information

Article

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

Tobias Nießner *, Stefan Nießner and Matthias Schumann

Faculty of Business and Economics, University of Goettingen, 37073 Goettingen, Germany
* Correspondence: tobias.niessner@uni-goettingen.de

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

check for updates

Citation: Nießner, T.; Nießner, S.; Schumann, M. Is It Worth the Effort? Considerations on Text Mining in AI-Based Corporate Failure Prediction. *Information* **2023**, *14*, 215. https://doi.org/10.3390/info14040215

Academic Editors: Johannes Winter, Katsuhide Fujita and Ralf Krestel

Received: 13 January 2023
Revised: 13 March 2023
Accepted: 30 March 2023
Published: 1 April 2023

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?

Within this study, we follow a classical data mining process model, i.e., Cross-industry Standard Process for Data Mining (CRISP-DM) [7], 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 the development of statistical models in the context of machine learning-based bankruptcy prediction models follows in the second section. 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, highlighting contributions to research and practice and addressing limitations and approaches for future research before a conclusion is drawn.

2. Related Research

Research into the prediction of corporate failure or bankruptcy initially showed a clear trend toward the analysis of financial ratios [8]. While these models were based on stochastic methods, such as multivariate discriminant analysis in their early days [8–10], the use of machine learning algorithms is playing an increasingly important role in current research [11], 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 [12]. Due to the constant demand from practice, based on ever-increasing risks due to potential bankruptcies, the interest in the evaluation of these data is high [6]. 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 to the presentation of 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. This limitation is due to the fact that structured data would only be suitable for looking into the future by representing estimates. It should be noted here that the analysis of textual data and, insofar, 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 [13] or the readability of financial statements [14,15]. 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 [16], 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 [17]. The study by Loughran [18] 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 regarding 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 [19]. 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 [20].

In addition, initial results suggest that there is a high degree of standardization, particularly for German financial statements [11]. 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 [1].

Specifically, 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 [21]. While in current research, justified by the successful use in comparable use cases, Ensemble Learning is given a preference in algorithm selection processes; we argue for the choice of XGBoost in the following as the examined algorithm to train our model [12]. 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 failure or bankruptcy prediction.

3. Data Presentation, Understanding, and Preparation

To answer the research question, below we present the methodology based on a standardized data mining project according to CRISP-DM. Following Figure 1, we present the steps separately, starting with data selection. First, we should say that we used two basic data sets, one being a data set of 2,163,147 German Financial Statements from 2017–2021, which are in XML format provided by a practice partner and contain additional meta-information, and second, company-related information exported from Bureau van Dijk's Amadeus database matching the first data set [22].

In the first step, the meta-data were 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 1) [23,24]. Sentiment analysis was approached in two different ways. On the one hand, a general German sentiment dictionary [25] was used, and on the other hand, a context-related one [19]. Furthermore, using a translated version of a hedging dictionary, a score regarding the word choice of obfuscating terms was evaluated [26], as well as an analysis of the extent to which the text authors used passive constructions in their sentences [24]. For Part-of-Speech (POS) tagging, the spaCy Python package was used with the help of a German NLP model based on the TIGER corpus [27]. 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 submitted late, using the assigned fiscal year as a reference. In addition, the quarter in which the report was published was recorded to reflect different time cycles of publishing behavior. According to the granularity of the extracted features,

for reasons of proportionality, ratios were calculated and normalized by the total number of words in the case of words and by the total number of sentences in the case of sentences.

