Data-driven decision-making in credit risk management: The information value of analyst reports

Jan Roeder [a], Matthias Palmer [a], Jan Muntermann [a,b,*]

[a] *Faculty of Business and Economics, University of Goettingen, Germany*
[b] *Faculty of Business and Economics, University of Augsburg, Germany*

ABSTRACT

Other than banks, non-financial companies also continuously monitor and analyze their credit risk exposure to avoid possible counterparty defaults. Credit default swaps are commonly used financial instruments that provide information on a counterparty's creditworthiness. Although this metric can provide crucial insights, the underlying price dynamics often remain unknown and require further explanation. Data-driven decision-making is a key concept for identifying these reasons and supporting and justifying decisions. In this paper, we provide such justifications by applying sentiment and topic analysis to company-related financial analyst reports. While the contents of financial news have been analyzed in the past, analyst reports can offer additional insights, as seasoned analysts use them to disseminate in-depth research to experienced investors. This analysis examines 3386 analyst reports covering constituents of the Dow Jones Industrial Average Index in the period from 2009 to 2020. The results suggest that even when established credit risk indicators and financial news are considered, the sentiment and a subset of topics are correlated with changes in the credit default swap spread, indicating a fundamental relationship between quantitative risk metric and analyst reports. We find that analyst reports contain information related to the change in credit default swap spreads, an insight that helps to improve our understanding of existing risk assessments. The outcome indicates that banks or corporate risk managers can benefit from complementing established financial metrics and even financial news data with new insights derived from analyst reports.

1. Introduction

Banks and finance departments regularly find themselves in a position where it is necessary to assess the credit risk of counterparties on an ongoing basis, i.e., the risk that a party cannot service its debt obligations [1]. For this purpose, in addition to more fundamental accounting measures, credit derivatives such as the spread of credit default swaps (CDS) are used [2]. It serves as an indicator of creditworthiness because the CDS spread expresses the market-based cost of insuring against a negative credit event (i.e., the default of an entity) [3]. Compared to credit ratings, this risk indicator is market-based and therefore updated at a higher frequency.

For decision-makers assessing credit risk, it is desirable to supplement quantitative risk measures with qualitative information [4]. Text mining can help bridge this gap and identify patterns in large and heterogeneous data [5]. The practical relevance of this challenge is evident because this research project was initiated in cooperation with a globally operating constituent of the STOXX 50 index. During several interviews and a use case definition, it became clear that it is of high practical relevance to identify subjects associated with CDS spread movements. Therefore, this study analyzes how unstructured data can provide decision support for credit risk assessments and thereby support "data-driven decision-making" (DDD) [6]. This also provides implications for the model component of a decision support system used to augment the decision-making process [7].

In the finance context, unstructured or semi-structured data sources, such as company reports, forms 10-K, quarterly conference calls, financial news, social media, and analyst reports are open for analysis and can provide meaningful insights [8]. This study focuses on analyst reports created by recognized industry experts with several years of professional experience, who act as intermediaries between companies and investors. They address market environments, and their work has gained a positive reputation in the capital market literature [9,10].

The intersection of text data with CDS spreads and credit ratings has

been addressed in different ways. Liebmann et al. [11] analyzed how prices react to financial news and identified words that significantly impact the decisions of stock or CDS traders. Lu et al. [12] combined structured and unstructured data sources and used news from the Wall Street Journal and financial ratios to model credit rating changes. Following Liebmann et al. [11] and Galil and Soffer [13], this analysis focuses on explaining CDS spread changes. To our knowledge (see 2.3 for more details), textual analyst reports (i.e., extracted sentiment and topics) have not yet been used to explain CDS spreads. Additionally, we explore whether a relationship persists even when news sentiment, which was used in previous studies, is considered. Against this background, we aim to answer the following research questions:

- **RQ 1**: To what extent are the sentiment and topics contained in analyst reports useful for explaining CDS spreads and can therefore support data-driven decision-making in credit risk management?
- **RQ 2**: If analyst reports are useful for explaining CDS spreads, which topics are related to CDS spread changes, and what properties do they exhibit?

The paper begins with an introduction to credit risk management, data-driven decision-making, and financial analysts. Subsequently, the research design, including the dataset and model specification, is defined. The empirical analysis uses panel regression to connect the CDS spread, quantitative variables, and measures extracted from analyst reports. Finally, the implications of our findings and their limitations are discussed. After concluding, we highlight future research directions.

2. Theoretical background

2.1. Credit risk management and credit default swaps

It is essential for a company to manage the financial risks it is exposed to and to hedge these risks at all times. McNeil et al. [14] name four tasks of credit risk management: 1) determining the capital requirement to absorb credit risks, 2) monitoring the credit risks inherent to the balance sheets, 3) portfolio monitoring of traded credit derivatives, and 4) assessing the risk exposure from contracts and trades with counterparties. These essential tasks create the demand for credit scoring and forecasting of financial distress to identify high-risk counterparties [15]. Risk monitoring can use ratings from rating agencies or credit derivative prices [2]. CDS contracts, which are credit derivatives, can be understood as insurance against specific credit events, such as bankruptcy and payment default [16]. They link the payment of premiums to the creditworthiness of an entity, for example, a company or country. Longstaff et al. [3] provide an example that illustrates the mechanisms of a CDS: the buyer of the protection (i.e., the CDS) buys insurance against the default of a bond (reference obligation) issued by a company (reference entity) by paying a recurring premium (i.e., the CDS spread). The seller of the derivative agrees to buy the bond at face value in case of a default.

The term duration of CDS ranges from a few months to several years. The most common time horizon is five years [3]. Retail and investment banks usually act as protection buyers and hedge funds and investment banks take the position of insurance sellers [14]. Among credit events such as bankruptcy, failure to pay, obligation default, and obligation acceleration, restructuring is the most controversial and most discussed credit event [17]. The International Swaps and Derivatives Association (ISDA) defines which restructuring convention is used depending on local laws for bankruptcy [16].

2.2. Financial analysts

The fundamental task of financial analysts is to analyze companies and evaluate their financial statements to provide investment recommendations to other market participants. Essentially, financial analysts use their expertise to develop purchase recommendations, company valuations, earnings forecasts, and future stock price estimates of specific companies [18]. Analysts are generally credited with two major roles: information discovery and information interpretation [19]. Analysts fill these roles by evaluating publicly available information with their expertise. They also use direct contact with company representatives to provide new information to the market. For example, they discuss product-specific topics with the middle management or contact top-level management to learn about strategic goals [20]. The role of analysts can, therefore, be considered an indispensable intermediary between companies and investors.

