Caserio et al. 2020), readability indices (Luo and Zhou 2020), and the use of specific expressions (Magnusson et al. 2005) play a major role here. In addition, the extent to which this can unite quantitative financial business forecasting with qualitative

prediction is investigated (Kloptchenko et al. 2004a). Moreover, the overall analysis of all components of a financial statement is not the only option. Further research has shown that externally sourced corporate metrics also have the potential to more accurately classify the information extracted from financial statements and thus support financial business forecasting models (Jones 2017). The assumption that the consideration of different company sizes is a factor that has a decisive effect on the quality of the bankruptcy prediction of companies has been proven by various studies (El Kalak and Hudson 2016; Pamuk et al. 2021). A more detailed examination of which factors play a role in the different classes of company size has not yet been carried out. In their research on bankruptcy prediction based on German financial statements, Lohmann and Ohliger (2020) showed that the feature of text length, as well as sentiment and the mentioning of various key phrases, e.g, if existential risks endangering the company or whether there is an implemented, ongoing, or planned restructuring, in risk reports of financial statements are suitable to distinguish solvent from bankrupt as well as from defaulting companies. However, there is a lack of consideration of features concerning all quantitative components of financial statements. Furthermore, there are currently no studies that examine a company's local environment based on the industry affiliation in terms of its financial performance.

Data set

In this paper, we consider an industry-independent dataset consisting of 860,234 financial statements of 370,679 different German companies. This dataset, consisting of individual XML files, was supplemented by data from the *Amadeus* database (Bureau van Dijk 2021) to classify the financial status of the companies and thus label the financial statements. The data show whether a company became bankrupt according to German Bankruptcy Law (InsO 2022) and filed for bankruptcy (InsBekV 2021) or defaulted before the end of 2021. Descriptive statistics of the data set underlying this paper can be viewed in Table 14. The financial statements were published in the period 2017-2019. Furthermore, data describing the internal company structures, as well as its external partnerships were collected in this step. In addition, it was possible to ascertain the industry affiliation of each company and to distinguish them according to their location selection in features, i.e., address, federal state. The selected locations could be

characterized more precisely by using an additional data set (Zensus 2022) on the distribution of the number of inhabitants in districts within Germany, as well as, data from Statista (2022) about the size of federal states in Germany. For the analysis, the text of the financial statements was extracted using the BeautifulSoup package in Python (Richardson 2021). In this regard, we have used the python integrated multiprocessing package to allow an acceptable runtime based on parallelization of the process even with the large mass of textual data.

Year	Number of companies	Solvent	Defaulting	Bankruptcy
2017	286,764	283,546	2,601	617
2018	302,001	299,275	2,236	490
2019	271,469	270,114	1,076	279

Table 14. Descriptive statistics of the data set

In terms of data analysis, we extracted the following independent variables (see Table 15) and preprocessed them according to the use of statistical procedures, i.e., Random Forest (Nassirtoussi et al. 2014). As the Random Forst can handle nondichotomous dependent variables, no further adjustment of the financial status is needed. The independent variables were differentiated into three areas, i.e., competition-based variables, company-related personnel structure variables, and report-based variables. A differentiation of companies in their industry is realized in the following by the competitor density, which shows the ratio in relation to a federal state. While Jones (2017) analyzed the ownership structure of companies we consider independent variables from the human resource structure of a company, e.g., the number of employees. For the consideration of the report-based variables, the text of the financial statements was quantified to examine their impact on the financial position of a company. General variables such as text length and punctuation were considered, as well as content-related indicator variables that represent and quantify the use of selected words within annual financial statements. These words have been chosen based on the

content of annual financial statements by the German Commercial Code §289 (HGB 2021b).

Independent variables	Description
Population density	Number of inhabitants of a federal state divided by its total size
Competitor density	Number of companies operating in the same industry in the same federal state divided by the number of all companies in the state
Number of employees	Facts and figures have been collected using the *Amadeus* database (Bureau van Dijk 2021)
Number of shareholders	
Number of branches	
Number of subsidiaries	
Text length	The length of the textual data within the financial statements available in XML format in characters
Punctuation ratio	The ratio of occurrence of dots, commas, question marks, and exclamation marks to the total length of the report
Report components ratio	Describes the ratio of occurrence of words, i.e., risks, opportunities, problems, challenges, possibilities, projects, research or bankruptcy, to the total length of the report

Table 15. Feature selection

Hypotheses Development and Research Methodology

Based on existing literature toward the use of various data sources, e.g., company-related structure data and qualitative financial statement data, that extends existing models for financial business forecasting, we developed a research model to study the effectiveness of three different data examples (see Figure 11). The research model aims to investigate whether variables related to the location of a company compared to variables describing the structure of a company are more suitable for the development of AI-based models. Considering

worldwide globalization, companies are increasingly exposed to new competitors that compete with them and have a corresponding influence on their financial success. In this respect, it could be shown that specific competitive situations influence the choice of the digital business strategy of companies (Mithas et al. 2013). Given this, the question arises to what extent competition as such differs within different industries. An approach to differentiate characteristics for bankruptcy prediction of companies based on the consideration of different industries was shown by Tanaka et al. (2019).

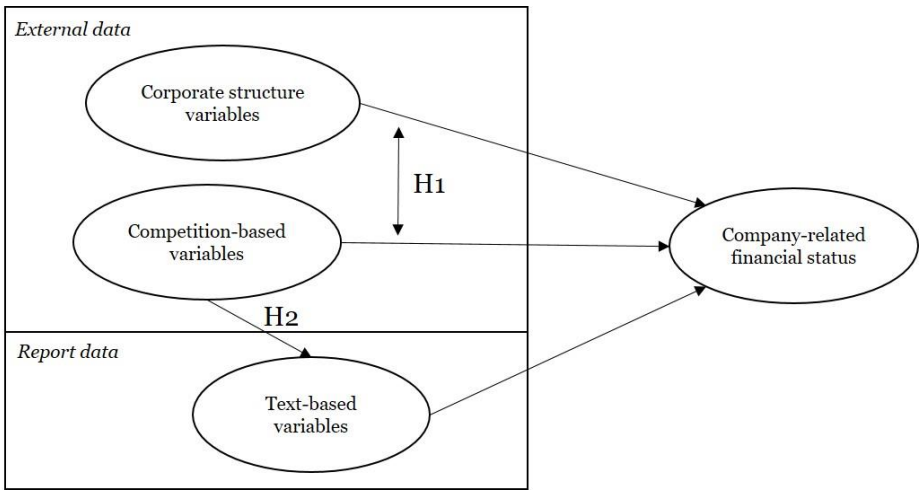


Figure 13. Research model for analyzing the influence of industry affiliation

Jones (2017) examined the ownership structure of companies in a multidimensional framework and showed that it has a high weight in the prediction of bankruptcies. However, the study did not analyze the internal personnel structure. Regarding the size of the company, the study validated the findings of Precourt and Oppenheimer (2015) that the probability of bankruptcy is higher for companies with a smaller shareholding than for companies with a high shareholding, ceteris paribus. In the following, we assume that the size of a company is correlated with its personnel structure and conclude that variables describing this structure influence the probability of bankruptcy. However, since an industry-specific view creates a more suitable comparative framework for comparing companies of the same size, we define the following hypothesis to be tested:

H1. Competition-based variables have a greater impact on predicting the financial status of a company than variables describing the personnel and company structure.

In the literature, the consideration of qualitative financial statement data is considered to be of great importance for optimizing the classic bankruptcy prediction of companies based on financial ratios (Kirkos 2015; Loughran and McDonald 2016; Luo and Zhou 2020). Extracted features from texts can be distinguished by their granularity, i.e., character, word, sentence, or document, as well as the type of linguistic analysis, i.e., morphological, lexical, syntactic, or semantic. Now, while finely granular features only allow for a one-dimensional view, coarser approaches, such as sentiment analysis, collocation-based approaches, or even dictionary-based ones that refer to entire sentences or the document as a whole can allow for a multidimensional view (Fromm et al. 2019). If we compare the consideration of current research regarding the development of AI-based models on quantitative financial metrics, initial studies have been able to show that an industry-specific focus when evaluating data has a positive effect on the predictive accuracy of corporate bankruptcies (Tanaka et al. 2019; Roumani et al. 2020). As we know, companies have an interest in informing stakeholders about, e.g., their IT-related activities. Different methods are used to do this, depending on the industry (Zmud et al. 2010). So we also have to ask to what extent competition affects reporting in financial statements. In this respect, we consider the following second hypothesis:

H2. Competitor density concerning the specific industry of a company has an impact on the text length of a financial statement.

Considering the use of the data to train AI-based models for corporate bankruptcy prediction, we decided to investigate the suitability through an application of the Random Forest model (Breiman 2001), since it is precisely the class of tree-based ensemble learners to which it belongs that is the focus of recent studies (Kirkos 2015; Kim 2018; Hsu and Lee 2020). In this regard, we used a stratified test-train split to divide the dataset and used the random forest implementation in the scikit-learn library for Python (scikit-learn 2022). To evaluate feature importance, we do not choose the classic *Mean Decrease in impurity* (MDI) as the risk of overfitting cannot be ruled out. Considering this possible problem due to the data used to train the random forest, we calculate the *Mean Decrease in Accuracy* (MDA). In the evaluation of the MDA, the association between the

feature and the target is removed. The values of the feature are randomly interchanged and the resulting increase in error is measured. In particular, the influence of the correlated features is removed.

Data Analysis and Results

In a first step, based on the features presented in Table 15 we trained a random forest that aims to classify the financial status of a company. Here, Figure 14 shows how much accuracy the trained model loses by excluding each variable. The greater the value of a feature, the more important it is for the model. Contrary to expectations, the evaluation of the permutation significance of the considered characteristics shows that the employee development, as well as the shareholding structure of a company, has a higher significance for the differentiation of solvent, defaulting, and bankrupt companies than the predefined competitor density as well as its location, represented by its population density.

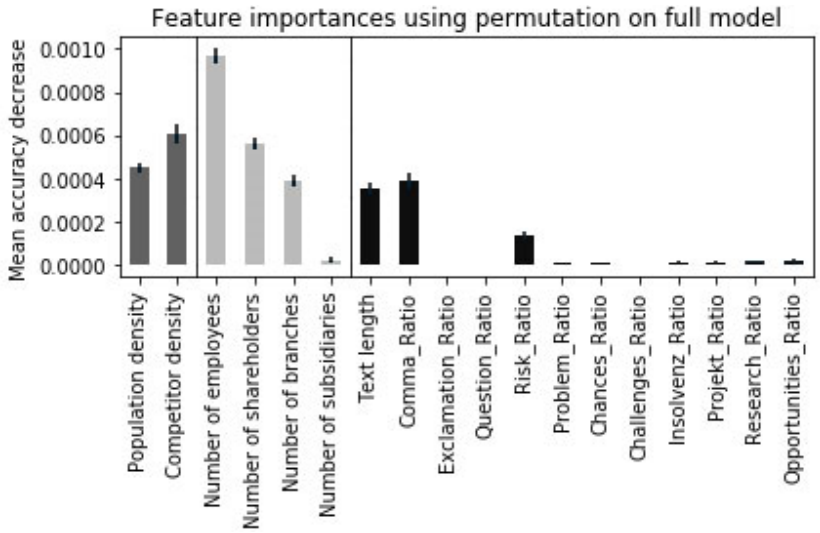


Figure 14. Mean decrease in accuracy (Permutation importance)

A clear trend emerges also about the report components represented by indicator variables about the report length. In doing so, we can validate that the risk report has a special position relative to other components in distinguishing between solvent, distressed, and bankrupt companies.

In the consideration of the text length, as well as the used punctuation, it shows that in the corporate language only the usage of comma provides a useful statement regarding punctuation. This in turn ties in with existing research findings regarding the readability of reports, as the heavy use of commas and thus the use of nested sentences is an indication of a complex sentence structure in use (Kloptchenko 2004; Ajina et al. 2016; Le Maux and Smaili 2021).

To analyze the data concerning hypothesis H2, we first considered the texts in general. We were able to determine that, due to an industry-unspecific consideration about the character length of the texts, only 55,250 texts of different lengths were included in the data set in total. To clean up this situation, we removed financial statements of the same length based on the differentiation of the data about their industry classification. In the following, we have examined a data set with 273.631 financial statements, divided into 78 different industry classes based on NACE Rev. 2 about the correlation of competitor density and text length (eurostat 2022). Based on the hierarchical representations in NACE Rev. 2, the first two digits of the coding were used to distinguish the respective industries. The median is 2686 financial statements per considered industry. In light of the findings regarding the consideration of Figure 12, in addition to considering the correlation between text length and competitor density, we conducted an analysis of other correlations, i.e., number of employees, comma ratio within the text, population density, and the amount of occurrence of risk statements in the text. An examination of the frequency of correlations of varying strength concerning industries can be seen in Table 16. In this regard, we calculated Pearson r and, accordingly, mirrored a division into strong ($r > 0.5$), moderate ($0.5 < r < 0.3$), weak ($0.3 < r < 0.1$), and no correlation (corresponding to the same scale for negative correlations).

Correlation with Competitor density	Strong corr. (negative)	Moderate corr. (negative)	Weak corr. (negative)	No corr.
Population density	18 (7)	8 (19)	8(13)	5
Number of employees	1 (0)	2 (0)	1 (4)	70
Text length	0 (0)	0 (0)	4 (6)	68

Comma_Ratio	0 (0)	0 (0)	3 (3)	72
Risk_Ratio	0 (0)	0 (0)	1 (8)	69

Table 16. Results of the correlation analysis about industries (n=78)

As can be seen, just 10 (4 positively correlated, 6 negatively correlated) of the 78 industries considered show a correlation, weak at all, between competitor density and the text length of companies' annual financial statements. Although competitor density has a positive influence on the forecast of a company's financial status (see Figure 14) it can be concluded that there is no influence on corporate reporting, which accordingly does not show any quantitative reflection of competitive pressure. A similar result can be seen for the influence of competitor density on the number of employees, as well as the ratio of commas and risk statements to text length, which also shows only weak correlations, if any, in a few sectors (see Table 16). Only about population density can stronger correlations be found for specific industries.

Discussion and Implications

First of all, we have to note that no feature of the text analysis was actively influenced by the competition a company faces. Nevertheless, we were able to show that competition as such has an important role in financial business forecasting, although marginally less than the corporate structure as such. While working with real-typical data, we did not perform any cleaning of outliers within our data set. Therefore, we performed an analysis of the median report length (in characters), which showed that it differs for the solvent (M=357), defaulting (M=397), and bankrupt (M=464) classes. This is an indication that it is defined by influencing factors other than a company's competition. Regardless of the industry affiliation of the companies studied in the dataset, we additionally found that the competitor density of companies varies given their financial status and location. Again, based on median analysis, it was found that the competitor density of solvent companies differs locally from that of companies in arrears and bankrupt companies in all the federal states considered. With regard to a data selection for AI-based models for financial business forecasting, it might thus be

advisable that industry-specific ratios are formed that document the competition of companies. Accordingly, the influence of an increase or decrease in competition in an industry can be taken into account over time in future forecasts.

Limitations and Future Research

Like any other research paper, this study is subject to limitations. First of all, the dataset has to be considered, which consists of the linkage of a dataset of XML-based financial statements and has been appropriately matched with another independent data source from Bureau van Dijk for the classification of financial status. We thus have only one dependent point in time for determining financial status, rather than determining it dynamically over the 2017-2019 period under consideration. Secondly, only publicly available data could be used to consider each data dimension of the research model presented. Since data acquisition is not subject to the same standards internationally, research findings related to company-based bankruptcy forecasting are scarce compared to English-language studies. Nevertheless, the results of this paper provide starting points for future research. The focus on competition-oriented factors and the thus concretized consideration of industry-specific data showed clear suitability for distinguishing the financial status of companies. Furthermore, future research should not focus solely on qualitative financial statement data but should investigate to what extent quantified features of the textual data improve financial business forecasting. In light of the trend of using ensemble learning methods in AI-based bankruptcy prediction, the use of such variables in the training process should be investigated.