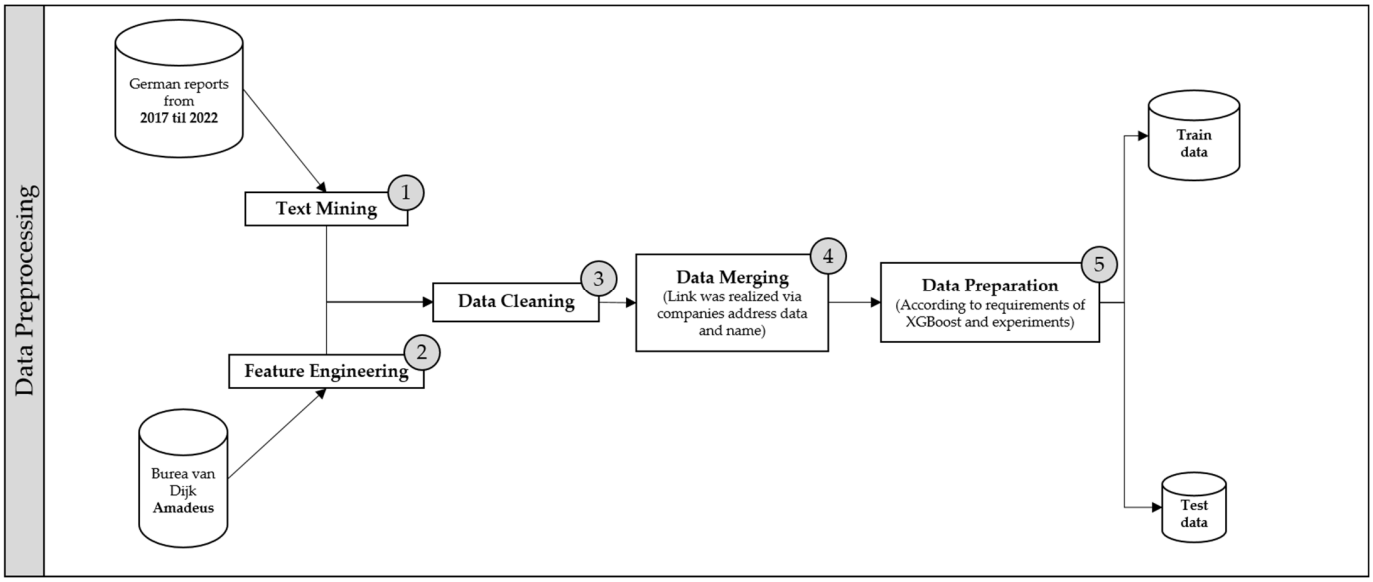

Figure 1. Summary of the methodical procedure and the associated steps.

Table 1. Feature sets for XGBoostClassifier model development.

Accounting-Based Features (F_A)	Text-Related Features	(F_T)	Company Features (F_C)
Asset coverage ratio	Meta-Information:	Dictionary approaches:	Count_branch_offices
Capital Debt ratio	Publication quarter ^{M}	Hedging score ^{D} (see [24])	Count_companies_group
Debt structure	Text_length ratio ^{D}	SentiWS_Hits_ratio ^{D}	Count_employees
Debt-to-Equity ratio	Time_difference_publication ^{M}	SentiWS_POS_Hits_ratio ^{D}	Count_shareholders
EBITDA	Character/Word matching (Count):	SentiWS_NEG_Hits_ratio ^{D}	Count_subsidiaries
Equity multiplier	Comma ratio ^{C}	SentiWS_POS_Score_ratio ^{D}	Count_CEO_change
Equity ratio	Excalamation_Marks ratio ^{C}	SentiWS_NEG_Score_ratio ^{D}	Count_CEO_last_10_years
Leverage ratio	Question_Marks ratio ^{C}	SentiWS_Score ^{D}	Company_age
Long-term liabilities	Bankruptcy_words ratio ^{W}	BPW_UNC_ratio ^{W}	Company_size
Net loss for the year	Chances_words ratio ^{W}	BPW_POS_ratio ^{W}	NACE_identification
Profit margin	Opportunities_word ratio ^{W}	BPW_NEG_ratio ^{W}	Competitor_density
Return on Assets ratio	Project_words ratio ^{W}	Semantic POS-tagging (see [23]):	Population_density
Return on Capital Employed ratio	Research_words ratio ^{W}	Evidential_Strategies_ratio^{S}	
Return on Equity ratio	Risk_words ratio ^{W}	Evidential_Strategies_Pos_ratio^{S}	
Short-term debt ratio	POS-Tagging (Count):	Evidential_Strategies_Neg_ratio^{S}	
Short-term liabilities	ADJ_ratio ^{W}	POS-Pattern matcher:	
Working capital ratio	ADP_ratio ^{W}	Bool_Deficit ^{D}	
	ADV_ratio ^{W}	All_Sentence_ratio ^{S}	
	AUXVERB_ratio ^{W}	Passive_Sentence_ratio ^{S} (see [24])	
	CONJ_ratio ^{W}		
	DET_ratio ^{W}		
	NOUN_ratio ^{W}		
	NUMERAL_ratio		
	PROPNOUN_ratio ^{W}		
	VERB_ratio ^{W}		

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

In the second step, various financial ratios (see Table 2) 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. We have chosen a limited number of financial ratios that have achieved positive results in the application context in previous research, and that can also be calculated consistently and presented transparently on the basis of the data [10,28]. In addition, supplementary information about the respective companies was collected, which dealt with the organizational and personnel level [1,29]

and the competitive level [30]. In this regard, we used additional demographic and region-specific data sets for the calculation [31,32].