A study by Huang et al. [21] sees the most important added value of financial analysts in interpreting information. The study provides three reasons why financial analysts' work is beneficial: First, analysts help investors filter unimportant information in corporate disclosures. Second, analysts help interpret the statements of the managers. Finally, analysts should be independent to be able to question the credibility of executives' statements. In most cases, analyst reports are published before and after financial statement data [21]. In addition to balance sheets and company telephone calls, analysts can take advantage of their years of experience and knowledge of capital market specificities. This extensive experience is also reflected in the depth and scope of analyst reports, specifically when compared to conventional financial news. Nevertheless, whether prices on capital markets follow analysts' opinions is partially questioned in the scientific literature [22,23]. The qualitative explanations in analyst reports provide an additional source of information that is particularly valuable for the assessment of the current situation and future development of a company, especially concerning credit risks [24].

2.3. Intersection of unstructured data and credit risk management

The intersection of unstructured data and credit risk has been analyzed in multiple disciplines, ranging from finance to information systems to computer science. In a literature review, Roeder [25] identified studies that examined credit risk in connection with unstructured data, e.g., financial news, 10-K filings, social media posts, or even search engine queries. The two most commonly used risk metrics are a categorical classification and CDS spreads [25]. However, no paper was identified that relates the textual content of analyst reports to CDS spreads. Related to our study is the research by Bao and Datta [26] who developed a topic model to identify risk types from 10-K forms. Another study analyzed the coverage and sentiment of financial news and found that it provides incremental value to the content of 10-K forms [27]. Wei et al. [28] identified 21 bank risk factors based on the analysis of 10-K forms using a custom semi-supervised model. Noteworthy is the work of Huang et al. [21] who examine the added value of analyst reports to quarterly earnings announcements. The relevance of financial news in the analysis of CDS spread changes is indicated by Smales [29]. Additionally, based on financial news data, a strong negative relationship between news sentiment and CDS spreads can be ascertained [30]. While analyst price targets have been linked to CDS spreads [31], we are not aware of research that analyzes the text-based sentiment or content of analyst reports in a credit risk context. Considering differences between analyst reports and financial news regarding 1) the expertise of the author, 2) the audience of the research, and 3) the depth of the analysis, it is reasonable to hypothesize analyst reports could provide incremental value. Overall, the textual content of analyst reports and their use for credit risk management is still underexplored and represents a research gap.

2.4. Data-driven decision-making

It has been shown that appropriate analytical techniques can draw meaningful conclusions from the data. The enormous speed at which economic and social transactions can be recorded, stored, and made

available digitally is novel [5]. The resulting large amounts of data represent a challenge, as data are diverse, structured, unstructured, and constantly supplemented by new data types and sources. This development offers the potential to advance data-driven decision support systems, i.e., systems that use internal and external time series data for retrospective and predictive data analysis [32,33]. Facilitating the processing and manipulation of unstructured data is particularly important because knowledge workers are not replaced by technology in many cases. Rather, their work is augmented and the decision support systems enable a more profound analysis of the task at hand [7]. DDD magnifies the potential of organizational data collection and affects how corporate strategy processes are shaped [34]. The diverse and often large amounts of data disrupt the traditional information value chain and bring the processing and analysis of alternative data sources to the forefront of corporate decision-making [35].

Research finds that companies that consistently use DDD are generally more successful than their competitors [6,36]. It has also been observed that, on average, long-standing multi-unit companies switch to DDD earlier than young single-establishment companies [37]. Grover et al. [38] list example areas in which companies carry out analytical initiatives and provide decision support for management. These examples include identifying the root causes of outages in near real-time, anomaly detection, or the refinement of in-house processes. Furthermore, Davenport [39] emphasizes that companies can gain a decisive competitive advantage by analyzing in-house and external data. In our view, the advantages of DDD should be leveraged to assess credit risk.

3. Research design

3.1. Analysis setup

The analysis setup in this study (Fig. 1) is based on the data mining process [40,41]. To account for the specifics of text mining, we divide the transformation step from the data mining process into feature extraction and feature representation. Feature extraction transforms the text into a numerical format, and feature representation processes and transforms the numerical representation (e.g., weighting). Therefore, the analytical approach is guided by the following six steps: (1) data selection, (2) pre-processing, (3) feature extraction, (4) feature representation, (5) data analysis, and (6) interpretation and evaluation.

In the first step, appropriate data sources and subsets are chosen (Section 3.2). *Pre-processing* includes the preparation of text data, such as phrase detection (Section 3.3). *Feature extraction* helps to put the text into a structure that can be automatically processed and analyzed (Section 3.3). The *feature representation* step in Section 3.4 transforms the extracted features, that is, calculating the sentiment and estimating the topic distribution. Additionally, transformations for the quantitative

measures were applied (Section 3.5). In the *analysis* step, the variables are linked and analyzed (Sections 3.6 to 4.3). Finally, the results were *interpreted* and *evaluated* (Section 4.4).

3.2. Data set

The analyst reports stem from companies that are part of the Dow Jones Industrial Average (DJIA) index as of 01/01/2014, i.e., the 30 largest US companies. These companies are important counterparties and their CDS spreads are sufficiently liquid. The reports under analysis were obtained from Refinitiv and ranged from June 2009 to December 2020. This includes the aftereffects of the global financial crisis (GFC), the phase of economic recovery in recent years, and the height of the COVID-19 crisis. The GFC was defined as the period from August 2007 to June 2009 [42,43]. The chosen time period is a compromise between the coverage of different macroeconomic conditions and data availability. This initial selection resulted in a dataset of 28,784 analyst reports.

CDS spreads were obtained from Refinitiv EIKON [44]. Following established literature, senior CDS spreads with a maturity of five years were used [45]. For North American companies, the no-restructuring (XR14) clause has been prevalent since April 2009. No CDS data were available for *E. I. du Pont de Nemours and Company*, *United Technologies*, and *Visa Inc.* These companies were excluded from the analysis. Following Das et al. [46], financial companies are excluded because their fundamentals are not comparable in a credit risk context. Therefore, 22 out of 30 constituents of the DJIA (shown in Table A.1) can be linked to CDS Spreads and analyst reports. For each earnings announcement date, the CDS spread is linked to analyst reports published up to ten days before the earnings date. Therefore, the included reports represent the analyst's initial assessment without being influenced by earnings results, which in turn helps to prevent endogeneity issues. The final selection step resulted in 3386 analyst reports containing 1,767,488 words.