Practical and Theoretical Contributions

Based on the expectations for improved AI models for corporate financial forecasting, practitioners are not the only ones asking for transparency research about how, why, where, when, and what data can be used by companies for public benefit (Meadows et al. 2022). Regarding the classification of our research in the field of business analytics, the research results presented here contribute to different research streams in this area, i.e., the role of data actors and data capacities and availability (Pappas et al. 2018). By understanding the importance of presenting competitive descriptive metrics in financial business forecasting, we

show how a participatory bottom-up approach can be designed between analysts of financial statements regarding data acquisition. Improved communication about data needs can support both automated model development and maintenance of a forecasting model. Especially given the high value of managing partnerships in practice (Deloitte 2018b), this research offers an approach to determine the value of data and can thus help to minimize strategic and operational risks of data-sharing partnerships. Since we show results that are limited to a specific country in terms of data analysis, our research shows a starting point to consider international competition in future research. This could also identify possible cross-national network effects of industries that influence the success of companies. According to our results, industry considerations, as well as direct competition, should be investigated through the formation of appropriate metrics. Overall, from a research perspective, our paper contributes to a broader understanding of how competition can be applied to traditional financial statement analysis in terms of financial business forecasting, i.e., we improved the understanding of industry affiliation based on a view on competition as a feature within future model development.

Conclusion

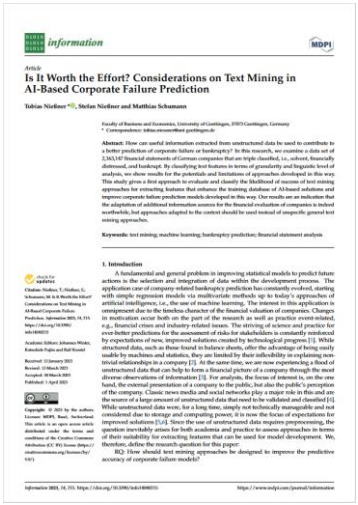
This study was motivated by the interest in using additional data sources in AI-based financial business forecasting. Since current models are mostly trained using financial ratios from the balance sheet of a financial statement, the question arises to what extent new data sources could be suitable to provide insights into the financial situation of companies that do not become clear by looking at a balance sheet alone. Although there are already several studies dealing with diverse data sources, we identified a research gap in the consideration of German-language reports and investigated a large data set in this respect. Here, we were able to show through statistical data analysis using random forest that the consideration and generation of local competition-related variables is a useful addition to research that already considers industry-specific factors to be suitable for optimizing bankruptcy forecasting models. Nevertheless, we have to reject hypothesis H1 because, as shown, employee development has higher feature

importance concerning MDA. About the second hypothesis, we investigated the extent to which correlations exist between local competitor density and the text length of a financial statement. In particular, we also analyzed the component of risk statements, commas used, and the local population. About the second hypothesis, we investigated the extent to which correlations exist between local competitor density and the text length of a financial statement. In particular, we also analyzed the component of risk statements, commas used, and the local population. By looking at 78 different industry classes according to NACE Rev. 2, we concluded that the quantification of local competitive pressure of a company influences its reporting. We thus also need to reject hypothesis H2. and, in the next step, propose that a multidimensional analysis connected with financial ratios reported in the annual financial statements should be done. All in all, we were able to show insights that enable analysts to narrow down the view of potential data sources for the evaluation of companies and to classify the information content of competition in this context.

VI Text Mining in AI-based Corporate Failure Prediction

“Is it worth the effort? Considerations on Text Mining in AI-based Corporate Failure Prediction”

April 2023 – Information



information

MDPI

Article

Is It Worth the Effort? Considerations on Text Mining in AI-Based Corporate Failure Prediction

Tobias Nießner *, Stefan Nießner and Matthias Schumann

Abstract: How can useful information extracted from unstructured data be used to contribute to a better prediction of corporate failure or bankruptcy? In this research, we examine a data set of 2,163,147 financial statements of German companies that are triple classified, i.e., solvent, financially distressed, and bankrupt. By classifying text features in terms of granularity and linguistic level of analysis, we show results for the potentials and limitations of approaches developed in this way. This study gives a first approach to evaluate and classify the likelihood of success of text mining approaches for extracting features that enhance the training database of AI-based solutions and improve corporate failure prediction models developed in this way. Our results are an indication that the adaptation of additional information sources for the financial evaluation of companies is indeed worthwhile, but approaches adapted to the context should be used instead of unspecific general text mining approaches.

Keywords: text mining; machine learning; bankruptcy prediction; financial statement analysis

1. Introduction

A fundamental and general problem in improving statistical models to predict future actions is the selection and integration of data within the development process. The application case of company-related bankruptcy prediction has constantly evolved, starting with simple regression models via multivariate methods up to today's approaches of artificial intelligence, i.e., the use of machine learning. The interest in this application is omnipresent due to the timeless character of the financial valuation of companies. Changes in motivation occur both on the part of the research as well as practice event-related, e.g., financial crises and industry-related issues. The striving of science and practice for ever-better predictions for the assessment of risks for stakeholders is constantly reinforced by expectations of new, improved solutions created by technological progress [1]. While structured data, such as those found in balance sheets, offer the advantage of being easily usable by machines and statistics, they are limited by their inflexibility in explaining non-trivial relationships in a company [2]. At the same time, we are now experiencing a flood of unstructured data that can help to form a financial picture of a company through the most diverse observations of information [3]. For analysts, the focus of interest is, on the one hand, the external presentation of a company to the public, but also the public's perception of the company. Classic news media and social networks play a major role in this and are the source of a large amount of unstructured data that need to be validated and classified [4]. While unstructured data were, for a long time, simply not technically manageable and not considered due to storage and computing power, it is now the focus of expectations for improved solutions [5,6]. Since the use of unstructured data requires preprocessing, the question inevitably arises for both academia and practice to assess approaches in terms of their suitability for extracting features that can be used for model development. We, therefore, define the research question for this paper:

RQ: How should text mining approaches be designed to improve the predictive accuracy of corporate failure models?

Authors: Tobias Nießner, Stefan Nießner and Matthias Schumann

Outlet: *Information 14 (2023) 4*.
https://doi.org/10.3390/info14040215

Abstract: How can useful information be extracted from unstructured data that can contribute to a better prediction of corporate bankruptcies? In this research, we examine a dataset of 2,163,147 financial statements of German companies that are triple classified, i.e., solvent, financially distressed, and bankrupt. By classifying text mining features in terms of granularity and linguistic level of analysis, we show results for the potentials and limitations of approaches developed in this way. With our study, we thus give a first approach to evaluate and classify the likelihood of success of text mining approaches for extracting features that enhance the training database of AI-based solutions and improve the models developed in this way.

Our results are an indication that the adaptation of additional information sources for the financial evaluation of companies is indeed worthwhile, but innovative approaches adapted to the application context should be used instead of generic approaches.

Keywords: Text Mining, Machine Learning, Bankruptcy Prediction, Financial Statement Analysis

Introduction

A fundamental problem in improving statistical models to predict future actions is the use and integration of data within the development process. The application case of company-related bankruptcy prediction has constantly evolved, starting with simple regression models via multivariate methods up to today's approaches of artificial intelligence, i.e., the use of machine learning. The interest in this application is omnipresent due to the timeless character of the financial valuation of companies. Changes in motivation occur both on the part of the research, as well as practice event-related, e.g., financial crises and industry-related issues. The striving of science and practice for ever better predictions for the assessment of risks for stakeholders is constantly reinforced by expectations of new, improved solutions created by technological progress (Jones 2017). While structured data, such as those found in balance sheets, offer the advantage of being easily usable by machines and statistics, they are limited by their inflexibility in explaining non-trivial relationships in a company (Kloptchenko et al. 2004a). At the same time, we are now experiencing a flood of unstructured data that can help to form a financial picture of a company through the most diverse observations of information (Nassirtoussi et al. 2014). For analysts, the focus of interest is on the one hand the external presentation of a company to the public, but also the public's perception of the company. Classic news media, but also social networks, play a major role in this and are the source of a large amount of unstructured data that needs to be validated and classified (Schumaker et al. 2012). While unstructured data was for a long time simply not technically manageable and not considered due to storage and computing power, it is now the focus of expectations for improved solutions (Kloptchenko et al. 2004b). Since the use of unstructured data requires preprocessing, the question inevitably arises for both academia and practice to assess approaches in terms of their suitability for extracting features that can be used for model development. We, therefore define the research question for this paper:

RQ: How should text mining approaches be designed to improve the predictive accuracy of corporate failure models?

Within this study, we follow a classical data mining process model, i.e., Cross-industry Standard Process for Data Mining (CRISP-DM) (Wirth and Hipp 2000), to present our approach in a standardized and structured way. While we defined and motivated the goal of our project in the first step within the introduction, a review of related research dealing with the use of data and development of statistical models in the context of machine learning-based bankruptcy prediction models follows in the second chapter. Building on this, the presentation of the data used, as well as a description of the further processing based on the insights gained from an analysis of it, follows. In the following, we describe the procedure in the model development before moving on to an evaluation of it. This is followed by a discussion of the results, which highlights contributions to research and practice and addresses limitations and approaches for future research before a conclusion is drawn.

Related Research

Research into the prediction of corporate failure or bankruptcy initially showed a clear trend toward the analysis of financial ratios (Altman 1968). While these models were based on stochastic methods such as multivariate discriminant analysis in their early days (Altman 1968; Altman et al. 1977; Ohlson 1980), the use of machine learning algorithms is playing an increasingly important role in current research (Kirkos 2015), driven by the availability of larger amounts of data and computing power. These enable more sophisticated analyses to be performed in a short period, and the capture of complexity in developing a predictive model is thereby transferred from humans to machines. While companies have to prepare their financial statements according to different paradigms depending on their size and origin, the textual components of reports remained unnoticed by such models for a long time (Lohmann and Ohliger 2020). Due to the constant demand from practice, based on ever-increasing risks due to potential bankruptcies, the interest in the evaluation of these data as well as stirring (Veganzones and Severin 2021). While the analysis of financial ratios, which are calculated based on a company's balance sheet to evaluate the company's performance in relative terms, is limited by the presentation of only past reliable events, the analysis of qualitative data allows a look at assumptions and expectations of the company, as well as an explanation of the current

financial situation. It should be noted here that the analysis of textual data and insofar as the generation of suitable features can proceed in different dimensions. Up to now, a large number of publications exist in the area of the analysis of the usability of document-related parameters, such as the sentiment (Caserio et al. 2020) or the readability of financial statements (Ajina et al. 2016; Le Maux and Smaili 2021). If one reflects, particularly on the analysis of readability, that a large number of such readability indices have been developed based on newspaper and book publications (Bjornsson 1983), the question arises as to whether these can be applied to financial statements in a meaningful way at all. It should be noted that these documents are in part quite standardized in their choice of words and preparation, which makes the application of these metrics appear questionable from a linguistic point of view as well (Loughran and McDonald 2011). The study by Loughran (2014) stands out in this respect, as it examined a wide variety of metrics for assessing the readability of financial disclosures and came to the conclusion that log(file size) is the most suitable for classifying them. Considering this result from a practical point of view, the question arises to what extent the size of the company and, in particular, the associated disclosure requirements make this result appear useless about the failure prognosis of German companies. Furthermore, the selection of graphics and the associated use of storage appears to be largely subject to a random principle in the selection of the file format and software used to create it. From our point of view, the question of a possible apparent causality arises, since from a comparison of metrics in the literature review it is not evident whether an explainable correlation based on the cause-effect principle exists. Regarding sentiment analysis, on the other hand, there are research contributions that address this problem and develop their metrics to classify the company-specific choice of words accordingly, since it is known that misinterpretations can otherwise occur (Bannier et al. 2019). Furthermore, a taxonomy for the characterization of text mining features exists which allows differentiating the approaches described here concerning, e.g., their degree of linguistic analysis or their granularity (Wambsganss et al. 2021).

In addition, initial results suggest that there is a high degree of standardization, particularly for German financial statements (Lohmann and Ohliger 2020). It can be assumed that companies reuse timeless text components such as frequently used phrases that are common in the corporate context. This would explain one approach of the lack of suitability of document-related text features, as they explain too much irrelevant information in their measurements. However, from the discussion regarding the suitability of sentiment analysis for bankruptcy prediction also arose contributions that consider sentiment analysis on a finer granular level and apply it only to specific components or patterns in the text to semantically assess very concrete developments and thus also provide better explanations for its use from a cause-effect perspective.

Beyond the pure consideration of the approaches for the utilization of textual data, the differentiated consideration of information according to the industry affiliation of companies is playing an increasingly important role, since especially personnel developments, but also possible risks for the financial development of companies, as current effects of various crises show, have a strong influence on the strategic orientation and thus consequently on the future situation of companies (Jones 2017).

Especially the classification of external information from third parties on companies, but also the overall economic situation is therefore of research interest to be able to explain and evaluate developments in more detail, since these, if at all non-trivial, can not be extracted from internal company publications (Zhao et al. 2020). While in the current research, justified by the successful use in comparable use cases, Ensemble Learning is given a preference in algorithm selection (Veganzones and Severin 2021). To the best of our knowledge and belief, no study exists that examines different dimensions of text mining features and evaluates their usefulness in comparison to the classical financial ratio-based corporate bankruptcy or failure prediction.

Data Presentation, Understanding, and Preparation

To answer the research question, we present below the methodology based on a standardized data mining project according to CRISP-DM. Following Figure 15, we present the steps separately starting with data selection. First, we should say

that we used two basic datasets, one being a dataset of 2,163,147 German Financial Statements from 2017-2021, which are in XML format and contain additional meta-information, and company-related information exported from Bureau van Dijk's Amadeus database (Bureau van Dijk 2021).

In the 1. step, the meta-data was extracted and the text was processed separately so that various parameters considered in the current research could be extracted and differentiated by linguistic analysis level and granularity (see Table 19) (Nießner et al. 2022b; Nießner et al. 2022a). Sentiment analysis was approached in two different ways. On the one hand, a general German sentiment dictionary (Remus et al. 2010) was used, and on the other hand, a context-related one (Bannier et al. 2019). Furthermore, using a translated version of a hedging dictionary, a score regarding the word choice of obfuscating terms was evaluated (Humpherys 2009), as well as an analysis of the extent to which the text authors used passive constructions in their sentences (Nießner et al. 2022a). For POS tagging, the spaCy Python package was used with the help of a German NLP model based on the TIGER corpus (Brants et al. 2004). If liability items in a balance sheet exceed asset items, a "deficit not covered by equity" must be reported in the balance sheet, which is included in our analysis as a Boolean Feature if it was reported in the text. Concerning the meta-information on the financial statements, we recorded whether and how long a financial statement was filed late, as well as the quarter in which it was received.

In the 2. step, various financial ratios (see Table 17) were calculated from balance sheet items, as these form the basis for analyzing and evaluating the extent to which the addition of additional ratios offers added value. In addition, supplementary information about the respective companies was collected, which dealt with the organizational and personnel level (Jones 2017; Brédart et al. 2021) and the competitive level (Nießner et al. 2022c). In this regard, we used additional demographic and region-specific data sets for the calculation (Statista 2022; Zensus 2022).

In the 3. step, the data were cleaned. In a first analysis, it was found that among the 2,163,147 financial statements, 1,238,244 were exact duplicates in terms of

textual content, which can be attributed to the fact that the smallest companies, whose financial statements were also available, did not insist on publishing textual information or that boilerplates exist in text form. These data are not relevant to our study as they do not provide any added value for the evaluation of text mining approaches.