In the third 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.

In the fourth 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 the fifth step, 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 [33]. For labeling the data with respect to the use case, we decided to adopt the financial situation decisions (bankruptcy, defaulting, solvent) captured 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 1.

4. 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 [1,10]. 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 three-fold cross-validation. Furthermore, the objective function "rank:pairwise" was determined using GridSearchCV [34]. 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 data set. 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 1), in addition to considering feature importance concerning text mining approaches, we decided to train different models based on differentiated training data sets in each case. Accordingly, a base model was trained based on the feature set of accounting variables, which was extended accordingly in different forms. We considered the effects of adding the text-related feature set (F_T), the company features set (F_C), and the addition of both on the performance of the prediction.

5. Evaluation

A stratified train-test-split was used to evaluate the models, with a ratio of 4:1 since it preserves the proportions of examples in each class and the data set is large enough to enable this recently common split. 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 in 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.

Furthermore, the addition of both feature groups distinguished in this study results in an additional improvement of both differentiated additions. Overall, the best model (F_A + F_T + F_C) was able to correctly identify about 81% of the reports of companies with financial problems (True-Positive-Rate = 1 − FPR) and about 74% (True-Negative-Rate = 1 − FNR) of the solvent companies (see Table 2).

Table 2. Evaluation of the differently trained XGBoost models.

	Evaluated Models			
Training Data (See Table 1)	**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 (FPR)	0.2608	0.2482	0.2119	0.1930
False-Negative-Rate (FNR)	0.3169	0.2993	0.2800	0.2559

6. 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. 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 3). 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 Section 2). 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 [20,35]. 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 of using 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 feature's content character from the language perspective, 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 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 2 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, as well as meta-information about the text, also have a high value in comparison. Here we also see that the application of the SentiWS has both a semantic and lexical component, while the Hedging dictionary, as well as the BPW dictionary, are applied purely lexically. If, on the other hand, we reflect on the results of the feature importance (see Figure 2), we see that a lexical analysis seems to be quite superior to a semantic one. One

reason for this could be the context-independent evaluation of the measured values of the words in SentiWS [25], but it suggests that such values are not blindly transferable to texts of different origins. In particular, it shows that the construction of text features in the area of syntactic analysis can indeed achieve promising results at the word and sentence level, considering the use of passive constructions, evidential strategies but also auxiliary verbs or conjunctions. When analyzing the latter, it should be questioned to what extent other parameters, such as sentence length, result from the use of conjunctions (POS-tagging).

Table 3. Classification of extracted text mining features.

		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} • Sentiment Score $^{L, Se}$ (SentiWS, BPW) • Descriptive Information ^{M}

Legend: Linguistic Analysis Level {^{N}: Non-linguistic; ^{L}: Lexical; ^{Sy}: Syntactic; ^{Se}: Semantic, ^{M}: Meta}.

6.1. 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 those approaches. Both German dictionaries, such as SentiWS (scaled polarities between 0 and 1 are available) and the context-adapted BPW (words are only classified into three 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, such as 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.