Sentiment data extracted from financial news is used to examine whether the results of the analyst report analysis persist when financial news are taken into account. The data was obtained from the Ravenpack News Analytics database. Selecting data points with an entity relevance of 100 and a novelty of more than 50 (out of 100) resulted in 14,215 data points.

3.3. Document pre-processing and feature extraction

The first step is to transform analyst reports into a standardized format. We obtained the reports as PDF files and transformed them into Excel files, where each paragraph is represented as an Excel cell. This structure enables detailed filtering. The heuristics in Fig. 2 yielded the most consistent results for removing residuals of non-essential
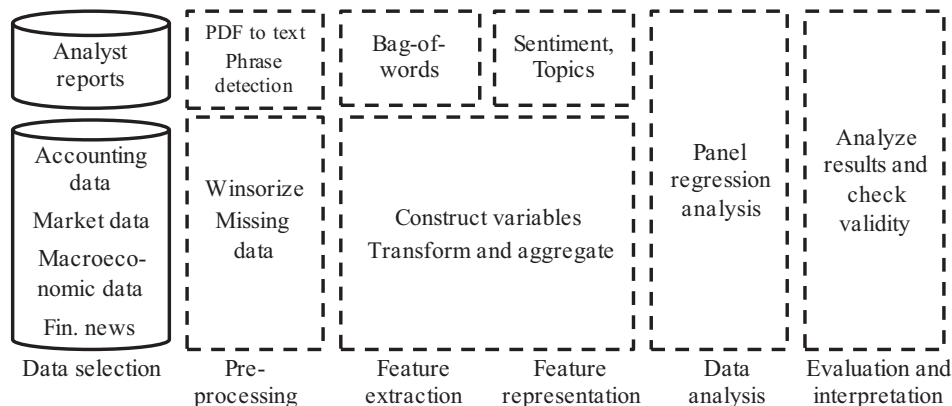

Fig. 1. Analysis setup overview.

Word count > 4	Word count > (number count × 2)	Word count > (punctuation × 1.2)	Word count > (whitespaces × 3)	Not starting with "disclaimer" etc.

Fig. 2. Necessary condition per paragraph to not be removed.

boilerplates and tables. The first condition, that is, the minimum word count, is necessary to remove figure captions or table residuals from the original PDF. The ratio of words to numbers and punctuation can identify remnants of tables or formulas. A cluster of spaces is another accurate indicator of table remnants for the file format used. Finally, the phrase "disclaimer," if positioned at the beginning of a paragraph, also indicates the presence of a disclaimer.

To remove duplicates, reports were compared on a paragraph basis. Those that exceeded a similarity threshold of 70% were removed. As shown in Fig. 3, *raw text* is lowercased, whitespaces are removed, and company- and broker-specific phrases are removed. Then the text is *tokenized* and lemmatized, and the numbers and stop words are removed. Commonly occurring two-word *phrases* (i.e., bigrams) are identified using a bigram score, which was proposed by Mikolov et al. [47] and implemented by Řehůřek and Sojka [48]. This helps to keep the dimensionality of the resulting term-document matrix at a moderate level. The identified phrases were concatenated with an underscore.

For *feature extraction*, the text is transformed into numerical features using a bag-of-words (BoW) vector representation. Regarding the maximum document frequency, we start from the assumption that a word that is a distinct indicator of the credit risk level likely does not appear in more than half of the documents. The maximum document frequency was set to 50% and is analyzed per individual company. First, this helps to remove terms that frequently occur in the whole corpus, that is, corpus-specific stop words. Second, this approach removes terms that apply to a single firm but cannot capture the broader risk environment. To further validate this value, we analyze which words would be removed if a value of 45% was used instead (5% is a common step size). Since this choice would result in the elimination of the following words, which we believe are important to assess the credit risk situation, a value of 50% is chosen: *gain*, *strength*, *offset*, *negative*, *supply chain*, *strategy*, *pressure*, and *restructuring*. The resulting maximum document frequency is 39.61% across all companies.

3.4. Sentiment and topic analysis

The goal of sentiment analysis is to capture the mood expressed in a text. It is a complex task, as it can be necessary to understand the syntactic, semantic, and pragmatic layers of text [49]. Dictionary-based approaches rely on lists of words assigned to a specific category. Typical categories are "positive" or "negative," and they can be extended further [8]. Sentiment dictionaries can be developed for different types of texts [50,51]. Machine learning models tend to be trained in a supervised manner, that is, labeled data are used to train and evaluate the model.

We use the state-of-the-art FinBERT transformer model to determine the sentiment of analyst reports [52]. The architecture corresponds to that of the well-established BERT model [53]. FinBERT has been trained on financial texts, including analyst reports. The paper shows a higher accuracy for financial data compared to BERT. Before using FinBERT, it is vital to understand the model's accuracy. Palmer et al. [54] manually assigned three categories to 1904 randomly sampled sentences from analyst reports of companies in the DJIA. This was the class distribution: positive (819), neutral (668), and negative (417). To account for class imbalance, a micro-averaged *F1* score of 75.7% was reported. This is significantly higher than the commonly used dictionary by Loughran and McDonald [8], which was evaluated and achieved 47.8%. For the following analysis, three labels are assigned using the uncased FinBERT model: negative (-1), neutral (0), and positive (1). For each sentence in each document, the numerical value is assigned based on the largest unnormalized log probability, that is, the output value of the last layer of the model. Finally, the mean sentiment polarity per document is calculated, resulting in a numerical measure of the sentiment of each analyst report.