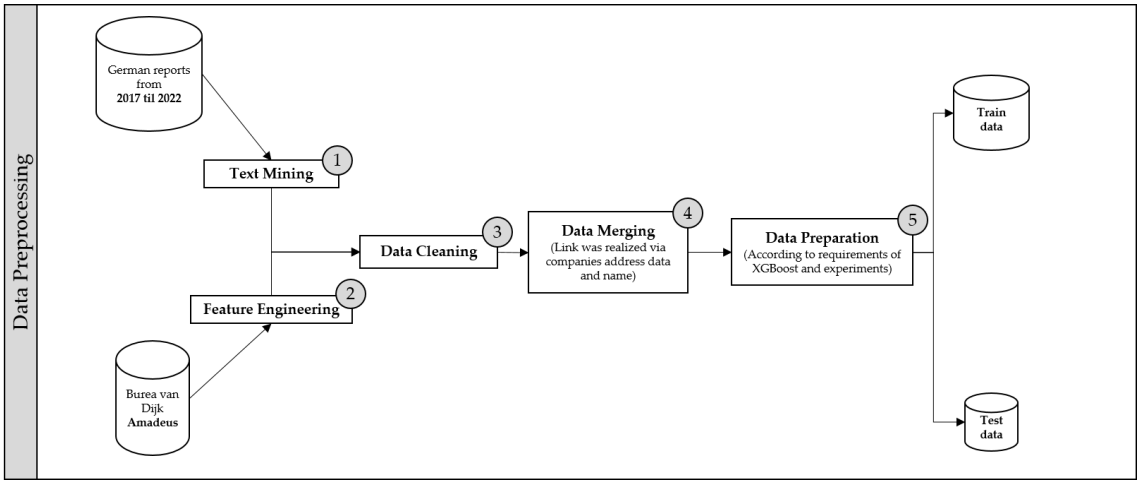


Figure 15. Summary of the methodical procedure and the associated steps

In the 4. step, the different data sets had to be linked with each other. For this purpose, the company name and address data were used in both data sets, so that a merge could be ensured by comparing this information. A match was achieved for 855,559 financial statements, reducing the data volume in the following.

In step 5., the features extracted in steps 1. and 2. were subjected to appropriate processing, such as normalization and relativization concerning the number of words or sentences of the respective financial statements. In particular, since the cardinality of the features was small, categorical features were preprocessed using one-hot encoding. This included features such as, e.g., company size and industry affiliation, the latter being classified according to NACE Rev. 2, i.e., the classification of economic activities in the European Community (eurostat 2022). For labeling the data with respect to the use case, we decided to adopt the financial situation decisions made by Amadeus. While the literature often differentiates between a binary and a non-binary decision problem, we decided to distinguish only between solvent and financially distressed or bankrupt companies due to the impact of the classification on a user of the model, since

the margin from defaulting to bankrupt did not allow a differentiation based on the financial ratios. In the following, we thus consider a binary decision problem between solvent and a merge of defaulting and bankrupt companies. An overview of the features extracted and used for model development is shown in Table 17.

Accounting-based features (F_A)	Text-related features (F_T)		Company features (F_C)
Asset coverage ratio	Publication quarter^{M}	Count_ES_ratio^{S}	Count_branch_offices
Capital Debt ratio	Text length ratio^{D}	Count_ES_Pos_ratio^{S}	Count_companies_group
Debt structure	Time_difference_publication^{M}	Count_ES_Neg_ratio^{S}	Count_employees
Debt-to-Equity ratio			Count_shareholders
EBITDA	Count_Comma^{C}	All_Sentence_ratio^{S}	Count_subsidiaries
Equity multiplier	Count_Excalamation_Marks^{C}	Passive_Sentence_ratio^{S}	
Equity ratio	Count_Question_Marks^{C}		Count_CEO_change
Leverage ratio	Count_Bankruptcy_words^{W}	Hedging Score^{D}	Count_CEO_last_10_years
Long-term liabilities	Count_Chances_words^{W}		
Net loss for the year	Count_Opportunities_words^{W}	SentiWS_Hits_ratio^{D}	Company_age
Profit margin	Count_Project_words^{W}	SentiWS_POS_Hits_ratio^{D}	Company_size
Return on Assets ratio	Count_Research_words^{W}	SentiWS_NEG_Hits_ratio^{D}	NACE_identification
Return on Capital Employed ratio	Count_Risk_words^{W}	SentiWS_POS_Score_ratio^{D}	
Return on Equity ratio	Bool_Deficit^{D}	SentiWS_NEG_Score_ratio^{D}	Competitor_density
Short-term debt ratio		SentiWS_Score^{D}	Population_density
Short-term liabilities	Count_ADJ_ratio^{W}		
Working capital ratio	Count_ADP_ratio^{W}	BPW_UNC_ratio^{W}	
	Count_ADV_ratio^{W}	BPW_POS_ratio^{W}	
	Count_AUXVERB_ratio^{W}	BPW_NEG_ratio^{W}	
	Count_CONJ_ratio^{W}		
	Count_DET_ratio^{W}		
	Count_NOUN_ratio^{W}		
	Count_NUMERAL_ratio		
	Count_PROPNOUN_ratio^{W}		
	Count_VERB_ratio^{W}		

Legend: Granularity of Text Features {^{C}: Character, ^{W}:Word, ^{S}:Sentence, ^{D}:Document, ^{M}:Meta}

Table 17. Feature sets for XGBoostClassifier model development

Modeling

Since gradient boosting algorithms are currently attracting great interest among researchers and offer many advantages, such as training time and accuracy, over more classical algorithms due to their structure, we decided to use an XGBoostClassifier for our experiments (Kirkos 2015; Jones 2017). For parameter selection, a GridSearchCV was performed, determining the learning rate (eta),

the regularization parameter (gamma), and the maximum depth of the individual trees, based on a 3-fold cross-validation. Furthermore, the objective function "rank:pairwise" was determined using GridSearchCV (scikit-learn 2022). Thus, the binary decision problem is transferred to a ranking, which generates the model based on a pairwise comparison of all instances within the training dataset. In this respect, contrary to the more common approaches of supervised learning, i.e., regression or classification, the decision is made about placement in a ranking, as it is known from search engines. Thus, adapted to the problem, it can be simplified to say that a model is trained that uses a list ordered according to financial performance to make decisions. Given the use of features from different origins (see Table 17), in addition to considering feature importance concerning text mining approaches, we decided to train different models based on differentiated training datasets in each case. We thus show in the following comparison for the suitability to predict financial difficulties of companies. We consequently train a base model with financial balance sheet ratios and compare this with extensions to textual, company-related, and, an extension of both.

Evaluation

A stratified train-test-split was used to evaluate the models, with a ratio of 4:1. Basic measures for the evaluation of binary decision problems are used, with a balanced accuracy calculated due to the imbalance of the data set in terms of the occurrence of companies with financial problems. The evaluation of the models shows that the variation of the training data basis has an impact on the performance. While the classical base model with a balanced accuracy of 0.71 already achieves a value that is clearly above a random assignment, it is shown that the addition of textual features improves the performance, although this does not happen to the same extent as with an addition of company-related features.

	F_A	$F_A + F_T$	$F_A + F_C$	$F_A + F_T + F_C$
Precision	0.9953	0.9957	0.9964	0.9968
Recall	0.6831	0.7007	0.7200	0.7441
Balanced Accuracy	0.7112	0.7262	0.7541	0.7755
F1-Score	0.8102	0.8225	0.8360	0.8521
False-Positive-Rate	0.2608	0.2482	0.2119	0.1930
False-Negative-Rate	0.3169	0.2993	0.2800	0.2559

Table 18. Evaluation of the differently trained XGBoost models

Furthermore, the addition of both feature groups distinguished in this study results in an additional improvement of both differentiated additions. Overall, the best model was able to correctly identify about 81% of the reports of companies with financial problems and about 74% of the solvent companies.

Discussion and Implications

In the following, we discuss aspects of the used text features and their usefulness in the context of the use case of financial failure prediction of companies based on financial statements.

	Feature Expressions	
Granularity Level	**Character**	• Number of punctuation marks used in a report [N]
	Word	• Number of specific words within a report [L] • POS-Tagging [Sy]
	Sentence	• Evidential Strategies [Sy] • Passive Voice [Sy]
	Document	• Hedging Score [L, Se] • Sentiment Score [L, Se] (SentiWS, BPW) • Descriptive Information [M]
Legend: Linguistic Analysis Level {[N]: Non-linguistic; [L]: Lexical; [Sy]: Syntactic; [Se]: Semantic, [M]: Meta }		

Table 19. Classification of extracted text mining features

The text features used in the model development can be analyzed according to the level of lexical analysis they exhibit, but also according to the granularity they use within the text (see Table 19). Contrary to the popularity of readability indices, we did not analyze them because, as argued in the review of the literature, there is no clear deductive evidence for their usefulness and actual correlation in the financial context (see Related Research). In line with the granularity, we have not considered the meta-level as an independent level, as such information can be directly related to the whole document and a basis for a finer granular diversification is missing. Conversely, it cannot be logically concluded that publication and author information are the results of natural language processing (Fromm et al. 2019; Wambsganss et al. 2021).

Such information is obtained by linking additional data, but not directly from the existing text. Reflecting on the different levels of granularity, the combinatorics, but also the possibility to use external data sources, suggests that a positive correlation between the level of granularity and the number of extractable metrics can be assumed. The linguistic level of the analysis, on the other hand, provides information about the content character of the feature from the perspective of the language, which also allows the approaches to be adapted to other languages according to different grammatical peculiarities. It should be noted, however, that no validation of actual deductive evidence based solely on the empirical observation is possible in this study and therefore only correlations and no causalities are analyzed. A look at the feature importance calculated with the F-Score, which measures the number of occurrences of a feature within the trees generated by the XGBoost algorithm, shows indications that the text mining approaches used can be limited in terms of their usefulness. Figure 16 shows that certain financial features, i.e., those that relate to existing debts of the company, play a very important role within the model, but the textual features also have a high value in comparison.

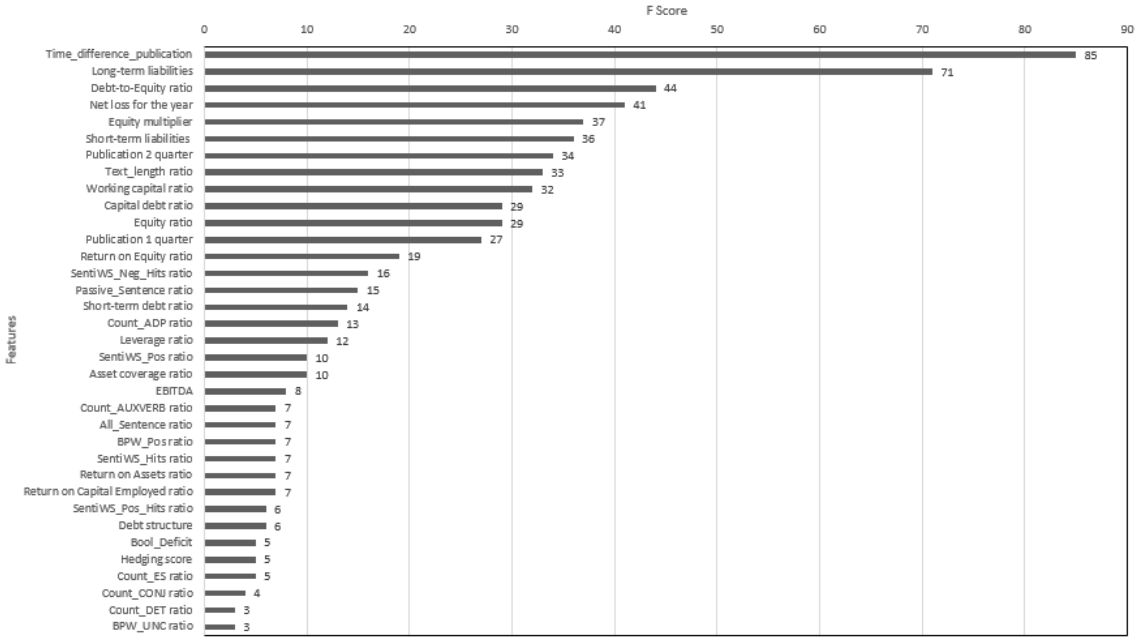


Figure 16. Feature Importance based on supplemented textual feature set ($F_A + F_T$)

Contributions to Research and Practice

A close look at feature importance suggests various conclusions for research on the inclusion of textual features in the prediction of corporate financial failure. While sentiment analysis of texts is often realized in research via dictionary-based approaches, our evaluation reveals a weakness in the definition of these. Both German dictionaries, such as SentiWS (scaled polarities between 0 and 1 are available) and the context-adapted BPW (words are only classified into 3 classes, e.g., uncertainty, positive, negative) has the disadvantage that no methodology exists for their application. Therefore, a list of words can be used to form and analyze different metrics according to the creativity of the user. Within this study, we, therefore, did not only calculate a classical score over the weighted words but also considered the ratio of the hits to the total number of words in the text. The consideration of the feature importance suggests that it is much more relevant for the assessment of the financial situation in which a ratio of positive and especially negative words are used within the text than the calculation of a value for a polarity that does not appear in the top 35 of the most prominent features (out of a total of 61 features). Reflecting the different reporting obligations and voluntary options of companies, it is also of interest that the min-max-normalized length of the reports has a high weighting, as this suggests that the development of completeness measures for financial statements offers an interesting starting point for future research. Furthermore, it is of interest that ratios describing sentence length, the number of conjunctions used, as well as passive constructions have a comparatively high influence. Here we have to consider to what extent such constructions may express an abstract construct like that of readability anyway. We further see that besides the semantic level of analysis, the syntactic level, represented by more complex constructions (Count_ES ratio and Bool_Deficit), seems to be of greater interest compared to features based purely on lexical analysis. Nevertheless, it is also shown that meta-information such as the difference between the described year and the actual publication date outperforms, for example, even financial metrics. For practical purposes, we could show using the results of this study that the effort of analyzing textual as well as other data sources is worthwhile for the improvement

of forecast models. However, given the feature's importance, it must also be reflected that the concept of forecasting based on financial statements also has a weakness in that it depends on companies publishing their financial statements by the deadlines, which is at least doubtful in our view.

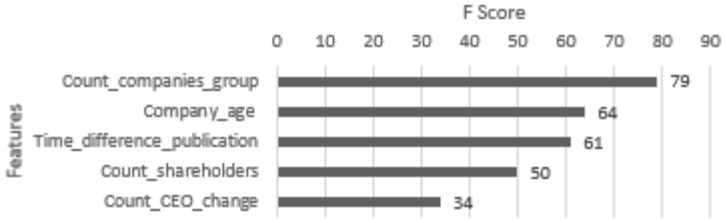


Figure 17. Feature Importance based on all available data($F_A + F_T + F_C$)

In addition, an examination of the top 5 features (see Figure 17) of the overall model ($F_A + F_T + F_C$) based on all available data shows that, on the one hand, there is no financial indicator and, on the other hand, there is also only one feature that relates to the meta-information on the report. The other features describe the structure and changes within a company and suggest that a reassessment of the financial situation of a company can not only be based on a new financial statement, but can also be triggered by changes such as acquisitions of other companies, closures of existing locations, but also changes in the management board.

Limitations and Future Research Opportunities

We are aware of the fact that certainly, other machine learning approaches exist that have the potential to provide higher prediction accuracy. The focus of the conducted study, however, is to evaluate the suitability of additional data sources with a specific view on text mining features to meaningfully extend the training database of AI-based approaches in corporate failure prediction. Therefore, and due to the strong specific data-oriented dependency of the trained models, we refrain from an evaluation in practice scenario as current german data is not accessible via a public API. Another reason to be mentioned here is the lack of comparability of studies, as to our knowledge, no german benchmark data set exists that allows the evaluation of AI-based approaches. It is worth mentioning that this problem also exists for English-language reports.