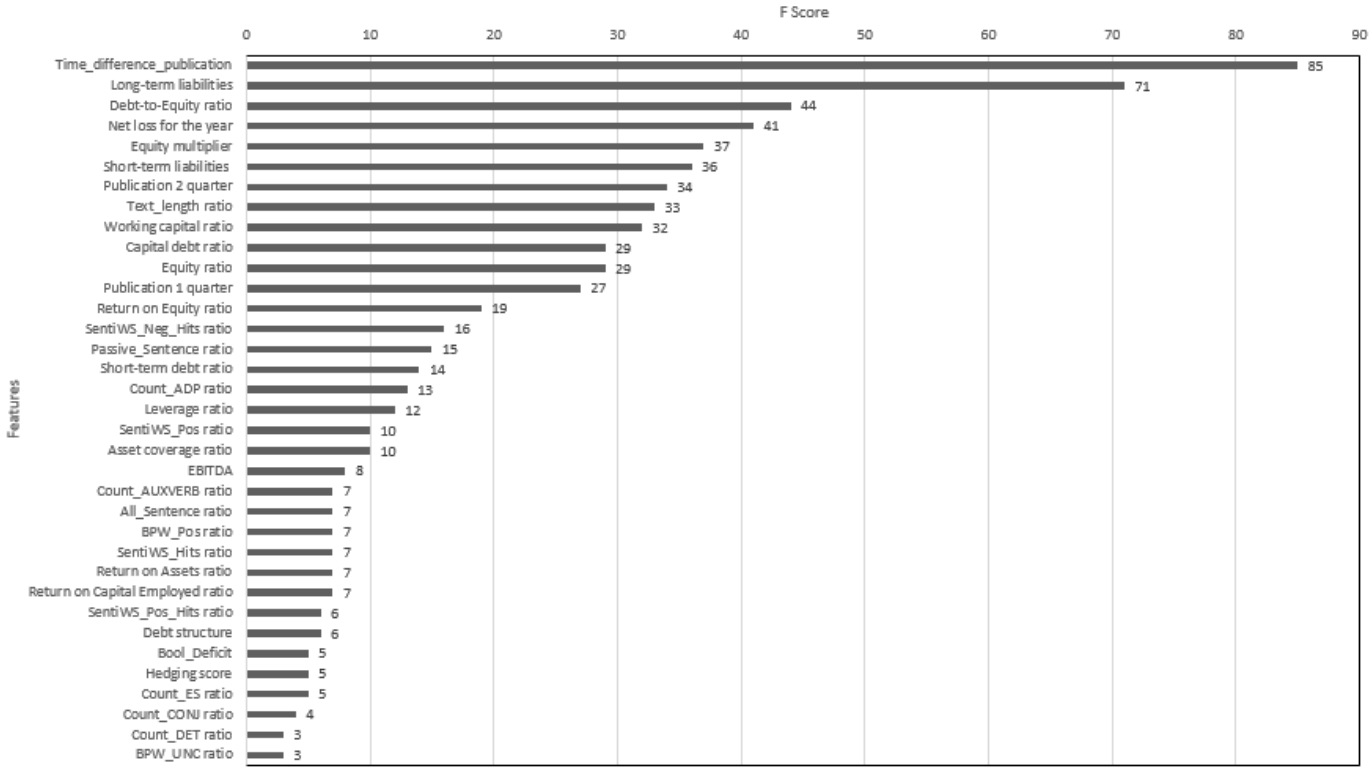

Figure 2. Feature Importance (Top 35) based on supplemented textual feature set ($F_A + F_T$).

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 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 view of a company's financial difficulties and the associated existence of various additional problems. In doing so, a prediction would only be made in response to the filing of a report and would thus be equally delayed in time.

In addition, an examination of the top five features (see Figure 3) 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, and changes in the management board. The classification options shown for text mining features are further useful for practitioners as guidance for the comparability of trained models and, in this respect, support the understanding of the development of the model. Conceivable here is, besides the identification of financially distressed companies, enabled comparisons of competitors as well as spotting of investment opportunities.

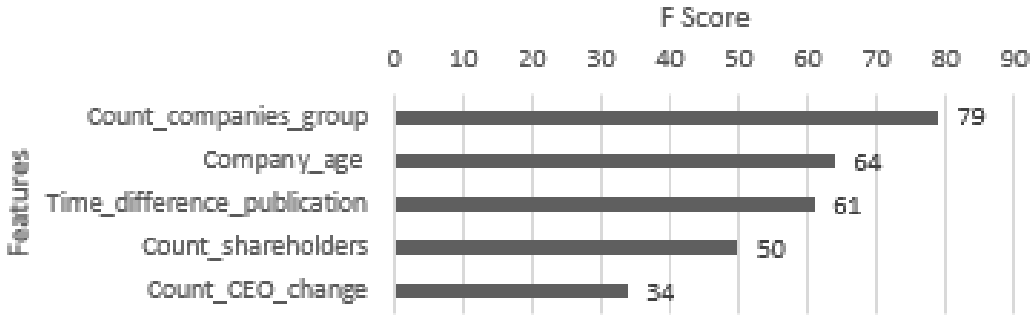

Figure 3. Feature importance (Top 5) based on all available data ($F_A + F_T + F_C$).