Topic modeling represents the content of documents using latent topics. A preferred method is the generative and probabilistic model latent Dirichlet allocation (LDA) [55]. We are interested in the distribution of topics for each document d (θ_d), in this case, an analyst report, and the word distribution for topic k (β_k). θ characterizes the meaning of each document by assigning topics and β is used to interpret the inferred topic. Both of these random variables are Dirichlet distributed, which helps to prevent overfitting compared with prior approaches [55]. The detailed generative process assumed by the LDA model is defined in Blei et al. [55]. On an abstract level, the imagined generative process creating each document d can be roughly expressed as follows [56]:

1. Randomly choose a distribution over topics
2. For each word in the document
   a) Randomly choose a topic from the distribution over topics in step #1
   b) Randomly choose a word from the corresponding distribution over the vocabulary

This process highlights that each document is understood as a mixture of multiple latent topics [56]. This view aligns with our intent to create a better understanding of the contents of analyst reports, which also address multiple topics simultaneously, such as financial and strategic aspects. Before the distributions can be estimated, it is necessary to define K, α, and η. K defines the number of topics, α is the prior distribution for θ, and η is the prior distribution for β. The process of identifying the appropriate number of topics K is described in detail in Section 4.2. For the prior distributions, the initial *MALLET* [57] configuration was used. Finally, posterior estimation of the LDA model was performed. The modified Gibbs sampling-based approach of *MALLET* was employed to approximate the distributions. For practical implementation, we want to emphasize that *MALLET* yielded noticeably superior results compared
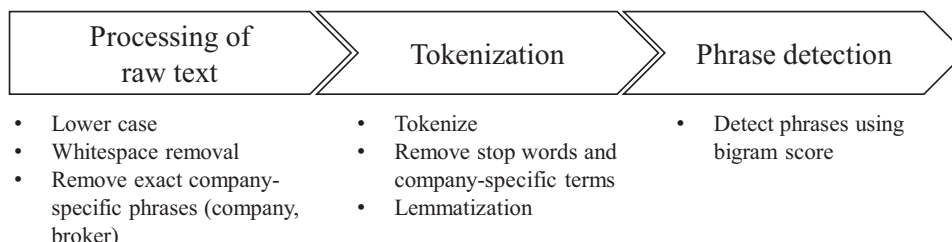
Processing of raw text	Tokenization	Phrase detection
• Lower case • Whitespace removal • Remove exact company-specific phrases (company, broker)	• Tokenize • Remove stop words and company-specific terms • Lemmatization	• Detect phrases using bigram score

Fig. 3. Pre-processing steps in detail.

to *gensim* [48] in essentially all the configurations that we analyzed.

3.5. Variable construction

For the dependent variable, that is, the CDS spreads, theoretical considerations suggest that a transformation using the natural logarithm is appropriate [46]. However, the following assumptions are necessary: The premium payments from the buyer to the seller are equal to the expected present value of the payment exchanged in the case of a default. Furthermore, the CDS spread is influenced by variables such as interest rates, default intensities, and recovery rates. Das et al. [46] note that empirical studies show that a better model fit is achieved using the natural logarithm [58].

The control variables represent risk factors at the corporate, market, and macro levels. This helps to ensure that established quantitative variables do not already reflect the findings from analyst reports. In choosing and transforming the indicators, we are guided by the comprehensive studies of Das et al. [46] and Tsai et al. [27]. The data were primarily obtained from Refinitiv Datastream, and the risk-free rate was retrieved from the US Department of the Treasury. The different variables and a brief definition are listed in Table 1. In addition, the following characteristics must be considered. The rolling four-quarter average for *return on assets* and *revenue growth* is used to reduce the impact of seasonal effects [46]. The calculation of the naïve distance to default (*dtd*) is based on the bond-pricing model of Merton [59]. A firm's equity can be understood as a call option on the value of a firm, where the strike price is equal to the face value of the firm's debt [60]. Neither the underlying value of the firm nor its volatility is directly observable [60]. Therefore, these measures are typically derived from the equity value and other observable variables by iteratively solving a system of nonlinear equations [60]. Bharath and Shumway [60] proposed a naïve *dtd* measure and empirically demonstrated a high correlation with the traditional measure combined with a stronger predictive performance.

The market value of a firm's debt is approximated by its value (*F*). The total firm volatility *(naïve σ_v)* also depends on the equity value (*E*) and equity volatility (σ_E). The stock return of the previous year, r_{it-1}, was also incorporated. *T* is the forecasting horizon, which is set to one year.

The total firm volatility (naïve σ_v) is defined as shown in Eq. (1):

$$\text{naïve } \sigma_V = \frac{E}{E+F}\sigma_E + \frac{F}{E+F}\ (0.05+0.25*\sigma_E) \tag{1}$$

The naïve *dtd* is then defined as Eq. (2) describes:

$$\text{naïve } dtd = \frac{ln[(E+F)/F] + \left(r_{t-1} - 0.5\text{ naïve }\sigma_V^2\right)T}{\text{naïve }\sigma_V\sqrt{T}} \tag{2}$$

3.6. Model specification

The panel regression model is chosen to account for the hierarchical structure of the data at the firm and year levels. It connects the risk metric to established market-based variables and the sentiment and topics extracted from analyst reports. The suitability of the panel model was also demonstrated in related research. For example, Smales [29] examined the relationship between risk measures and non-scheduled news events using panel regression. Tsai et al. [27] analyzed news coverage and risk disclosures and their relationship to CDS spreads. In both cases, cluster-robust standard errors were used in conjunction with the panel model. Since the goal is to determine whether the qualitative data contained in analyst reports can provide additional insights in the context of credit risk management, accounting for these quantitative measures is essential.

As discussed in Section 3.5, the inclusion of CDS spreads transformed with the natural logarithm is theoretically justified and empirical results indicate that it provides good explanatory power. We want to reiterate that the focus of this study is to analyze how an established credit risk metric relates to insights from a qualitative data source. The predictive component was not at the forefront of this study. The quantitative variables were winsorized at the 1% level. Eq. (3) shows the regression model. For brevity, the independent variables are included as a vector per category. The same applies to the topics. α_i is the unobservable individual-specific effect, and λ_t is the corresponding time-specific effect. Additionally, the standard errors are adjusted to account for clustering in firms and years (clustered standard errors). The model was implemented using Python package *linearmodels*.

$$\begin{aligned} log(CDS_{it}) = {}&\beta_1^T \text{ACCOUNTING}_{it} + \beta_2^T \text{MARKET}_{it} + \beta_3^T \text{MACRO}_{it} + \beta_4^T \text{TOPICS}_{it} \\ &+ \beta_5 \text{AnalystSentiment}_{it} + \beta_6 \text{NewsSentiment}_{it} + \alpha_i + \lambda_t + \epsilon_{it} \end{aligned} \tag{3}$$

4. Empirical results

4.1. Descriptive statistics

Table 2 shows the number of data points and the mean, median, first, and third quartiles to provide an overview of the distribution of the quantitative variables. The data are shown after performing the respective transformations. For example, in the case of the CDS spread, the natural logarithm has already been applied. Overall, 646 fiscal quarters across all firms were analyzed. Table 2 shows that analyst report's mean sentiment is close to zero after the standardization, as expected.