Finally, a further limitation arises from the distribution of the data for the respective companies. While an approach motivated by the literature was chosen for the consideration of financial statements, which consider a financial statement as a single data point, it can be questioned whether this modeling does justice to the analysis of companies. First of all, it is clear from this decision that the variables under consideration are not analyzed in a time-related manner and are thus limited by the fact that only past developments can be represented by a balance sheet. However, if one considers, for example, linguistic constructs such as the sentiment of a document, it would be presumptuous to argue that this is not also dependent on factors of the author and that a comparison of these based on individual financial statements of different companies cannot be objective. In turn, the focus on the statistical comparison of financial statements, rather than companies, leads to the problem that differences in the development and presentation of companies may not be adequately reflected by a forecasting model. Due to the lack of suitable consecutive financial statements, the study carried out is not able to model changes in a company for itself and thus make them usable for comparison. Reflecting on the data collection process in the literature on the topic, it is noticeable that few studies focus on the comparison of company profiles developed from financial statements and thus view the topic from a different level of abstraction. It should be noted, however, that the problem of acquiring suitable data is not the sole responsibility of researchers, but also requires the cooperation of practitioners. This would also enable aspects for future research, such as industry-specific model building, for which there are already various motivating research results.

Here we tie up with our consideration to the past research and must mention however for future research also that there are further possibilities of the classification of annual financial statements into AI-based solutions, i.e., a contrary company-based view on those report data over a period of time as mentioned above.

Conclusion

Researchers are developing a variety of AI-based approaches to address the use case of predicting corporate bankruptcies, but often face the same issues, i.e., data accessibility, industry- and region-specific differences, as well as the comparability of approaches to evaluate results in the overall context. When considering the data, a distinction must also be made between internal and external information from third parties about the company to identify independent objective parameters that can be used to characterize the financial situation of companies. As with each specific use case, we have shown the extent to which an AI-based approach can be improved using additional data sources, and have narrowed down which manifestations of text mining approaches, which have recently received much interest, have the potential for this.

Our results clearly show that the classical training process of an AI-based approach for corporate failure or bankruptcy prediction, which is purely based on financial ratios, can benefit from various new perspectives. In addition to the well-known semantic text mining approaches, we also showed that the contextual extraction of features from reports can provide new incentives for forecasting developments. Interestingly, we found that not only these approaches, which are often used in research, influenced our model, but also syntactic and meta-textual parameters can be attributed to a corresponding benefit. Considering the research question defined at the beginning, we can conclude that even though the context-specific sentiment dictionary performed worse than the general dictionary, we can do without features that represent sentiment, even though the relative number of polarized hits (negative) seems to add more value to the model than a classical score. Nevertheless, we were able to show that research from exploratively developed approaches, which are extracted in context, can keep up with the well-known tools of text mining, such as sentiment analysis, and even have a greater value within the failure prediction of companies. The text analysis also showed that meta-information regarding the financial statements as well as the addition of external information about the company should play a major role in future research. We need to work on further analyzing syntactic constructions in terms of semantic, as well as contextual information to collect data from the

evaluation of such patterns, but also meta-information, which will allow us to further improve prediction models.

C Contributions

The studies presented previously aim to improve the understanding in the area of data acquisition, as well as data preprocessing up to the development of AI-based models for corporate failure prediction. In particular, the different substeps of the CRISP-DM are reflected, starting from the *Business Understanding* phase, through *Data Understanding* and *Data Preparation*, to *Modeling* and *Evaluation*.

In the following, C.1 provides an aggregated discussion of the conducted studies, focusing on a summarizing reflection of the contents with respect to classification, both in current research and the development process of a bankruptcy prediction model. Subsequently, in C.2 the results are presented in relation to the research questions defined at the beginning of the dissertation. Finally, in C.3, based on limitations theoretical and practical implications are reflected.

1 Discussion

Reflecting on the phases of the CRISP-DM, it should first be noted, starting with the phase of business understanding, that both, the literature review conducted in Section A and the other studies conducted indicate that the specific goal definition for an AI-based solution is problematic. While a majority of research articles comprehensively declare their success by improving metrics, e.g., precision, accuracy, F1-score, there is a lack of comparability of those individual results (cf. A.2). There is currently no metric to meaningfully distinguish and evaluate AI-based approaches from one another, as the data basis often varies and the data selection processes are usually imprecise. Consequently, there is a problem regarding the reproducibility of results since data is often not public. However, this identified weakness is not solely attributable to researchers, but can also be explained by the limited availability of data in practice. While it is comparatively easy to conduct research on a selection of 10K reports of American companies, as these are publicly available, the opposite picture emerges often as financial statements of other countries are to be considered. It should therefore be noted that real improvements cannot be validated on a comprehensible base

across different cultures and countries. The vast majority of studies themselves use a comparison with native base models in order to validate their assumptions. This shortcoming, pointed out by the analysis of the literature (see A.2), can be addressed by using one or more quality criteria for the evaluation of decisions as higher transparency can positively influence the trustworthiness of AI-based decision-making (Ashoori and Weisz 2019). In particular, this supports the comparability of different approaches. For example, the comparison to human decisions as status quo could be better suited to classify the performance but also error rates of an AI-based model than a purely technical view. Often this results in a possibly impossible quest for perfect accuracy of AI-based solutions. These, possibly exaggerated, hopes are reflected, for example, by superlatives such as "AI will change everything" in connection with the use of AI, which are propagated by large companies (IBM 2021). On the one hand, new approaches are constantly being examined by the research community, which deal with ever new data sources and the classification of their usefulness for the improvement of models, on the other hand, the term "improvement" has not yet been sufficiently discussed scientifically as reflected by means of the comparability. This can be explained in particular by, as noted in the literature review (see A.2), a lack of communication with practitioners. The evaluation of models is mostly based on common metrics of AI-based approaches for binary decision problems like Accuracy (and the balanced version), Precision, Recall or the F1 score, where purely technical interpretations of the metrics are in the foreground instead of those that also reflect a practical benefit. For example, the analysis of errors of a model could be shifted more into the focus of the evaluation, since from a practical point of view in a non-binary decision problem an error is less weighty if it is in a neighboring class (financially distressed, when true prediction would be bankrupt) than further away from it (solvent, when true prediction would be bankrupt). Science and practice must work together to define a status quo that makes it possible to compare the currently unstructured research approaches via universal measures. This prospect is currently far away from reflecting the state of art, but holds the greatest potential for making results from all perspectives perceivable as useful. Considering the standardization of the labeling of textual components of a financial statement, new scenarios of adaptation arise from the approaches shown in Part B. Linking the approach of Lohmann and Ohliger

(2020) with studies III and IV, not only analyses of individual specific report sections are conceivable, but beyond that also their development in terms of textual characteristics, e.g., linguistic, content, structure (internally and within the report as a whole).

In summary, the research processes are, regarding the lack of specific requirements and thus a goal, in a non-finite iteration process according to CRISP-DM. The taxonomy presented in Study I of this cumulative dissertation can be helpful especially with regard to the communication of researchers and analysts to classify approaches and thus shape status quos with regard to diverse developments of AI-based approaches. Especially with regard to the definition of benchmarks, it can be very useful to classify and thus differentiate between different approaches with regard to their evaluation.

In the context of the Data Understanding and Data Preparation phases, Studies II-VI are discussed next. It has to be reflected that these phases are crucial for the success of a model, because the work, which takes place within these process steps, has a significant influence on the data and therefore the later model development, i.e., through a training process in case of machine learning. The phase of deployment is not discussed here, since a continuous data flow would be necessary to implement a dynamic reacting AI-based prototype, for which publicly available interfaces would have to be created. Since the extension of the training dataset of AI-based models based on the analysis of textual financial statement data is a central part of this dissertation, a reflection of this based on a taxonomy of text mining features is done first. The characteristics of text mining features can, according to Wambsganss et al. (2021), generally be divided into different dimensions, i.e., *dimensionality*, *representation*, *linguistic analysis level*, *granularity level,* and *information source* of a feature. Regarding the *dimensionality* of features, an earlier publication by Fromm et al. (2019) argues that this dimension seems trivial to most researchers. However, it should be noted at this point that the distinction in terms of *dimensionality* does not necessarily represent added value from a technical perspective. Looking at the use of text mining features for the development of AI-based models, it can be

seen that multi-dimensional word features, e.g., *bag-of-words*, *bag-of-n-grams,* and *bag-of-POS-grams,* are technically used element-wise for training. Nevertheless, the distinction at the causal level raises the question of the extent to which a parameter of a text is to be represented by one or more values. The sentiment as an example could be represented in the form of a scalar as the sum of individual word evaluations, whereas in the form of a vector, the possibility arises to use further characteristics, e.g., frequencies of individual polarizations. The distinction as to whether a feature is represented as a scalar or a vector is thus of no essential importance for the training process. Also, considering the level of granularity, it should be questioned to what extent the characteristic *beyond document* can be derived from the consideration of NLP features or, as suggested by the example of Wambsganss et al. (2021), via call data of a text on the web, by linking it to meta-data. Meta-data is not natively related to NLP. For this reason, we have deliberately refrained from speaking of a *text feature set* in Study VI and have chosen the term *text related* to assign the inclusion of meta-information. Therefore, we suggest that a granularity level called *corpus* based on the reflection of the studies performed and the reference to feature extraction based on NLP needs to be added. This will allow future studies to go beyond the analysis of a single document to form metrics based on the analysis of multiple texts.

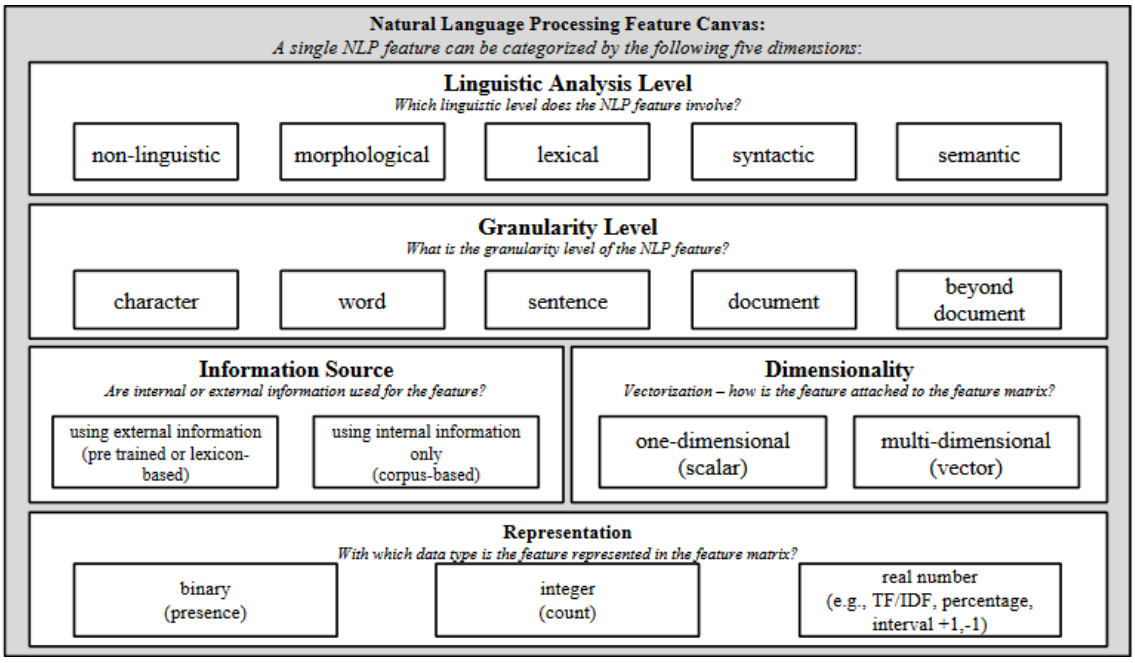


Figure 18. Taxonomy of NLP-based features from Wambsganss et al. (2021)

Nevertheless, the taxonomy approach developed in the context of emerging research is suitable to classify and reflect the conducted text mining approaches presented in this dissertation.

Before discussing the specifics of text mining approaches in the context of identifying financially distressed companies, it should be noted that the studies conducted differ according to the data basis and thus the options for extracting information. While studies III and IV are based on the analysis of consecutive financial statements and offer the possibility to analyze the granularity level *corpus* with regard to changes in more detail, studies II and VI deal with approaches that also use such classified key figures, but do not integrate the change in a company's reporting over time in comparison (see Figure 16). In particular, the studies have made it clear that the field of text mining cannot be applied to the area of detecting financially distressed companies by only adapting known methodologies. Established methods such as sentiment analysis can be reinterpreted in this context. Novel approaches can be distinguished from the classical consideration of entire documents with regard to the text excerpts to be analyzed. Furthermore, linking this finely granular view with levels of linguistic analysis also provides further opportunities to evaluate textual patterns that can improve the prediction success rate of AI-based models in the area of enterprise failure prediction. An advantage resulting from the more specific definition of such text features can also be seen in the context of explainable AI. There is the possibility that a detailed consideration of the existence of causal relationships can be enabled, as such approaches analyze text in a more elementary way. Furthermore, exploratory approaches resulting from theories of linguistics could have the potential to enable new approaches to the interpretation of expectations and to provide added value through the semantic construction of defined patterns.

Considering the steps of data understanding from capturing and describing to discovering the data, it was shown that textual data opens up various possibilities for analysis and that new ways have to be found to enable corresponding improvements. The model case considered here shows an indication that the simple use of known procedures is not sufficient to extract specific information

and, in this respect, to extract new knowledge from the analysis of the reports. This was equally evident within the studies conducted when using dictionaries to evaluate the sentiment of a text. Since there is no clear methodology for the application of the dictionaries, various application possibilities can be constructed whose output cannot be reliably assigned to a causality. It is also questionable in view of the results regarding the granularity of the features, whether such *document-related* applications of dictionaries could not be replaced by more fine-grained applications. Thus, more specific words or even topics fitting the application context could be further evaluated syntactically as well as semantically in order to be able to deal with their use in a more comprehensible way. Furthermore, when analyzing consecutive financial statements, in addition to analyzing existing words or topics in such approaches, it is also possible to discuss whether the absence of either or both also has a cause that is associated with the financial situation of a company. Furthermore, this consideration illustrates another aspect with regard to the problem of the application of dictionaries, which often goes hand in hand with the lack of specification of a context.

The data preparation phase came to the fore in Studies V and VI. The results of both studies indicate that the predictive performance of AI-based models is positively affected by expanding the training database. However, based on Study II, Study V, and Study VI, it could be shown that the construction of new features based on different data sources, both from the financial statement itself, but also beyond, should be considered, a successful improvement of the prediction accuracy.

In the modeling phase, it quickly became clear that the XGBoostClassifier algorithm cannot train a model based on the data that can make a fine-grained distinction between companies in default and bankrupt companies. This is not surprising, considering that there is a certain period of time that is not necessarily represented in the data in which a company may recovers economically or files for bankruptcy. While the financial condition of a company must be considered critical in both cases, the question here is whether and in what form companies communicate ways of avoiding bankruptcy, irrespective of the development of the model. Based on the results of the studies conducted in this dissertation, starting points can be identified, but a generally valid distinction between the statuses of

bankruptcy and inability to pay based on the analysis of financial statements alone is not possible. The model developed in Study VI was unable to make a distinction in this respect. An interesting aspect to consider here would be to what extent the analysis can be seen as part of a company profile, which is continuously updated based on different developments, whether external or internal and thus enables updated assessments at shorter intervals in order to be able to satisfy measures and reactions instantaneously.