6.2. Limitations and Future Research Opportunities

We are aware of the fact that, certainly, other machine learning approaches exist that could have the potential to provide higher prediction accuracy. Nevertheless, we used XGBoost because it has shown promising results in recent studies [12]. 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 a practice scenario as current German data are 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 these are independent of individual characteristics of companies, which, hence, cannot be objectively compared while ignoring those companies' own development and presentation. Future research must, therefore, weigh the advantages and disadvantages of focusing on companies or financial statements. 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 our consideration to the past research and must mention how, for future research, 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.

7. Conclusions

Researchers are developing a variety of AI-based approaches to address the use case of predicting corporate failure or 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 a company. 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 show that the contextual extraction of features from financial statements 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, the relative number of polarized hits (positive, negative, or uncertain classified) seems to add more value to the model than a calculated 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 semantics, 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. Regarding meta-information, it is also worth mentioning that studies already exist that deal with the analysis of the change in text mining features in consecutive reports [24]. Such features, which can be classified as beyond-document, can also help to explore causalities behind identified correlations and to specify the extraction of textual metrics.

Author Contributions: Conceptualization, T.N.; methodology, T.N.; formal analysis, T.N. and S.N.; data curation, T.N., S.N. and M.S.; writing—original draft preparation, T.N. and S.N.; writing—review and editing, T.N., S.N. and M.S.; visualization, T.N. and S.N.; validation, T.N. and S.N.; project administration, M.S.; All authors have read and agreed to the published version of the manuscript.

Funding: We acknowledge support by the Open Access Publication Funds of the Göttingen University.

Data Availability Statement: Partial publicly available data sets were analyzed in this study. This data can be found here: https://www.bvdinfo.com/en-us/our-products/data/international/amadeus, accessed on 25 July 2022. The financial statement data are not publicly available due to data protection reasons of an external partner.

Acknowledgments: We would like to thank the reviewers for their thoughtful comments and efforts towards improving our manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Jones, S. Corporate bankruptcy prediction: A high dimensional analysis. *Rev. Account. Stud.* **2017**, *22*, 1366–1422. [CrossRef]
2. Kloptchenko, A.; Eklund, T.; Karlsson, J.; Back, B.; Vanharanta, H.; Visa, A. Combining data and text mining techniques for analyzing financial reports. *Intell. Syst. Account. Financ. Manag.* **2004**, *12*, 29–41. [CrossRef]
3. Nassirtoussi, A.; Aghabozorgi, S.; Ying Wah, T.; Ngo, D.C.L. Text mining for market prediction: A systematic review. *Expert Syst. Appl.* **2014**, *41*, 7653–7670. [CrossRef]