To further improve our understanding of the relationship between independent variables and CDS spread, the data was split into four

Table 1

Description of the used variables, mainly following Das et al. [46]. Includes accounting-, market-, macroeconomic-, and text-based measures.

Accounting	Description	Exp. sign
Return on assets (*roa*)	Percentage that represents the income after taxes for the past twelve months divided by the average total assets	−
Revenue growth (*rg*)	Change from period to period in trailing twelve months revenue in percent	−
Leverage (*lev*)	The ratio of total debt to total assets	+
Retained earnings (*earn*)	The ratio of retained earnings to total assets	−
Net income growth (*nig*)	Net income growth normalized by the total assets	−
Market	**Description**	**Exp. sign**
Equity return (*ret*)	Annualized 100 trading day equity return	−
Equity volatility (*vola*)	Annualized 100 trading day equity volatility	+
Index return (*index*)	Prior year S&P 500 return	−
Distance to default (*dtd*)	Distance to default model ("naïve") based on the functional form Merton distance to default model [60]	−
Macroeconomic	Description	
Risk-free rate (*rfr*)	The 3-Month constant maturity US Treasury bill rate	−
Credit rating (*rating*)	The long-term issuer credit rating assigned by Standard and Poor's. The ordinal scaled ratings are transformed to the range from 0 to 1 following [61] according to the following schema: AAA (0), AA+ (0.056), …, D (1).	+
Textual	Description	
Analyst sentiment	Averaged analyst sentiment determined using the FinBERT transformer model (Z-score)	−
Topics	Predicted topic distribution based on LDA, which is estimated using optimized Gibbs sampling	+ or −
News sentiment	Composite sentiment score from the Ravenpack database (Z-score)	−

Table 2
Descriptive statistics for relevant regression variables.

	Count	Mean	Median	First quartile	Third quartile
CDS spread	646	3.608	3.565	3.282	3.931
Return on assets	646	0.101	0.092	0.064	0.134
Revenue growth	646	0.026	0.021	−0.023	0.06
Leverage	646	0.62	0.595	0.477	0.744
Retained earnings ratio	646	0.488	0.412	0.164	0.738
Net income growth	646	−0.001	0	−0.004	0.003
Stock return	646	0.159	0.15	−0.005	0.33
Stock volatility	646	0.198	0.178	0.144	0.228
Index return	646	0.123	0.132	0.064	0.186
Distance to default	646	14.168	13.972	10.422	17.609
Risk-free rate	646	0.489	0.11	0.04	0.52
Credit rating	646	0.188	0.167	0.111	0.278
Analyst sentiment	646	−0.001	0.01	−0.586	0.637
News sentiment	646	−0.007	0.096	−0.36	0.521

segments of equal size. The split threshold is determined based on the three quartiles, that is, the 25%, 50%, and 75% quantile. Table 3 provides an overview of the average values per CDS spread segment. This overview can help create a first intuition regarding the relationship between CDS spread and the respective variables.

For example, for *return on assets*, the first three segments do not provide a clear signal regarding the relationship with CDS spreads. However, in critical cases, where the CDS spread is high, the data clearly indicates that the *return on assets* tends to be low. In the case of *leverage*, the data suggest that the average values in the second and third segment are close. However, for the fourth segment, we observe a higher *leverage* on average. The example of *credit ratings* shows a positive correlation. In particular, a high CDS spread is associated with a substantially higher *credit rating* (i.e., higher risk). For the *analyst sentiment* variable, it is apparent that a higher CDS spread is associated with a lower sentiment value. This is sensible because we would expect a negative assessment to occur with higher CDS spreads. This pattern can also be observed for *news sentiment*.

4.2. Topic analysis of the analyst reports

One major challenge when using topic models is the goal-oriented identification of an appropriate topic number. For this purpose, several factors must be considered. The first aspect is the hierarchical structure of the data since each report belongs to a company and a fiscal quarter. Loughran and McDonald [8] point out that it is important to question whether the identified patterns could be proxies for underlying factors, such as time or company. We shed light on this issue in Section 4.4.

The number of evaluated topics started at 20 and ended at 100. From a theoretical perspective, more than one topic per company is

Table 3
Means per CDS spread quartiles.

	25%	50%	75%	100%
Return on assets	0.11	0.111	0.111	0.071
Revenue growth	0.018	0.026	0.021	0.038
Leverage	0.624	0.597	0.594	0.665
Retained earnings ratio	0.634	0.513	0.534	0.272
Net income growth	0	−0.001	0	−0.001
Stock return	0.174	0.164	0.146	0.151
Stock volatility	0.185	0.192	0.202	0.212
Index return	0.12	0.128	0.125	0.119
Distance to default	13.883	14.164	14.874	13.756
Risk-free rate	0.588	0.745	0.433	0.191
Credit rating	0.149	0.159	0.179	0.267
Analyst sentiment	0.081	0.238	−0.183	−0.139
News sentiment	0.198	0.117	−0.124	−0.218

reasonable since the model should capture multiple aspects per company. At the same time, choosing too many topics poses the risk that mostly small details are covered, while failing to capture the broad concepts. The appropriate number of topics was determined based on the work of Röder et al. [62]. The authors identified four dimensions to construct coherence measures and systematically evaluated different configurations. Coherence measures are useful because they help to identify topic models that are interpretable by humans and correspond to human judgment [62]. The four dimensions of interest are 1) segmentation, 2) probability calculation, 3) confirmation measure, and 4) aggregation. Established coherence measures such as normalized pointwise mutual information (NPMI) can be represented in this framework [62]. The authors found that a combination of word-to-word set comparison (for 1), a large Boolean sliding window (for 2), indirect cosine similarity (for 3), and the arithmetic mean (for 4) outperform alternative measures [62]. The superior result is indicated by the highest correlation to human assessment, and we refer to this coherence measure as C_v. Fig. 4 shows that the model with 55 topics exhibits the same C_v value of 0.619 as the model with 65 topics, while requiring fewer topics. Therefore, a less complex model with fewer parameters, which consists of 55 topics, was chosen.

We characterize the topics with word relevance rather than the raw topic-term probability from the LDA model. The relevance metric by Sievert and Shirley [63] was controlled by the λ parameter. It helps to create a balance between the topic-word probability and lift, which incorporates the marginal probability of the terms in the corpus. This helps to identify topic-specific terms more precisely. λ was set to 0.6, which yielded the best empirical results in the original paper. A full list of the extracted topics, topic labels, and ten words with the highest relevance per topic can be found in Table A.2.