The final study VI, which is positioned in the evaluation phase, shows, that we can achieve results for a binary decision problem that are far above random decisions and with a False-Positive-Rate (FPR) of approx. 19.3% and False-Negative-Rate (FNR) of approx. 25.6% are able to deliver results that can probably be used in practice. An evaluation of this usability assumption has yet to be conducted. It is clear that the confidence in AI-based models by the broad mass is currently a subject of research in various application contexts and this can probably also be transferred to this, but it must also be questioned here what accuracy a human analysis is able to provide in comparison. The decision of an AI-based model in the context considered here must always be seen from two sides, from the point of view of the classified company and from the point of view of an analyst. There are advantages and disadvantages on both sides that arise from a financial assessment, such as on the one hand the availability of financial resources from third parties, but on the other hand also the protection of these resources in the economic exchange with companies that are no longer financially trustworthy.

2 Conclusion

A literature review was conducted to assess the current state of research for the prediction of corporate failure or bankruptcy. While the research needs of this dissertation were derived from the review, cues can be derived from the literature review that indicate that current research on the adaption of AI-based solutions currently still in its proverbial infancy.

Research Area: Assessment of Corporate Failure Approaches	
RQ. 1	How should the current state of research on the use of AI in corporate failure prediction by means of financial statements be assessed?

A large proportion of the studies identified are explorative in nature. Classification of these using the taxonomy from Study I shows that they are characterized by the fact that new data is always available for analysis, and thus more and more studies are using supplementary data for analysis and consider existing approaches based on financial ratios from balance sheets to be "researched". However, it must be borne in mind that while such research results can provide indications for decisions in the further development of forecasting models, valid causal explanations of correlations have so far often been disregarded and only correlations have been considered due to the limited nature of data. Reflecting the developments in research, we are at a point where it is possible to analyze ever larger amounts of data, but the publicly accessible availability of German data in the context of usability for the development of AI-based approaches is becoming significantly more difficult. Whereas until a few years ago it was still possible to acquire large quantities of annual financial statements by machine via the Bundesanzeiger homepage, the new design presents hurdles that make such acquisitions impossible, i.e., by using randomized IDs for the reports and captcha queries in the content management logic. At this point, research and practice must ask themselves to what extent the interface between knowledge, evaluation and subsequently also deployment can be realized. In summary, a status quo with regard to prediction models, as suggested in the discussion, would be desirable.

Research Area: Knowledge Extraction	
RQ. 2	How can quantitative financial statement analysis be supported by using textual data with respect to German companies failure prediction?

The studies II-V have shown to what extent time-dependent as well as time-independent text considerations can be used to predict the occurrence of financial difficulties in companies. It was shown that the combination and further development of approaches already used in research, which consider the semantic or syntactic level of a text, are suitable, but should be specified contextually to enable the extraction of new knowledge. It should be noted that observation over a specific period of time has proven useful in assessing changes in the characteristics of the textual design of reports (Study III). However, evidence exists that document-related parameters may be less affected by this than wording at the fine granular level (Study IV), as companies may tend to recycle large portions of their text from previous years for new financial statements. Comparing the results of Studies V and VI, it becomes clear that company-related features are suitable to allow specific distinctions, but there is evidence that this does not affect the amount of textual data in financial statements. In this respect, we can assume that not only approaches such as those shown in Study II and Study III for the extraction of text mining features can produce a comprehensive picture, but that it also remains to be questioned on the meta-level of the text which information is consciously or subconsciously not included in the financial statement. In summary, a new interdisciplinary, as well as several context-specific approaches to text mining were presented in this dissertation. Especially with respect to the identification of specific structures, e.g., passive constructions or evidential strategies, the use and modeling of POS tagging, which was applied in Studies II, III, and VI, is worth mentioning. Furthermore, this procedure allows the identification of context- or subject-specific formulations and patterns beyond the application purpose within this

dissertation. The use case shown here can thus be seen as a motivation for the use in further text analyses regardless of the language in which they were written.

Research Area: Artificial Intelligence	
RQ. 3	What is the value of textual data in the development process of an AI-based solution in German companies failure prediction?

In study VI, the usefulness of different text mining approaches in the context of corporate failure prediction was considered. These were classified according to their linguistic analysis and granularity level in order to derive design paradigms. It was shown that a syntactic view of the text based on both lexical and semantic contexts can provide added value for the assessment of the financial situation of companies and seems to be superior to less complex approaches in terms of linguistic analysis and granularity level. Nevertheless, it has to be said that although a desired improvement of the prediction accuracy of an AI-based approach could be demonstrated, greater success can be achieved in comparison to the performance gain, e.g., when adding meta-text information as well as company structure data. Furthermore, it should be mentioned that the classification of financially distressed and bankrupt companies, which is often differentiated in the literature, is also not possible in the prediction based on the examined text features. From a practical point of view, however, the differentiation is of less interest, since the default of a company also has similar negative effects on stakeholders.

From a research point of view and reflecting Occam's razor, the value of the text features presented in this dissertation compared to classical financial ratios, if one considers the percentage improvement, turns out to be low. The addition of such features would contradict the principle of parsimony, but the presented studies also explained that the classical prediction of financial difficulties in companies based on financial ratios is limited in terms of its ability to explain several developments and their causal relationship. In this respect, even small improvements are an important step towards closing this gap in explaining developments from the point of view of a person reading and analyzing a financial statement.

3 Limitations and Future Perspectives

As with any research endeavor, the results of this dissertation are subject to limitations. The following presents an aggregated overview focused on these limitations. Based on a short critical discussion, this enables entry points for further research and work within the field of corporate failure prediction.

First, the introductory literature review conducted as part A of the dissertation, as well as each individual study to identify research gaps and assess the current state of research are systematically limited by the scope of the databases searched. Nevertheless, the articles considered reflect a current picture of research to date, as they have all been subject to a scientific review process, according to the information provided by the publishers or conferences. It should be noted that the evaluation, systematization, and discussion of literature reviews always involve a certain degree of subjective decisions, which we are aware of, but which we have minimized by choosing the CRISP-DM as leading framework. Because of its foundation in the analysis of literature, the taxonomy developed in Study I is subject to the same limitation. It is noteworthy that the taxonomy has not yet been subjected to a third-party evaluation that could validate its usefulness (Szopinski et al. 2019). Reflecting on the discussion, the lack of structure in the research area of corporate failure as well as bankruptcy prediction offers opportunities for future research on the usage of such models in real-world scenarios. It should also be questioned for future research to what extent broader literature research and the consideration of interdisciplinary application contexts are suitable to generate added value, especially with regard to the development of new text mining approaches.

Second, due to the form of its acquisition, the data considered is limited in terms of its use and evaluation in context. While the financial statements used in Studies III and IV were acquired and processed manually, Studies V and VI refer to financial statements randomly collected and provided by an external partner. The linkage with data regarding the companies, collected in the Amadeus database of the Bureau van Dijk, allows an allocation of the reports to financial ratios based on the respective year, but there are no observations that enable a shorter time frame. In addition, it was not possible to form a view of companies in the data set based on the analysis of various financial statements over a larger period of

several years. A consideration of the suitability of text mining approaches related to the specific analysis of corpora (proposed granularity level) is another starting point to be mentioned for future research. Reflecting the point that statistical methods of analysis for the prediction of future developments can only make decisions that are known from the past and are subject to a trigger, such as the new publication of a financial statement, the question arises for future research as to what extent shorter-term predictions are possible. Regarding the unpredictable economic reactions to events like the COVID-19 pandemic and the overall limited number of bankrupt companies compared to solvent ones, the generation of artificial data might be used as starting point for further research. The focus is on the validation of forecasting models with respect to new developments that could be modeled, since current observations, as presented in this dissertation, can only consider historical data.

Third, the limitation of a missing benchmark has to be mentioned, which is caused by the previous limitation of data availability, but is not only a specific feature of this dissertation but is transferable to different research directions. Especially with respect to the evaluation of AI-based models, a deficit was shown, which already results from the phase of business understanding in the CRISP-DM, since without uniform data accessibility no comparable studies are carried out, which allows comparing the performance of such solutions among each other. The results of this dissertation are therefore limited to the time periods considered and the reports of a selection of companies examined cannot, provide information on how the model developed in Study VI compares with others. Thus, future research should also address the question of the extent to which reference data sets can be generated. This research need is a finding of this thesis, which can be generalized, as research can only add value in this way. In summary, the potential is given to refocus research strands that are broad in content due to the availability and linkage of diverse data sources. When establishing a benchmark, it must be dynamic and potentially industry-specific applicable to reflect a variety of possible events, which can have financial consequences for companies. Furthermore, it must also be reflected at this point that an assessment of the performance of an AI-based approach is not possible from the perspective of

comparison with human capabilities, since no research results are available for comparison to date.

Fourth, it must be kept in mind that the use of text mining in the context of financial report analysis has nothing to do with the general limitation of the AI-based model with respect to a trigger. A model always generates a prediction based on the publication of a new financial statement and thus only once for a defined fiscal year. In this respect, an assessment of a company is made in parallel to the publication and therefore does not allow a short-term reaction to new developments in a company that happens in the meantime. Consequently, it could be questioned on which level of abstraction models are trained to assess the financial situation of companies and whether thought must also be given to linked AI-based approaches that can be used as an aggregate at company-level to capture the full potential for changes.

Fifth, the studies conducted are based on an analysis of German-language financial statements, which means that the analyzed language is a limitation of the approaches. Nevertheless, such approaches are also rare in English-language text analysis. An adaptation as well as evaluation with 10K reports or new approaches, which bring differently granular text levels and/or patterns in connection with semantic and/or syntactic analyses are conceivable and could create new research beginnings. General constructs are even conceivable beyond the use case presented here.

In summary, many research needs in general AI research arise from the pursuit of additional data that could be studied in context. Fundamentally, however, it should be noted that in forecasting the financial development of companies we are faced with problems that can be very well abstracted to other contexts. Text mining approaches carried out in the studies may well be used for the analysis of other texts, since such pattern recognition, e.g., evidentiality, may also play a role in many other contexts, depending on their causality. In our research, we focus on issues related to the phases of *data understanding*, *data preparation,* and *modeling*, but we must not forget that in order to transfer and apply the knowledge gained, we also need the phases of business understanding and a clear definition of evaluation in order to create added value for society. Last but not least, it should be emphasized that the presented results and classifications of possible

information to be used are not necessarily limited to the use of AI-based solutions, but can also be used in particular by other evaluation algorithms or models. If the results of this dissertation are placed in the context of recently published contributions according to corporate failure prediction (see A.2), it can be argued that the allegorical step back to restructuring research on company profiles can point to a major step forward. The design of the financial valuation of a company based on the analysis of financial statements alone is due to the use of AI under scrutiny.

Appendix

Findings of the Literature Analysis Process

Source	Used AI-based methods										
	Machine Learning										
		Supervised Learning								Unsupervised Learning	
			Linear Regression	Logistic Regression	Decision Tree	k-Nearest Neighbor	Naive Bayes	Support Vector Machine	Neural Network	Ensemble Learning	Self-organizing map
(Olmeda and Fernández 1997)			○	○	○				○	●	
(Yang and Harrison 2002)			○	○					●		
(Brabazon and Keenan 2004)			○						●		
(Yim and Mitchell 2004)			○	○					○	●	
(Merkevicius et al. 2006)			●								○
(Wu et al. 2006)				○	○		○		●		
(Ozkan-Gunay and Ozkan 2007)									●		
(Ciampi and Gordini 2013)			○	○					●		
(Hsu and Pai 2013)										●	
(Hajek et al. 2014)				○	○			●	○		
(Kirkos 2015)			○	○	○			○	●	○	
(Camacho-Miñano et al. 2015)					●						
(Chung et al. 2016)									●		
(Huang et al. 2016)									●		
(Behr and Weinblat 2017)				○	○					●	
(Jones 2017)				○						●	
(Jones et al. 2017)			○	○						●	
(Zhao et al. 2017)				○		○		○	○	●	
(Zhou and Lai 2017)			○	○	○				○	●	
(Wyrobek and Kluza 2018)			○	○						●	

(Rustam et al. 2018)								●			
(Inam et al. 2019)			○	○					●		
(Slimene and Mamoghli 2019)			○	○						●	
(Tanaka et al. 2019)			○	○					○	●	
(Alaka et al. 2020)									●		
(Lohmann and Ohliger 2020)			●								
(Roumani et al. 2020)				○	○	○				●	
(Sankhwar et al. 2020)									●		
(Smith and Alvarez 2021)				○				○	○	●	
Legend:				● – Best model ○ – Used model							

Table 20. Used AI-based methods (RC. 1) – Bankruptcy prediction

Source	Used AI-based methods											
	Machine Learning										Unsupervised Learning	
	Supervised Learning											
	Linear Regression	Logistic Regression	Decision Tree	k-Nearest Neighbor	Naive Bayes	Support Vector Machine	Neural Networ	Ensemble Learning				Self-organizing map
(Platt and Platt 2002)		●										
(Bose 2006)	○		●									
(Mora et al. 2008)			●				○					○
(Xie et al. 2011)	○					●						
(Sun 2012)	○	○				○		●				
(Shie et al. 2012)		○	○		○	○	○	●				
(Pai et al. 2014)								●				
(Goo et al. 2016)			○			○	●					
(Salehi et al. 2016)				○	○	○	●					
(Huang et al. 2017)	○	○	○			○	○	●				
(Jiang and Jones 2018)								●				
(Kim 2018)								●				
(Ahmadi et al. 2018)							●					
(Farooq and Qamar 2019)		○	○	○	○	○	○	●				
(Balasubramanian et al. 2019)		●										
(Paule-Vianez et al. 2019)	○						●					
(Hsu and Lee 2020)		○	○			○	○	●				
(Clintworth et al. 2021)		○						●				
Legend:	● – Best model ○ – Used model											

Table 21. Used AI-based methods (RC. 2) – Financial distress prediction

Source	Use Case
(Kloptchenko et al. 2004a)	Analysis of dependencies of quantitative and qualitative financial statement data
(Magnusson et al. 2005)	Analysis of language as indicator of change in company's financial status
(Shirata et al. 2011)	Evidence for Word-based collocations and there effects on the financial situation of companies future financial situation
(Appiah et al. 2015)	Literature review on methodological issues in corporate failure prediction
(Loughran and McDonald 2016)	Creation of an english-based sentiment dictionary for financial texts
(Chen et al. 2016)	Literature review on intelligent methods for financial distress forecasting
(Chou et al. 2018)	Development of a text-mining approach based on TF-IDF
(Caserio et al. 2020)	Analysis of dependencies between language tone and financial performance
(Gupta et al. 2020)	Literature review on text-mining applications in finance
(Luo and Zhou 2020)	Literature review on textual tone in financial disclosures
(Myšková and Hájek 2020)	Analysis of risk-related sentiment and its effect on financial performance

Table 22. Financial Statement Analysis without use of AI (RC. 3)

References

(Abbasi et al. 2012): Abbasi, A., Albrecht, C., Vance, A., Hansen, J.: MetaFraud: A Meta-Learning Framework for Detecting Financial Fraud. In: *Management Information Systems Quarterly* 36 (2012) 4. pp. 1293–1327.

(Ahmadi et al. 2018): Ahmadi, Z., Martens, P., Koch, C., Gottron, T., Kramer, S.: *Towards Bankruptcy Prediction: Deep Sentiment Mining to Detect Financial Distress from Business Management Reports.* In: *Proceedings of the IEEE 5th International Conference on Data Science and Advanced Analytics.* Turin, Italy. pp. 293–302.

(Aikhenvald 2004): Aikhenvald, A. Y.: *Evidentiality.* 1. publ. Oxford: Oxford University Press.

(Ajina et al. 2016): Ajina, A., Laouiti, M., Msolli, B.: Guiding through the Fog: Does annual report readability reveal earnings management?. In: *Research in International Business and Finance* 38 (2016). pp. 509–516.

(Alaka et al. 2020): Alaka, H. A., Oyedele, L. O., Owolabi, H. A., Bilal, M., Ajayi, S. O., Akinade, O. O.: A framework for big data analytics approach to failure prediction of construction firms. In: *Applied Computing and Informatics* 16 (2020) 1/2. pp. 207–222.