4. Schumaker, R.P.; Zhang, Y.; Huang, C.-N.; Chen, H. Evaluating sentiment in financial news articles. *Decis. Support Syst.* **2012**, *53*, 458–464. [CrossRef]
5. Kloptchenko, A.; Magnusson, C.; Back, B.; Visa, A.; Vanharanta, H. Mining Textual Contents of Financial Reports. *Int. J. Digit. Account. Res.* **2004**, *4*, 1–29. [CrossRef]
6. Veganzones, D.; Severin, E. Corporate failure prediction models in the twenty-first century: A review. *Eur. Bus. Rev.* **2021**, *33*, 204–226. [CrossRef]
7. Wirth, R.; Hipp, J. CRISP-DM: Towards a standard process model for data mining. In Proceedings of the Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining, Manchester, UK, 11–13 April 2000; pp. 29–39.
8. Altman, E.I. Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *J. Financ.* **1968**, *23*, 589–609. [CrossRef]
9. Altman, E.I.; Haldemann, R.G.; Narayanan, P. ZETA™ analysis: A new model to identify bankruptcy risk of corporations. *J. Bank. Financ.* **1977**, *10*, 29–54. [CrossRef]
10. Ohlson, J.A. Financial Ratios and the Probabilistic Prediction of Bankruptcy. *J. Account. Res.* **1980**, *18*, 109. [CrossRef]
11. Kirkos, E. Assessing methodologies for intelligent bankruptcy prediction. *Artif. Intell. Rev.* **2015**, *43*, 83–123. [CrossRef]
12. Lohmann, C.; Ohliger, T. Bankruptcy prediction and the discriminatory power of annual reports: Empirical evidence from financially distressed German companies. *J. Bus. Econ.* **2020**, *90*, 137–172. [CrossRef]
13. Caserio, C.; Panaro, D.; Trucco, S. Management discussion and analysis: A tone analysis on US financial listed companies. *Manag. Decis.* **2020**, *58*, 510–525. [CrossRef]
14. Le Maux, J.; Smaili, N. Annual Report Readability And Corporate Bankruptcy. *J. Appl. Bus. Res.* **2021**, *37*, 73–80. [CrossRef]
15. Ajina, A.; Laouiti, M.; Msolli, B. Guiding through the Fog: Does annual report readability reveal earnings management? *Res. Int. Bus. Financ.* **2016**, *38*, 509–516. [CrossRef]
16. Bjornsson, C.H. Readability of Newspapers in 11 Languages. *Read. Res. Q.* **1983**, *18*, 480. [CrossRef]
17. Loughran, T.I.M.; McDonald, B. When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks. *J. Financ.* **2011**, *66*, 35–65. [CrossRef]
18. Loughran, T.I.M.; McDonald, B. Measuring Readability in Financial Disclosures. *J. Financ.* **2014**, *69*, 1643–1671. [CrossRef]
19. Bannier, C.; Pauls, T.; Walter, A. Content analysis of business communication: Introducing a German dictionary. *J. Bus. Econ.* **2019**, *89*, 79–123. [CrossRef]
20. Wambsganss, T.; Engel, C.; Fromm, H. Improving Explainability and Accuracy through Feature Engineering: A Taxonomy of Features in NLP-based Machine Learning. In Proceedings of the ICIS 2021, Austin, TX, USA, 12–15 December 2021.
21. 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. *IEEE Access* **2020**, *8*, 100426–100437. [CrossRef]
22. van Dijk, B. Amadeus Database. Available online: https://www.bvdinfo.com/de-de/unsere-losungen/daten/international/amadeus (accessed on 3 November 2021).
23. Nießner, T.; Gross, D.H.; Schumann, M. Evidential Strategies in Financial Statement Analysis: A Corpus Linguistic Text Mining Approach to Bankruptcy Prediction. *J. Risk Financ. Manag.* **2022**, *15*, 459. [CrossRef]
24. Nießner, T.; Wiederspan, O.; Schumann, M. Consideration of the Use of Language in Corporate Bankruptcy Prediction: A data analysis on German Companies. In Proceedings of the PACIS 2022, Virtual, 5–9 July 2022.
25. 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, 17–23 May 2010; pp. 1168–1171.
26. Humpherys, S. Discriminating Fraudulent Financial Statements by Identifying Linguistic Hedging. In Proceedings of the AMCIS 2009 Proceedings, San Francisco, CA, USA, 6–9 August 2009.
27. 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. *Res. Lang Comput.* **2004**, *2*, 597–620. [CrossRef]
28. Pamuk, M.; Grendel, R.O.; Schumann, M. Towards ML-based Platforms in Finance Industry—An ML approach to Generate Corporate Bankruptcy Probabilities based on Annual Financial Statements. In Proceedings of the 32nd Australasian Conference on Information Systems, Sydney, Australia, 6–10 December 2021; pp. 1–12.
29. Brédart, X.; Séverin, E.; Veganzones, D. Human resources and corporate failure prediction modeling: Evidence from Belgium. *J. Forecast.* **2021**, *40*, 1325–1341. [CrossRef]
30. Nießner, T.; Nießner, S.; Schumann, M. Influence of corporate industry affiliation in Financial Business Forecasting: A data analysis concerning competition. In Proceedings of the AMCIS 2022 Proceedings, Minneapolis, MN, USA, 10–14 August 2022.
31. 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 23 January 2022).
32. Zensus. Die Ergebnisse des Zensus. Available online: https://ergebnisse2011.zensus2022.de/datenbank/online/ (accessed on 24 January 2022).
33. 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 17 February 2022).

34. SciKit Learn. Machine Learning in Python. Available online: https://scikit-learn.org/stable/ (accessed on 28 January 2022).
35. Fromm, H.; Wambsganss, T.; Söllner, M. Towards a taxonomy of text mining features. In Proceedings of the 27th European Conference on Information Systems, Uppsala, Sweden, 8–14 June 2019; pp. 8–14.