Three key metrics are presented in Fig. 5 to improve our understanding of the characteristics of the topic. In each plot, the x-axis shows the topics sorted by the measure on the y-axis. The plot on the left contains the *maximum topic prevalence for a single company* (y-axis), which describes the largest aggregated topic probability value per topic (on a company basis). This plot illustrates the extent to which a topic focuses on a particular company. The figure in the middle shows the same relationship but in relation to years. This shows that a small percentage of topics are highly concentrated in an individual year. The overall topic prevalence in the right plot shows that some topics dominate across all years and companies. As an initial filter for our analysis, the following heuristics were adopted: Topics that are more than 50% attributable to a single company and thus have little tendency to generalize to more abstract concepts are excluded. Also, topics above the 90th percentile for max. *topic prevalence per year* and *relative topic prevalence overall* were excluded. Thereby, we can avoid topics that focus on a single year or are continuously present and dominant across all
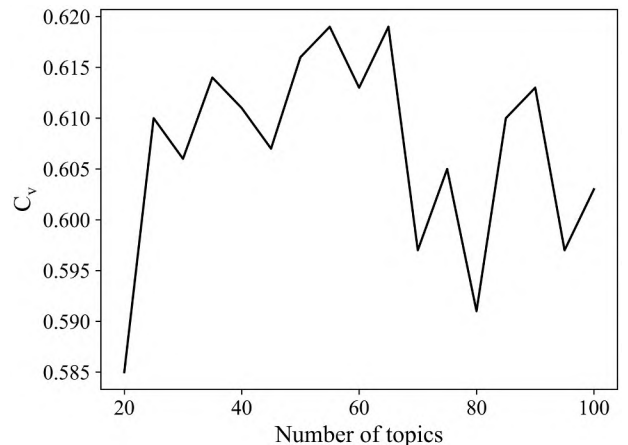
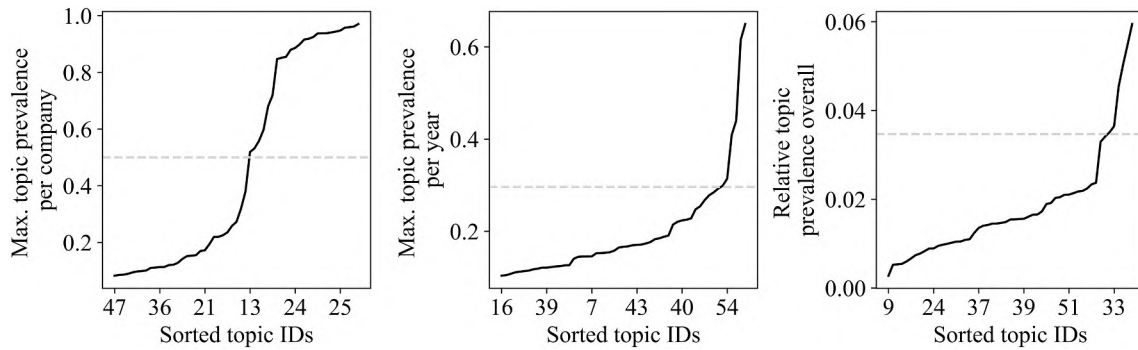**Fig. 4.** Topic coherence for all evaluated numbers of topics.

Fig. 5. The first two plots show the maximum topic prevalence for companies and years. The third plot shows the topic prevalence across companies and years. The cut-off is indicated by the dashed line.

companies and years.

4.3. Panel regression analysis

Table 4 presents the results of the panel regression analysis. The inclusion of established variables (see Section 3.5) helps to ensure that the extracted topics are not mere proxies for these variables. It is important to keep in mind that a CDS spread increase signals an increase in credit risk, as perceived by the market. A positive coefficient means that an increase in this variable is associated with an increase in credit risk. Both firm and time fixed effects are included to account for their specific characteristics. To ensure sufficient data coverage, a minimum of two analyst reports in combination with 500 words is required for a firm-quarter observation to be included.

Arguably, the most intuitive control variable is the *credit rating* because a strong association with default risk should be expected [27,46,64]. The positive coefficient estimate shows that a worse credit rating (i.e., a larger number) has a positive relationship with the *CDS spread*. *Return on assets* shows a significant inverse relationship with the *CDS spread*, which is in line with Donovan et al. [65] and Tsai et al. [27], thereby providing a good sanity-check. The results suggest that *revenue growth* shows no clear connection to *CDS spread* in our case. In the case of *leverage*, a clear correlation is expected because it expresses the relationship between debt and assets. The results support this assumption. In contrast, we do not find a statistically significant relationship for *retained earnings* and *net income growth* in the present study setup.

At the market variable level, *stock returns and index returns* show an

inverse relationship, and *stock volatility* is *positively* associated with CDS spreads, which is consistent with prior findings in the literature [27,46]. In the case of *distance to default*, this analysis does not indicate coefficients that are significantly different from zero, which would have been expected. At the macroeconomic level, the coefficient of the *risk-free rate* is *negative* but not statistically significant, which prior research would suggest [46,66]. In terms of textual data, the *sentiment* extracted from analyst reports shows a statistically significant link to *CDS spread*. The negative parameter estimate should be interpreted such that a more positive sentiment by financial analysts is associated with a lower CDS spread, signifying a lower credit risk. Since this parameter estimate is significantly different from zero, this result indicates that the content of analyst reports does indeed possess informational value when attempting to reason about CDS spreads. This also appears to be the case, while accounting for other control variables and fixed effects. With respect to the regression diagnostics, no noticeable heteroscedasticity was found. The residuals approximately represented a normal distribution, with some smaller outliers in the negative range. Furthermore, a good fit of the data is evident, as the R^2 of 75.9% indicates.

However, the question arises as to whether analyst reports offer incremental value even when established unstructured data sources such as financial news are included, for which previous studies have identified a relationship. For this purpose, financial *news sentiment* was also included in model [3R] to examine the robustness. The estimated coefficient for analyst sentiment remains negative and significant. Although one should be wary of interpreting the coefficients of the standardized sentiment variables as a precise measure of importance,

Table 4

Panel regression result of the logarithmic CDS spread regressed on [1] control variables, [2] control variables and analyst sentiment, [3] control variables, analyst sentiment, and topics, [3R] control variables, analyst and news sentiment, and topics. The regression shown in the table includes time and firm fixed effects. Clustered standard errors help to account for time and firm effects in the residuals; the *t*-statistics are shown in parentheses.