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

(Altman et al. 1977): Altman, E. I., Haldemann, R. G., Narayanan, P.: ZETA™ analysis: a new model to identify bankruptcy risk of cooperations. In: *Journal of Banking and Finance* 1 (1977) 1. pp. 29–54.

(Appiah et al. 2015): Appiah, K. O., Chizema, A., Arthur, J.: Predicting corporate failure: a systematic literature review of methodological issues. In: *International Journal of Law and Management* 57 (2015) 5. pp. 461–485.

(Ashoori and Weisz 2019): Ashoori, M., Weisz, J. D.: *In AI We Trust? Factors That Influence Trustworthiness of AI-infused Decision-Making Processes*. Available online: https://arxiv.org/pdf/1912.02675 (accessed on January 28, 2023).

(Baker et al. 2008): Baker, P., Gabrielatos, C., KhosraviNik, M., Krzyżanowski, M., McEnery, T., Wodak, R.: A useful methodological synergy? Combining critical discourse analysis and corpus linguistics to examine discourses of refugees and asylum seekers in the UK press. In: *Discourse & Society* 19 (2008) 3. pp. 273–306.

(Balasubramanian et al. 2019): Balasubramanian, S. A., Radhakrishna, G. S., Sridevi, P., Natarajan, T.: Modeling corporate financial distress using financial and non-financial variables. In: *International Journal of Law and Management* 61 (2019) 3/4. pp. 457–484.

(Bankamp et al. 2021): Bankamp, S., Neuss, N., Muntermann, J.: *Crowd Analysts vs. Institutional Analysts – A Comparative Study on Content and Opinion.* In: *Innovation Through Information Systems - Proceedings of WI 2021.* Duisburg-Essen, Germany. pp. 492–508.

(Bankamp and Muntermann 2022): Bankamp, S., Muntermann, J.: *Understanding the Role of Document Representations in Similarity Measurement in Finance and Accounting.* In: *PACIS 2022 Proceedings.* (virtual) Taipei-Sydney, Asia. pp. 1–16.

(Bannier et al. 2019): Bannier, C., Pauls, T., Walter, A.: Content analysis of business communication: introducing a German dictionary. In: *Journal of Business Economics* 89 (2019) 1. pp. 79–123.

(Barovick and Steele 2001): Barovick, B., Steele, C.: The location and site selection decision process: Meeting the strategic and tactical needs of the users of corporate real estate. In: *Journal of Corporate Real Estate* 3 (2001) 4. pp. 356–362.

(Behr and Weinblat 2017): Behr, A., Weinblat, J.: Default prediction using balance-sheet data: a comparison of models. In: *The Journal of Risk Finance* 18 (2017) 5. pp. 523–540.

(Benbya et al. 2020): Benbya, H., Davenport, T. H., Pachidi, S.: Artificial Intelligence in Organizations: Current State and Future Opportunities. In: *MISQ Executive* 19 (2020) 4. pp. 9–21.

(Benson 1989): Benson, M.: The Structure of the Collocational Dictionary. In: *International Journal of Lexicography* 2 (1989) 1. pp. 1–14.

(Besnard 2017): Besnard, A.-L.: BE likely to and BE expected to, epistemic modality or evidentiality? In: Marín Arrese, J. I., Haßler, G. & Carretero, M.

(eds.) *Evidentiality Revisited*, pp. 249–269. Amsterdam: John Benjamins Publishing Company.

(Bjornsson 1983): Bjornsson, C. H.: Readability of Newspapers in 11 Languages. In: *Reading Research Quarterly* 18 (1983) 4. p.480.

(Bloomfield 2002): Bloomfield, R. J.: The “Incomplete Revelation Hypothesis” and Financial Reporting. In: *Accounting Horizons* 16 (2002) 3. pp. 233–243.

(Boas 1938): Boas, F.: Language. In: Boas, F. (ed.) *General anthropology*. Boston: D.C. Heath and Company.

(Bose 2006): Bose, I.: Deciding the financial health of dot-coms using rough sets. In: *Information & Management* 43 (2006) 7. pp. 835–846.

(Brabazon and Keenan 2004): Brabazon, A., Keenan, P. B.: A hybrid genetic model for the prediction of corporate failure. In: *Computational Management Science* 1 (2004) 3-4. pp. 293–310.

(Brants et al. 2004): Brants, S., Dipper, S., Eisenberg, P., Hansen-Schirra, S., König, E., Lezius, W., Rohrer, C., Smith, G., Uszkoreit, H.: TIGER: Linguistic Interpretation of a German Corpus. In: *Research on Language and Computation* 2 (2004) 4. pp. 597–620.

(Brédart et al. 2021): Brédart, X., Séverin, E., Veganzones, D.: Human resources and corporate failure prediction modeling: Evidence from Belgium. In: *Journal of Forecasting* 40 (2021) 7. pp. 1325–1341.

(Breiman 2001): Breiman, L.: Random Forests. In: *Machine Learning* 45 (2001) 1. pp. 5–32.

(Brezina et al. 2015): Brezina, V., McEnery, T., Wattam, S.: Collocations in context. In: *International Journal of Corpus Linguistics* 20 (2015) 2. pp. 139–173.

(Brezina et al. 2021): Brezina, V., Weill-Tessier, P., McEnery, T.: *#LancsBox v. 6.0.* Available online: http://corpora.lancs.ac.uk/lancsbox/ (accessed on January 06, 2022).

(Bureau van Dijk 2021): Bureau van Dijk: *Amadeus database*. Available online: https://www.bvdinfo.com/de-de/unsere-losungen/daten/international/amadeus (accessed on November 03, 2021).

(Bushee et al. 2018): Bushee, B. J., Gow, I. A., Taylor, D. J.: Linguistic Complexity in Firm Disclosures: Obfuscation or Information?. In: *Journal of Accounting Research* 56 (2018) 1. pp. 85–121.

(Butler and Kešelj 2009): Butler, M., Kešelj, V.: *Financial Forecasting Using Character N-Gram Analysis and Readability Scores of Annual Reports.* In: *Proceedings of the 22nd Canadian Conference on Artificial Intelligence.* Kelowna, Canada. pp. 39–51.

(Camacho-Miñano et al. 2015): Camacho-Miñano, M.-M., Segovia-Vargas, M.-J., Pascual-Ezama, D.: Which Characteristics Predict the Survival of Insolvent Firms? An SME Reorganization Prediction Model. In: *Journal of Small Business Management* 53 (2015) 2. pp. 340–354.

(Campbell et al. 2008): Campbell, J. Y., Hilscher, J., Szilagyi, J.: In Search of Distress Risk. In: *The Journal of Finance* 63 (2008) 6. pp. 2899–2939.

(Caserio et al. 2020): Caserio, C., Panaro, D., Trucco, S.: Management discussion and analysis: a tone analysis on US financial listed companies. In: *Management Decision* 58 (2020) 3. pp. 510–525.

(Chafe 1986): Chafe, W. L.: *Evidentiality*. *The linguistic coding of epistemology.* Norwood: Ablex Publishing Corp.

(Chen et al. 2016): Chen, N., Ribeiro, B., an Chen: Financial credit risk assessment: a recent review. In: *Artificial Intelligence Review* 45 (2016) 1. pp. 1–23.

(Chou et al. 2018): Chou, C.-C., Chang, C. J., Chin, C.-L., Chiang, W.-T.: Measuring the Consistency of Quantitative and Qualitative Information in Financial Reports: A Design Science Approach. In: *Journal of Emerging Technologies in Accounting* 15 (2018) 2. pp. 93–109.

(Chung et al. 2016): Chung, C.-C., Chen, T.-S., Lin, L.-H., Lin, Y.-C., Lin, C.-M.: Bankruptcy Prediction Using Cerebellar Model Neural Networks. In: *International Journal of Fuzzy Systems* 18 (2016) 2. pp. 160–167.

(Church and Hanks 1989): Church, K. W., Hanks, P.: *Word association norms, mutual information, and lexicography.* In: *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics.* Vancouver, Canada. pp. 76–83.

(Ciampi and Gordini 2013): Ciampi, F., Gordini, N.: Small Enterprise Default Prediction Modeling through Artificial Neural Networks: An Empirical Analysis

of Italian Small Enterprises. In: *Journal of Small Business Management* 51 (2013) 1. pp. 23–45.

(Clemen 1998): Clemen, G.: Hecken in deutschen und englischen Texten der Wirtschaftskommunikation : eine kontrastive Analyse. Universität Siegen.

(Clintworth et al. 2021): Clintworth, M., Lyridis, D., Boulougouris, E.: Financial risk assessment in shipping: a holistic machine learning based methodology. In: *Maritime Economics & Logistics (*2021).

(Cooper 1988): Cooper, H.: Organizing knowledge syntheses: A taxonomy of literature reviews. In: *Knowledge in Society* 104 (1988) 1. pp. 104–126.

(Creditreform 2021): Creditreform: *Insolvenzen in Deutschland, Jahr 2021*. Available online: https://www.creditreform.de/muenster/aktuelles-wissen/pressemeldungen-fachbeitraege/news-details/show/insolvenzen-in-deutschland-jahr-2021 (accessed on January 05, 2022).

(Creditreform 2022): Creditreform: *Insolvenzen in Deutschland, 1. Halbjahr 2022*. Available online: https://www.creditreform.de/aktuelles-wissen/pressemeldungen-fachbeitraege/news-details/show/insolvenzen-in-deutschland-1-halbjahr-2022 (accessed on August 23, 2022).

(Daille 1995): Daille, B.: Combined approach for terminology extraction: lexical statistics and linguistic filtering. Université Paris VII.

(Delivery Hero 2020): Delivery Hero: *Annual report 2020*. Available online: https://ir.deliveryhero.com/download/companies/delivery/Annual%20Reports/DE000A2E4K43-JA-2020-EQ-E-01.pdf (accessed on February 22, 2022).

(Delivery Hero 2021): Delivery Hero: *Annual report 2021*. Available online: https://ir.deliveryhero.com/download/companies/delivery/Annual%20Reports/DE000A2E4K43-JA-2021-EQ-E-01.pdf (accessed on February 12, 2022).

(Deloitte 2018a): Deloitte: *AI and risk management*. Available online: https://www2.deloitte.com/global/en/pages/financial-services/articles/gx-ai-and-risk-management.html (accessed on January 14, 2021).

(Deloitte 2018b): Deloitte: *The New Physics of Financial Services - How artificial intelligence is transforming the financial ecosystem*. Available online: https://www.weforum.org/reports/the-new-physics-of-financial-services-how-

artificial-intelligence-is-transforming-the-financial-ecosystem (accessed on February 23, 2022).

(Deng et al. 2018): Deng, S., Huang, Z., Sinha, A., Zhao, H.: The Interaction Between Microblog Sentiment and Stock Returns: An Empirical Examination. In: *Management Information Systems Quarterly* 42 (2018) 3. pp. 895–918.

(DeWilde 2021): DeWilde, B.: *textacy*. Available online: https://pypi.org/project/-textacy/ (accessed on January 06, 2022).

(Diaz and Suriyawongkul 2020): Diaz, G., Suriyawongkul, A.: *German stopword list*. Available online: https://github.com/stopwords-iso/stopwords-de (accessed on December 10, 2021).

(El Kalak and Hudson 2016): El Kalak, I., Hudson, R.: The effect of size on the failure probabilities of SMEs: An empirical study on the US market using discrete hazard model. In: *International Review of Financial Analysis* 43 (2016) 4. pp. 135–145.

(ESMA 2020): ESMA: *Fast track peer review on the application of the guidelines on the enforement of financial information*. Available online: https://www.esma.europa.eu/sites/default/files/library/esma42-111-5349_fast_track_peer_review_report_-_wirecard.pdf (accessed on January 14, 2021).

(eurostat 2022): eurostat: *Aufstellung der statistischen System der Wirtschaftszweige*. Available online: https://ec.europa.eu/eurostat/de/web/-products-manuals-and-guidelines/-/ks-ra-07-015 (accessed on February 17, 2022).

(Farooq and Qamar 2019): Farooq, U., Qamar, M. A. J.: Predicting multistage financial distress: Reflections on sampling, feature and model selection criteria. In: *Journal of Forecasting* 38 (2019) 7. pp. 632–648.

(Fetzer 2014): Fetzer, A.: Foregrounding evidentiality in (English) academic discourse: Patterned co-occurrences of the sensory perception verbs seem and appear. In: *Intercultural Pragmatics* 11 (2014) 3. pp. 333–355.

(Fromm et al. 2019): Fromm, H., Wambsganss, T., Söllner, M.: Towards a taxonomy of text mining features. In: *Proceedings of the 27th European Conference on Information Systems (*2019). pp. 8–14.

(Goel and Gangolly 2012): Goel, S., Gangolly, J.: Beyond the numbers: Mining the annual reports for hidden cues indicative of financial statement fraud. In:

Intelligent Systems in Accounting, Finance and Management 19 (2012) 2. pp. 75–89.

(Goo et al. 2016): Goo, Y.-J. J., Chi, D.-J., Shen, Z.-D.: Improving the prediction of going concern of Taiwanese listed companies using a hybrid of LASSO with data mining techniques. In: *SpringerPlus* 5 (2016) 1. p.539.

(Gregor 2006): Gregor: The Nature of Theory in Information Systems. In: *MIS Quarterly* 30 (2006) 3. p.611.

(Gupta et al. 2020): Gupta, A., Dengre, V., Kheruwala, H. A., Shah, M.: Comprehensive review of text-mining applications in finance. In: *Financial Innovation* 6 (2020) 1. pp. 1–25.

(Hack 1999): Hack, G. D.: *Site selection for growing companies.* 1. publ. Westport, Conn.: Quorum Books.

(Hajek et al. 2014): Hajek, P., Olej, V., Myskova, R.: Forecasting Corporate Financial Performance using Sentiment in Annual Reports for stakeholders Decision Making. In: *Technological and Economic Development of Economy* 20 (2014) 4. pp. 721–738.

(Hajek and Henriques 2017): Hajek, P., Henriques, R.: Mining corporate annual reports for intelligent detection of financial statement fraud – A comparative study of machine learning methods. In: *Knowledge-Based Systems* 128 (2017). pp. 139–152.

(Hardman 1986): Hardman, M. J.: Data-source Marking in the Jaqi Languages. In: *Evidentiality: The Linguistic Coding of Epistemology (*1986). pp. 113–136.

(Hevner et al. 2004): Hevner, A., March, S., Park, J., Ram, S.: Design Science in Information Systems Research. In: *MIS Quarterly* 28 (2004) 1. pp. 75–105.

(HGB 2021a): HGB: *Handelsgesetzbuch §267 - Umschreibung der Größenklassen*. Available online: https://www.gesetze-im-internet.de/hgb/__267.html (accessed on November 15, 2021).

(HGB 2021b): HGB: *Handelsgesetzbuch §289 - Inhalt des Lageberichts*. Available online: https://www.gesetze-im-internet.de/hgb/__289.html (accessed on November 15, 2021).

(Hidalgo-Downing 2017): Hidalgo-Downing, L.: Evidential and epistemic stance strategies in scientific communication. In: Marín Arrese, J. I., Haßler, G. &

Carretero, M. (eds.) *Evidentiality Revisited*, pp. 225–248. Amsterdam: John Benjamins Publishing Company.

(Hinrichs 1999): Hinrichs, B.: Passivstrukturen im Spanischen und im Deutschen: ein kontrastiver Vergleich. In: *Informationen Deutsch als Fremdsprache* 26 (1999). pp. 348–354.

(Hsu and Lee 2020): Hsu, Y.-F., Lee, W.-P.: Evaluation of the going-concern status for companies: An ensemble framework-based model. In: *Journal of Forecasting* 39 (2020) 4. pp. 687–706.

(Hsu and Pai 2013): Hsu, M.-F., Pai, P.-F.: Incorporating support vector machines with multiple criteria decision making for financial crisis analysis. In: *Quality & Quantity* 47 (2013) 6. pp. 3481–3492.