Var. name/model	[1]	[2]	[3]	[3R]
Return on assets	−2.206*** (−3.237)	−2.349*** (−3.588)	−2.561*** (−4.223)	−2.605*** (−4.299)
Revenue growth	−0.001 (−0.01)	0.09 (0.719)	0.069 (0.528)	0.081 (0.611)
Leverage	0.57* (1.888)	0.639** (2.198)	0.588** (2.007)	0.586** (2.021)
Retained earnings ratio	0.244 (0.922)	0.238 (0.983)	0.238 (1.012)	0.254 (1.074)
Net income growth	0.024 (0.028)	0.174 (0.212)	0.278 (0.256)	0.369 (0.352)
Stock return	−0.189*** (−3.27)	−0.128** (−2.141)	−0.13** (−2.212)	−0.105* (−1.886)
Stock volatility	0.709*** (2.697)	0.666*** (2.62)	0.654** (2.536)	0.693*** (2.695)
Index return	−0.181*** (−3.049)	−0.143 (−1.63)	−0.177** (−2.064)	−0.147 (−1.391)
Distance to default	0.001 (0.24)	0.001 (0.284)	0.002 (0.449)	0.001 (0.255)
Risk-free rate	−0.016 (−0.174)	−0.013 (−0.146)	−0.009 (−0.111)	−0.009 (−0.112)
Credit rating	1.703** (2.026)	1.788** (2.178)	1.675** (2.157)	1.601** (2.069)
Analyst sentiment		−0.05*** (−3.726)	−0.05*** (−3.477)	−0.043*** (−3.558)
News sentiment				−0.03*** (−2.661)
Firm fixed effects	included	included	included	included
Time fixed effects	included	included	included	included
Observations	646	646	646	646
R^2	0.736	0.744	0.759	0.761
F Statistics	12.262***	13.006***	6.391***	6.415***

Note: *$p < 0.1$; **$p < 0.05$; ***$p < 0.01$.

Table 5
Panel regression result for the included topics. Model [3] is shown, i.e., the logarithmic CDS spread is regressed on control variables, analyst sentiment, and topics. This table shows the coefficient estimates and *t*-statistics for each relevant topic. *Note:* The statistical significance of topics 21 and 31 did not persist when financial news sentiment was included in model [3R].

Name	Coefficient	Name	Coefficient	Name	Coefficient
1 Application platform	−0.402** (−2.046)	21 Buy-hold-sell recomm.	0.329* (1.672)	43 Change prediction	0.518 (1.067)
2 Broker disclaimer	−0.16 (−0.626)	28 Patent lawsuit	0.611** (2.211)	45 Macroeco. exposure	0.648 (1.404)
7 Fin. statement analysis	0.158 (0.398)	30 Survey data	0.067 (0.306)	47 Strategic direction	1.186*** (3.78)
9 Earnings call	−1.762 (−1.423)	31 Executive managers	−0.594* (−1.778)	51 Meeting analysis	−0.266 (−0.954)
10 Debt to cash flow	0.515 (0.921)	34 M&A deal	0.298*** (4.433)	53 Sovereign energy demand	0.297 (0.908)
15 Scenario-based prediction	−1.006** (−2.392)	41 Option metrics	0.324 (1.268)		
18 Miscellaneous	−0.655 (−1.059)	42 Revenue estimate	0.148 (0.387)		

the result shows roughly that the change in standardized sentiment in analyst reports is reflected in the dependent variable in a comparable magnitude as is the case for financial news.

4.4. Properties of the identified topics

Table 5 shows the panel regression results for model [3], that is, the full model, which includes analyst sentiment and topics. The pre-selection of the topics was performed as described in Section 4.2. For the following topics, the panel regression analysis indicates a statistical significance: (1) application platform, (15) scenario-based prediction, (21) buy-hold-sell recommendation, (28) patent lawsuit, (31) executive managers, (34) mergers and acquisitions deal, and (47) strategic direction.

The regression model serves as a basis for a more in-depth analysis of the statistically significant topics. This is essential because this study aims to explain changes in CDS spreads. Therefore, in addition to the regression analysis, the plausibility of the topics is of prime importance. Topic 1 (Fig. 6) is called *application platform* and describes digital services and products by firms such as *Microsoft*, *Cisco*, and *IBM*, which support their customers in their digital infrastructure needs. The most relevant words such as "solution" and "public cloud" provide evidence for this. The negative sign indicates an inverse association with credit risk, even when considering the fixed effects and control variables. This relationship is plausible considering the dramatic increase in the relevance of digitization in virtually all industries. The topic 15 *scenario-based prediction* deals with the projection of future financial outcomes. It is distributed relatively evenly across firms and over time. The most relevant words indicate that not only positive but also negative developments are considered. The negative sign indicates that the presence of such an assessment is associated with a reduction in uncertainty.

Topic 21 (Fig. 7) captures the buy-hold-sell recommendations by analysts. It is prevalent during the earlier years included in the analysis. The top words indicate that the topic is concerned with the specific buy, hold, or sell recommendations, which provide the basis for actions of the broker's customers. The positive coefficient of the topic indicates a positive correlation with CDS spreads. Content dealing with investigations and *patent lawsuits* is identified when topic 28 is present. The words "litigation" and "settlement" show that this includes both the initial claim and the subsequent settlement. The topic is most present in

reports about *Johnson & Johnson*. Its distribution indicates relevance to multiple companies. The positive sign shows that the presence of reports dealing with patent lawsuits prior to earnings announcements correlates with a higher degree of credit risk. It also shows that the market does not perceive these lawsuits as a minor issue but as a significant threat.

Content dealing with the *executive managers* of a company is covered by topic 31 (Fig. 8). It includes not only the CEO but also the board of directors. For analyst reports, which generally cluster around earnings announcements, the executives of a company play a special role. Since earnings calls may include discussions with the CEO and CFO, this key personnel is analyzed carefully. The inverse relationship with CDS spreads suggests that topic 31 might come into play specifically when executive managers are perceived as confident or optimistic in the period leading up to the earnings announcement. The top words of topic 34 indicate that *mergers and acquisitions* (M&A) are the main concern. Since we analyze constituents of the DJIA, in most cases, these large companies acquire other businesses. While this can be a positive development from a risk perspective, it can also entail risks if the desired advantages (e.g., economies of scale) are not attained. The estimated coefficient suggests that discussions regarding M&A in the run-up to earnings announcements are associated with more uncertainty or risk.