(Huang et al. 2016): Huang, T.-H., Leu, Y., Pan, W.-T.: Constructing ZSCORE-based financial crisis warning models using fruit fly optimization algorithm and general regression neural network. In: *Kybernetes* 45 (2016) 4. pp. 650–665.

(Huang et al. 2017): Huang, J., Wang, H., Kochenberger, G.: Distressed Chinese firm prediction with discretized data. In: *Management Decision* 55 (2017) 5. pp. 786–807.

(Humpherys 2009): Humpherys, S.: *Discriminating Fraudulent Financial Statements by Identifying Linguistic Hedging.* In: *Proceedings of the 15th Americas Conference on Information Systems.* San Francisco, California. pp. 1–10.

(Hyland 1998): Hyland, K.: *Hedging in scientific research articles.* Amsterdam, Philadelphia: Benjamins.

(IBM 2021): IBM: *AI in the enterprise: Unleashing opportunity through data*. *Results from research conducted in 2021 by IBM Market Development & Insights*. Available online: https://www.ibm.com/downloads/cas/6DR9QRVQ (accessed on January 27, 2023).

(Inam et al. 2019): Inam, F., Inam, A., Mian, M. A., Sheikh, A. A., Awan, H. M.: Forecasting Bankruptcy for organizational sustainability in Pakistan. In: *Journal of Economic and Administrative Sciences* 35 (2019) 3. pp. 183–201.

(InsBekV 2021): InsBekV: *Insolvenzbekanntmachungen*. Available online: https://www.insolvenzbekanntmachungen.de/ (accessed on December 10, 2021).

(InsO 2022): InsO: *Insolvenzverordnung - §9 Öffentliche Bekanntmachung*. Available online: https://www.gesetze-im-internet.de/inso/__9.html (accessed on January 24, 2022).

(Janssen et al. 2020): Janssen, A., Passlick, J., Rodríguez Cardona, D., Breitner, M. H.: Virtual Assistance in Any Context. In: *Business & Information Systems Engineering* 62 (2020) 3. pp. 211–225.

(Jebara and Meila 2006): Jebara, T., Meila, M.: Machine learning: Discriminative and generative. In: *The Mathematical Intelligencer* 28 (2006) 1. pp. 67–69.

(Jiang and Jones 2018): Jiang, Y., Jones, S.: Corporate distress prediction in China: a machine learning approach. In: *Accounting & Finance* 58 (2018) 4. pp. 1063–1109.

(Jones 2017): Jones, S.: Corporate bankruptcy prediction: a high dimensional analysis. In: *Review of Accounting Studies* 22 (2017) 3. pp. 1366–1422.

(Jones et al. 2017): Jones, S., Johnstone, D., Wilson, R.: Predicting Corporate Bankruptcy: An Evaluation of Alternative Statistical Frameworks. In: *Journal of Business Finance & Accounting* 44 (2017) 1-2. pp. 3–34.

(Kim 2018): Kim, S. Y.: Predicting hospitality financial distress with ensemble models: the case of US hotels, restaurants, and amusement and recreation. In: *Service Business* 12 (2018) 3. pp. 483–503.

(Kirkos 2015): Kirkos, E.: Assessing methodologies for intelligent bankruptcy prediction. In: *Artificial Intelligence Review* 43 (2015) 1. pp. 83–123.

(Kloptchenko et al. 2004a): Kloptchenko, A., Eklund, T., Karlsson, J., Back, B., Vanharanta, H., Visa, A.: Combining data and text mining techniques for analysing financial reports. In: *Intelligent Systems in Accounting, Finance & Management* 12 (2004) 1. pp. 29–41.

(Kloptchenko et al. 2004b): Kloptchenko, A., Magnusson, C., Back, B., Visa, A., Vanharanta, H.: Mining Textual Contents of Financial Reports. In: *The International Journal of Digital Accounting Research* 4 (2004) 7. pp. 1–29.

(Kloptchenko 2004): Kloptchenko, A.: *Toward Automatic Analysis of Financial Reports: Readability of Quarterly Reports and Companies' Financial Performance.* In: *Proceedings of the 10th Americas Conference on Information Systems.* New York City, New York. pp. 3273–3280.

(KPMG 2019): KPMG: *Controlling AI*. Available online: https://advisory.kpmg.us/articles/2019/controlling-ai.html (accessed on January 18, 2021).

(Kumawat and Jain 2015): Kumawat, D., Jain, V.: POS Tagging Approaches: A Comparison. In: *International Journal of Computer Applications* 118 (2015) 6. pp. 32–38.

(Kupferschmidt 2022): Kupferschmidt, K.: *Studying—and fighting—misinformation should be a top scientific priority, biologist argues*. Available online: https://www.science.org/content/article/studying-fighting-misinformation-top-scientific-priority-biologist-argues (accessed on January 21, 2023).

(Lang and Stice-Lawrence 2015): Lang, M., Stice-Lawrence, L.: Textual analysis and international financial reporting: Large sample evidence. In: *Journal of Accounting and Economics* 60 (2015) 2-3. pp. 110–135.

(Le Maux and Smaili 2021): Le Maux, J., Smaili, N.: Annual Report Readability And Corporate Bankruptcy. In: *Journal of Applied Business Research* 37 (2021) 3. pp. 73–80.

(Levy and Ellis 2006): Levy, Y., Ellis, T. J.: A Systems Approach to Conduct an Effective Literature Review in Support of Information Systems Research. In: *Informing Science: The International Journal of an Emerging Transdiscipline* 9 (2006). pp. 181–212.

(Li 2008): Li, F.: Annual report readability, current earnings, and earnings persistence. In: *Journal of Accounting and Economics* 45 (2008) 2/3. pp. 221–247.

(Lin et al. 2015): Lin, C.-C., Chiu, A.-A., Huang, S. Y., Yen, D. C.: Detecting the financial statement fraud: The analysis of the differences between data mining techniques and experts' judgments. In: *Knowledge-Based Systems* 89 (2015). pp. 459–470.

(Lohmann and Ohliger 2020): Lohmann, C., Ohliger, T.: Bankruptcy prediction and the discriminatory power of annual reports: empirical evidence from financially distressed German companies. In: *Journal of Business Economics* 90 (2020) 1. pp. 137–172.

(Loughran and McDonald 2011): Loughran, T. I. M., McDonald, B.: When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks. In: *The Journal of Finance* 66 (2011) 1. pp. 35–65.

(Loughran and McDonald 2014): Loughran, T. I. M., McDonald, B.: Measuring Readability in Financial Disclosures. In: *The Journal of Finance* 69 (2014) 4. pp. 1643–1671.

(Loughran and McDonald 2016): Loughran, T. I. M., McDonald, B.: Textual Analysis in Accounting and Finance: A Survey. In: *Journal of Accounting Research* 54 (2016) 4. pp. 1187–1230.

(Luo and Zhou 2020): Luo, Y., Zhou, L.: Textual tone in corporate financial disclosures: a survey of the literature. In: *International Journal of Disclosure and Governance* 17 (2020) 2-3. pp. 101–110.

(Magnusson et al. 2005): Magnusson, C., Arppe, A., Eklund, T., Back, B., Vanharanta, H., Visa, A.: The language of quarterly reports as an indicator of change in the company's financial status. In: *Information & Management* 42 (2005) 4. pp. 561–574.

(Marín Arrese 2017): Marín Arrese, J. I.: Multifunctionality of evidential expressions in discourse domains and genres. In: Marín Arrese, J. I., Haßler, G. & Carretero, M. (eds.) *Evidentiality Revisited*, pp. 195–223. Amsterdam: John Benjamins Publishing Company.

(Mayer et al. 2020): Mayer, A.-S., Strich, F., Fiedler, M.: Unintended Consequences of Introducing AI Systems for Decision Making. In: *MIS Quarterly Executive* 19 (2020) 4. pp. 239–257.

(Mayew et al. 2015): Mayew, W. J., Sethuraman, M., Venkatachalam, M.: MD&A disclosure and the firm's ability to continue as a going concern. In: *The Accounting Review* 90 (2015) 4. pp. 1621–1651.

(Meadows et al. 2022): Meadows, M., Merendino, A., Dibb, S., Garcia-Perez, A., Hinton, M., Papagiannidis, S., Pappas, I., Wang, H.: Tension in the data environment: How organisations can meet the challenge. In: *Technological Forecasting and Social Change* 175 (2022). p.121315.

(Merkevicius et al. 2006): Merkevicius, E., Garšva, G., Girdzijauskas, S.: *A Hybrid SOM-Altman Model for Bankruptcy Prediction.* In: *Proceedings of the 6th*

International Conference on Computational Science. Reading, United Kingdom: Springer Verlag 3994. pp. 364–371.

(Mikalef et al. 2020): Mikalef, P., Pappas, I., Krogstie, J., Pavlou, P.: Big data and business analytics: A research agenda for realizing business value. In: *Information & Management (*2020).

(Miller 2020): Miller, G.: *As U.S. election nears, researchers are following the trail of fake news*. Available online: https://www.science.org/content/article/us-election-nears-researchers-are-following-trail-fake-news (accessed on January 21, 2023).

(Mithas et al. 2013): Mithas, S., Tafti, A., Mitchell, W.: How a Firm's Competitive Environment and Digital Strategic Posture Influence Digital Business Strategy. In: *MIS Quarterly* 37 (2013) 2. pp. 511–536.

(Moore et al. 2006): Moore, D. A., Tetlock, P. E., Tanlu, L., Bazerman, M. H.: Conflicts Of Interest And The Case Of Auditor Independence: Moral Seduction And Strategic Issue Cycling. In: *The Academy of Management Review* 31 (2006) 1. pp. 10–29.

(Mora et al. 2008): Mora, A. M., Castillo, Pedro A. Valdivieso, Guervós, J. J. M., Alfaro-Cid, E., Sharman, K.: *Discovering causes of financial distress by combining evolutionary algorithms and artificial neural networks.* In: *Proceedings of the Genetic and Evolutionary Computation Conference.* Atlanta, Georgia, USA. pp. 1243–1250.

(Myšková and Hájek 2020): Myšková, R., Hájek, P.: Mining risk-related sentiment in corporate annual reports and its effect on financial performance. In: *Technological and Economic Development of Economy* 26 (2020) 6. pp. 1422–1443.

(Nakatsu et al. 2014): Nakatsu, R. T., Grossman, E. B., Iacovou, C. L.: A taxonomy of crowdsourcing based on task complexity. In: *Journal of Information Science* 40 (2014) 6. pp. 823–834.

(Nartey and Mwinlaaru 2019): Nartey, M., Mwinlaaru, I. N.: Towards a decade of synergising corpus linguistics and critical discourse analysis: a meta-analysis. In: *Corpora* 14 (2019) 2. pp. 203–235.

(Nassirtoussi et al. 2014): Nassirtoussi, A., Aghabozorgi, S., Ying Wah, T., Ngo, D. C. L.: Text mining for market prediction: A systematic review. In: *Expert Systems with Applications* 41 (2014) 16. pp. 7653–7670.

(Ngai et al. 2011): Ngai, E., Hu, Y., Wong, Y. H., Chen, Y., Sun, X.: The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature. In: *Decision Support Systems* 50 (2011) 3. pp. 559–569.

(Nickerson et al. 2013): Nickerson, R. C., Varshney, U., Muntermann, J.: A method for taxonomy development and its application in information systems. In: *European Journal of Information Systems* 22 (2013) 3. pp. 336–359.

(Nießner et al. 2021): Nießner, T., Nickerson, R., Schumann, M.: *Towards a taxonomy of AI-based methods in Financial Statement Analysis.* In: *Proceedings of the 27th Americas Conference on Information Systems.* (virtual) Montreal, Canada. pp. 1–10.

(Nießner et al. 2022a): Nießner, T., Wiederspan, O., Schumann, M.: *Consideration of the Use of Language in Corporate Bankruptcy Prediction: A data analysis on German Companies.* In: *PACIS 2022 Proceedings.* (virtual) Taipei-Sydney, Asia. pp. 1-15

(Nießner et al. 2022b): Nießner, T., Gross, D. H., Schumann, M.: Evidential Strategies in Financial Statement Analysis: A Corpus Linguistic Text Mining Approach to Bankruptcy Prediction. In: *Journal of Risk and Financial Management* 15 (2022) 10. p. 459.

(Nießner et al. 2022c): Nießner, T., Nießner, S., Schumann, M.: *Influence of corporate industry affiliation in Financial Business Forecasting: A data analysis concerning competition.* In: *Proceedings of the 28th Americas Conference on Information Systems.* Minneapolis, Minnesota, USA. pp. 1-10

(Nießner et al. 2023): Nießner, T., Nießner, S., Schumann, M.: Is it worth the effort? Considerations on Text Mining in AI-based Corporate Failure Prediction. In: *Information 14 (*2023) 4.

(Nießner and Schumann 2022): Nießner, T., Schumann, M.: *Analysis of consecutive financial statements concerning bankruptcy prediction.* In: *Contributions to Risk Analysis: RISK 2022.* Barcelona, Spain. pp. 197–206.

(Nopp and Hanbury 2015): Nopp, C., Hanbury, A.: *Detecting Risks in the Banking System by Sentiment Analysis.* In: *Proceedings of the 2015 Conference on*

Empirical Methods in Natural Language Processing. Lisbon, Portugal. pp. 591–600.

(Oberländer et al. 2019): Oberländer, A. M., Lösser, B., Rau, D.: *Taxonomy Research in Information Systems: A Systematic Assessment.* In: *Proceedings of the 27th European Conference on Information Systems.* Stockholm, Sweden. pp. 1–17.

(Ohlson 1980): Ohlson, J. A.: Financial Ratios and the Probabilistic Prediction of Bankruptcy. In: *Journal of Accounting Research* 18 (1980) 1. p.109.

(Olmeda and Fernández 1997): Olmeda, I., Fernández, E.: Hybrid Classifiers for Financial Multicriteria Decision Making: The Case of Bankruptcy Prediction. In: *Computational Economics* 10 (1997) 4. pp. 317–335.

(Oswalt 1986): Oswalt, R. L.: The Evidential System of Kashaya. In: Chafe, W. L. & Nichols, J. (eds.) *Evidentiality: The Linguistic Coding of Epistemology*, pp. 20–29. Norwood: Ablex Publishing Corp.

(Ozkan-Gunay and Ozkan 2007): Ozkan-Gunay, E., Ozkan, M.: Prediction of bank failures in emerging financial markets: an ANN approach. In: *The Journal of Risk Finance* 8 (2007) 5. pp. 465–480.

(Pai et al. 2014): Pai, P.-F., Hsu, M.-F., Lin, L.: Enhancing decisions with life cycle analysis for risk management. In: *Neural Computing and Applications* 24 (2014) 7-8. pp. 1717–1724.

(Pamuk et al. 2021): Pamuk, M., Grendel, R. O., Schumann, M.: *Towards ML-based Platforms in Finance Industry - An ML approach to Generate Corporate Bankruptcy Probabilites based on Annual Financial Statements.* In: *Proceedings of the 32nd Australasian Conference on Information Systems.* (virtual) Sydney, Australia. pp. 1–12.

(Pappas et al. 2018): Pappas, I. O., Mikalef, P., Giannakos, M. N., Krogstie, J., Lekakos, G.: Big data and business analytics ecosystems: paving the way towards digital transformation and sustainable societies. In: *Information Systems and e-Business Management* 16 (2018) 3. pp. 479–491.

(Partington 2004): Partington, A.: "Utterly content in each other's company". In: *International Journal of Corpus Linguistics* 9 (2004) 1. pp. 131–156.

(Partington 2015): Partington, A.: Evaluative prosody. In: Aijmer, K. & Rühlemann, C. (eds.) *Corpus Pragmatics*, pp. 279–303. Cambridge: Cambridge University Press.

(Paule-Vianez et al. 2019): Paule-Vianez, J., Gutiérrez-Fernández, M., Coca-Pérez, J. L.: Prediction of financial distress in the Spanish banking system. In: *Applied Economic Analysis* 28 (2019) 82. pp. 69–87.

(Platt and Platt 2002): Platt, H. D., Platt, M. B.: Predicting corporate financial distress: Reflections on choice-based sample bias. In: *Journal of Economics and Finance* 26 (2002) 2. pp. 184–199.

(Pollach 2012): Pollach, I.: Taming Textual Data: The Contribution of Corpus Linguistics to Computer-Aided Text Analysis. In: *Organizational Research Methods* 15 (2012) 2. pp. 263–287.

(Porter et al. 2020): Porter, M., Chelli, A., Lakhdar, B.: *SnowballStemmer: "German"*. Available online: https://www.nltk.org/_modules/nltk/stem/-snowball.html (accessed on December 11, 2021).

(Prat et al. 2015): Prat, N., Comyn-Wattiau, I., Akoka, J.: A Taxonomy of Evaluation Methods for Information Systems Artifacts. In: *Journal of Management Information Systems* 32 (2015) 3. pp. 229–267.

(Precourt and Oppenheimer 2015): Precourt, E., Oppenheimer, H.: What Do Institutional Investors Know and Act on Before Almost Everyone Else: Evidence from Corporate Bankruptcies. In: *Journal of Accounting and Finance* 15 (2015) 7. pp. 103–127.

(PwC 2020): PwC: *How mature is AI adoption in financial services?* Available online: https://www.pwc.de/de/future-of-finance/how-mature-is-ai-adoption-in-financial-services.pdf (accessed on January 17, 2021).

(Ravisankar et al. 2011): Ravisankar, P., Ravi, V., Raghava Rao, G., Bose, I.: Detection of financial statement fraud and feature selection using data mining techniques. In: *Decision Support Systems* 50 (2011) 2. pp. 491–500.

(Reimann 2021): Reimann, M.: *Grundstufen-Grammatik für Deutsch als Fremdsprache*. *Erklärungen und Übungen : mit integriertem Lösungsschlüssel.* Aktualisierte Auflage, 1. Auflage. München: Hueber Verlag.

(Remus et al. 2010): Remus, R., Quasthoff, U., Heyer, G.: *SentiWS - A Publicly Available German-language Resource for Sentiment Analysis.* In: *Proceedings of the 7th International Language Resources and Evaluation.* Valletta, Malta. pp. 1168–1171.

(Richardson 2021): Richardson, L.: *BeautifulSoup v4.8.2*. Available online: https://www.crummy.com/software/BeautifulSoup/bs4/doc/ (accessed on December 11, 2021).

(Roumani et al. 2020): Roumani, Y. F., Nwankpa, J. K., Tanniru, M.: Predicting firm failure in the software industry. In: *Artificial Intelligence Review* 53 (2020) 6. pp. 4161–4182.

(Rustam et al. 2018): Rustam, Z., Nadhifa, F., Acar, M.: Comparison of SVM and FSVM for predicting bank failures using chi-square feature selection. In: *Journal of Physics Conference Series* 1108 (2018). p.12115.

(Salehi et al. 2016): Salehi, M., Mousavi Shiri, M., Bolandraftar Pasikhani, M.: Predicting corporate financial distress using data mining techniques. In: *International Journal of Law and Management* 58 (2016) 2. pp. 216–230.

(Sankhwar et al. 2020): Sankhwar, S., Gupta, D., Ramya, K. C., Sheeba Rani, S., Shankar, K., Lakshmanaprabu, S. K.: Improved grey wolf optimization-based feature subset selection with fuzzy neural classifier for financial crisis prediction. In: *Soft Computing* 24 (2020) 1. pp. 101–110.

(Schmid and Laws 2008): Schmid, H., Laws, F.: *Estimation of conditional probabilities with decision trees and an application to fine-grained POS tagging.* In: *Proceedings of the 22nd International Conference on Computational Linguistics.* Manchester, UK. pp. 777–784.

(Schöbel et al. 2020): Schöbel, S. M., Janson, A., Söllner, M.: Capturing the complexity of gamification elements: a holistic approach for analysing existing and deriving novel gamification designs. In: *European Journal of Information Systems* 29 (2020) 6. pp. 641–668.

(Schumaker et al. 2012): Schumaker, R. P., Zhang, Y., Huang, C.-N., Chen, H.: Evaluating sentiment in financial news articles. In: *Decision Support Systems* 53 (2012) 3. pp. 458–464.

(scikit-learn 2022): scikit-learn: *Machine Learning in Python*. Available online: https://scikit-learn.org/stable/ (accessed on January 28, 2022).

(Shearer 2000): Shearer, C.: The CRISP-DM Model: The New Blueprint for Data Mining. In: *Journal of Data Warehousing* 4 (2000) 5. pp. 13–22.

(Shie et al. 2012): Shie, F. S., Chen, M.-Y., Liu, Y.-S.: Prediction of corporate financial distress: an application of the America banking industry. In: *Neural Computing and Applications* 21 (2012) 7. pp. 1687–1696.

(Shirata et al. 2011): Shirata, C. Y., Takeuchi, H., Ogino, S., Watanabe, H.: Extracting Key Phrases as Predictors of Corporate Bankruptcy: Empirical Analysis of Annual Reports by Text Mining. In: *Journal of Emerging Technologies in Accounting* 8 (2011) 1.

(Sinclair 2004): Sinclair, J.: *Trust the text*. *Language, corpus and discourse.* 1. publ. London: Routledge.

(Slimene and Mamoghli 2019): Slimene, S. B., Mamoghli, C.: NeuroEvolution of Augmenting Topologies for predicting financial distress: A multicriteria decision analysis. In: *Journal of Multi-Criteria Decision Analysis* 26 (2019) 5-6. pp. 320–328.

(Smith and Alvarez 2021): Smith, M., Alvarez, F.: Predicting Firm-Level Bankruptcy in the Spanish Economy Using Extreme Gradient Boosting. In: *Computational Economics (*2021). pp. 1–33.

(Sommer 2020): Sommer, U.: *So wenige Insolvenzen wie seit Jahrzehnten nicht mehr – doch Hunderttausende Jobs betroffen*. Available online: https://www.handelsblatt.com/unternehmen/management/insolvenzen-so-wenige-insolvenzen-wie-seit-jahrzehnten-nicht-mehr-doch-hunderttausende-jobs-betroffen/26696884.html (accessed on January 05, 2022).

(spaCy v3.2 2021): spaCy v3.2: *spaCy: Industrial-strength NLP*. Available online: https://github.com/explosion/spaCy (accessed on November 15, 2021).

(Statista 2022): Statista: *Fläche der deutschen Bundesländer zum 31. Dezember 2020*. Available online: https://de.statista.com/statistik/daten/studie/154868/-umfrage/flaeche-der-deutschen-bundeslaender/
(accessed on January 23, 2022).

(Sun 2012): Sun, J.: Integration Of Random Sample Selection, Support Vector Machines And Ensembles For Financial Risk Forecasting With An Empirical Analysis On The Necessity Of Feature Selection. In: *Intelligent Systems in Accounting, Finance and Management* 19 (2012) 4. pp. 229–246.

(Szopinski et al. 2019): Szopinski, D., Schoormann, T., Kundisch, D.: *Because your Taxonomy is worth it: Towards a Framework for Taxonomy Evaluation.* In: *Proceedings of the 27th European Conference on Information Systems.* Stockholm, Sweden. pp. 1–17.

(Tamm et al. 2020): Tamm, T., Seddon, P. B., Shanks, G.: How Do Different Types of BA Users Contribute to Business Value?. In: *Communications of the Association for Information Systems* 46 (2020) 1. pp. 656–678.

(Tanaka et al. 2019): Tanaka, K., Higashide, T., Kinkyo, T., Hamori, S.: Analyzing industry-level vulnerability by predicting financial bankruptcy. In: *Economic Inquiry* 57 (2019) 4. pp. 2017–2034.

(Veganzones and Severin 2021): Veganzones, D., Severin, E.: Corporate failure prediction models in the twenty-first century: a review. In: *European Business Review* 33 (2021) 2. pp. 204–226.

(Vom Brocke et al. 2009): Vom Brocke, J., Simons, A., Niehaves, B., Niehaves, B., Reimer, K., Plattfaut, R., Cleven, A.: *Reconstructing the Giant: On the importance of rigour in documenting the literature search process.* In: *Proceedings of the 17th European Conference on Information Systems.* Verona, Italy. pp. 1–12.

(Vom Brocke et al. 2015): Vom Brocke, J., Simons, A., Riemer, K., Niehaves, B., Plattfaut, R., Cleven, A.: Standing on the Shoulders of Giants: Challenges and Recommendations of Literature Search in Information Systems Research. In: *Communications of the Association for Information Systems* 37 (2015) 1.

(Wambsganss et al. 2021): Wambsganss, T., Engel, C., Fromm, H.: *Improving Explainability and Accuracy through Feature Engineering: A Taxonomy of Features in NLP-based Machine Learning.* In: *ICIS 2021 Proceedings.* Austin, Texas, USA. pp. 1–17.

(Webster and Watson 2002): Webster, J., Watson, R.: Analyzing the past to prepare for the future: Writing a literature review. In: *Management Information Systems Quarterly* 26 (2002) 2. pp. 13–23.

(Wei et al. 2019): Wei, L., Li, G., Zhu, X., Li, J.: Discovering bank risk factors from financial statements based on a new semi-supervised text mining algorithm. In: *Accounting & Finance* 59 (2019) 3. pp. 1519–1552.

(Williams 1998): Williams, G.: Collocational Networks. In: *International Journal of Corpus Linguistics* 3 (1998) 1. pp. 151–171.

(Wirth and Hipp 2000): Wirth, R., Hipp, J.: CRISP-DM: Towards a standard process model for data mining. In: *Practical application of knowledge discovery and data mining (*2000). pp. 29–40.

(Wodak 2013): Wodak, R.: *Critical Discourse Analysis.* Los Angeles: SAGE.

(Wu et al. 2006): Wu, W., Lee, V. C. S., Tan, T. Y.: Data preprocessing and data parsimony in corporate failure forecast models: evidence from Australian materials industry. In: *Accounting and Finance* 46 (2006) 2. pp. 327–345.

(Wyrobek and Kluza 2018): Wyrobek, J., Kluza, K.: *Efficiency of Gradient Boosting Decision Trees Technique in Polish Companies' Bankruptcy Prediction.* In: *Proceedings of 39th International Conference on Information Systems Architecture and Technology.* Nysa, Poland. pp. 24–35.

(Xie et al. 2011): Xie, C., Luo, C., Yu, X.: Financial distress prediction based on SVM and MDA methods: the case of Chinese listed companies. In: *Quality & Quantity: International Journal of Methodology* 45 (2011) 3. pp. 671–686.

(Yang and Harrison 2002): Yang, Z. R., Harrison, R. G.: Analysing company performance using templates. In: *Intelligent Data Analysis* 6 (2002) 1. pp. 3–15.

(Yim and Mitchell 2004): Yim, J., Mitchell, H.: A comparison of Japanese failure models: Hybrid neural networks, logit models, and discriminant analysis. In: *International Journal of Asian Management* 3 (2004) 1. pp. 103–120.

(Zensus 2022): Zensus: *Die Ergebnisse des Zensus*. Available online: https://ergebnisse2011.zensus2022.de/datenbank/online/ (accessed on January 24, 2022).

(Zhang et al. 2002): Zhang, Y., Callan, J., Minka, T.: *Novelty and redundancy detection in adaptive filtering.* In: *Proceedings of the 25th annual international ACM SIGIR conference on research and development in information retrieval.* New York, New York, USA. pp. 81–88.

(Zhang et al. 2020): Zhang, Z., Nandhakumar, J., Hummel, J., Waardenburg, L.: Addressing the Key Challenges of Developing Machine Learning AI Systems for Knowledge-Intensive Work. In: *MIS Quarterly Executive* 19 (2020) 4.

(Zhao et al. 2017): Zhao, D., Huang, C., Wei, Y., Yu, F., Wang, M., Chen, H.: An Effective Computational Model for Bankruptcy Prediction Using Kernel Extreme Learning Machine Approach. In: *Computational Economics* 49 (2017) 2. pp. 325–341.

(Zhao et al. 2020): Zhao, W., Zhang, G., Yuan, G., Liu, J., Shan, H., Zhang, S.: The Study on the Text Classification for Financial News Based on Partial Information. In: *IEEE Access* 8 (2020). pp. 100426–100437.

(Zhou and Lai 2017): Zhou, L., Lai, K. K.: AdaBoost Models for Corporate Bankruptcy Prediction with Missing Data. In: *Computational Economics* 50 (2017) 1. pp. 69–94.

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

(Zmud et al. 2010): Zmud, R., Shaft, T., Zheng, W., Croes, H.: Systematic Differences in Firm's Information Technology Signaling: Implications for Research Design. In: *Journal of the Association for Information Systems* 11 (2010) 3. pp. 149–181.

Doctoral program "Wirtschaftswissenschaften"

Assurance upon admission of the doctoral examination

I assure,

1. that I have independently prepared the submitted dissertation "Reflections on Text Mining Approaches in Corporate Failure Prediction based on German Financial Statements" and have not used the help of third parties in a manner contrary to the right of examination and scientific honesty,
2. that I have complied with examination law, including scientific honesty - this includes strict adherence to the citation requirement, so that the adoption of foreign ideas in the dissertation is clearly marked,
3. that no intermediary has been engaged for remuneration in the present doctoral procedure, and in connection with the doctoral procedure and its preparation

   - no remuneration has been paid or services equivalent to remuneration have been rendered

   - no services were used free of charge that contradict the purpose of an examination procedure

4. that I have not otherwise applied for a corresponding doctoral degree and in doing so have submitted the submitted dissertation or parts thereof.

I am aware that untruthfulness with regard to the above assurance will preclude admission to the doctoral examination and, in the event that it is subsequently discovered, the doctoral examination may be declared invalid or the doctoral degree may be revoked.

Göttingen, 27.02.2023 Tobias Nießner

Overview of Author Contribution on the Conducted Studies

Outlet	Authors	Contribution (marked Author)
Study 1. "Towards a taxonomy of AI-based methods in Financial Statement Analysis" (Nießner et al. 2021)		
AMCIS 2021 (published)	**Nießner** Nickerson Schumann	conceptualization, taxonomy development, literature review, data analysis, writing
Study 2. "Evidential strategies in financial statement analysis: A Corpus linguistic Text Mining approach to bankruptcy prediction of German Companies" (Nießner et al. 2022b)		
Journal of Risk and Financial Management (published)	**Nießner** Gross Schumann	data preparation, data analysis, conceptualization, text mining, writing
Study 3. "Consideration of the Use of Language in Corporate Bankruptcy Prediction: A data analysis on German companies" (Nießner et al. 2022a)		
PACIS 2022 (published)	**Nießner** Wiederspan Schumann	data preparation, text mining, data analysis, hypothesis development, writing
Study 4. "Analysis of consecutive financial statements concerning bankruptcy prediction" (Nießner and Schumann 2022)		
RISK 2022 (published)	**Nießner** Schumann	data analysis, evaluation, writing
Study 5. "Influence of Corporate Industry Affiliation in Financial Business Forecasting: A Data Analysis Concerning Competition" (Nießner et al. 2022c)		
AMCIS 2022 (published)	**Nießner, T.** Nießner, S. Schumann	data preparation, data analysis, text mining, feature engineering, feature selection, hypothesis development, evaluation, writing

Study 6. “Is it worth the effort? Considerations on Text Mining in AI-based Corporate Failure Prediction” (Nießner et al. 2023)		
Information (published)	**Nießner, T.** Nießner, S. Schumann	data prepation, data analysis, methodology, model building, evaluation, writing

Göttingen, 27.02.2023 Tobias Nießner