The last topic (47) deals with strategic aspects, specifically the *strategic direction* (Fig. 9). The most relevant words show that not only the possibilities are being analyzed ("opportunity"), but also what can be achieved ("capability"). The high prevalence of this topic for *General Electric* and *Walmart* signals that topic 47 deals with established companies that must reinvent themselves as part of the ubiquitous digital transformation in the past decade. The synchronous relationship with CDS spreads shows that uncertainties regarding the future strategic direction are very prominent in topic 47.

5. Discussion

In the following, we discuss the implications for practitioners, researchers, and highlight the study's limitations. The practical relevance of this study is illustrated by the underlying use case. The first insight is that a statistical relationship between analysts' sentiment prior to the earnings date and CDS spreads is identified. At a fundamental level, this indicates that the textual output of analysts could improve our understanding of the credit risk of companies. It could also be a starting point
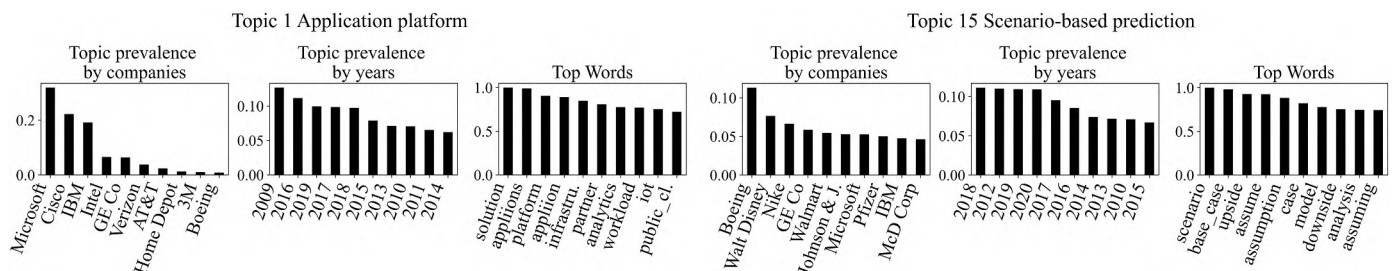

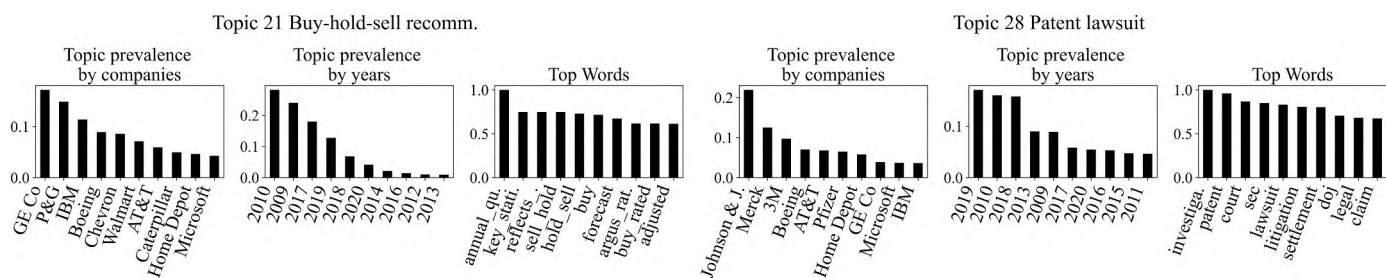**Fig. 6.** Property visualization of Topic 1 and 15.

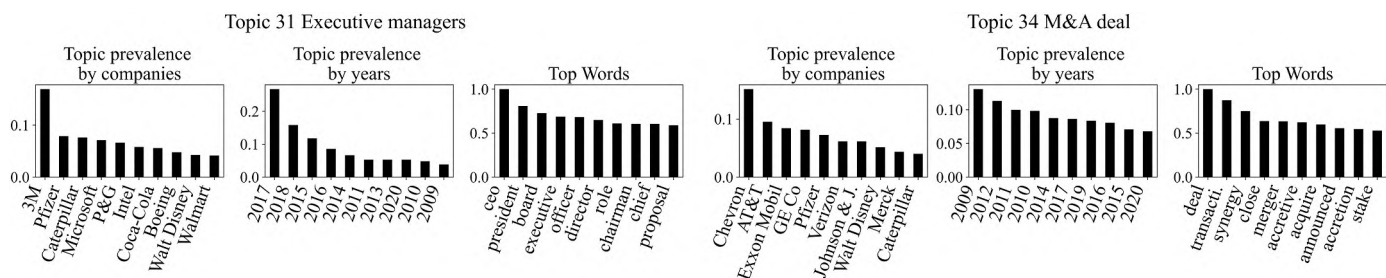Fig. 7. Property visualization of Topic 21 and 28.

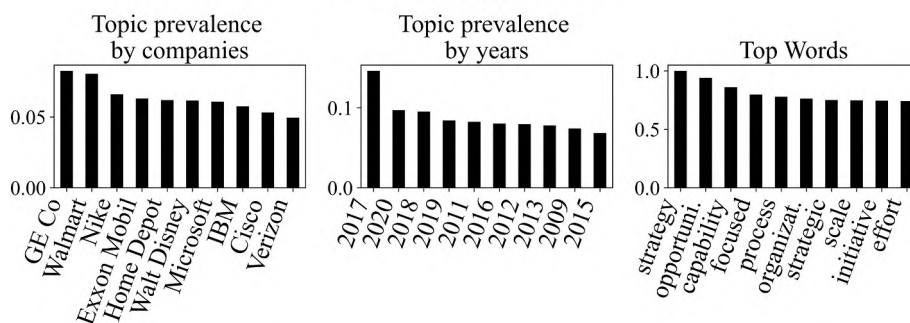Fig. 8. Property visualization of Topic 31 and 34.

Fig. 9. Property visualization of Topic 47.

to try and construct a risk proxy measure for companies without CDS [65]. The link between sentiment and CDS spread is useful for the complementary topic analysis since a missing correlation could call into question the relevance of analyst reports. The topics identified with the regression analysis can help to understand CDS spreads better. After accounting for fixed effects and established control variables, a statistically significant relationship was found for several latent topics that were identified with LDA.

Interestingly, these topics include subjects that can apply to multiple companies (e.g., *strategic direction*) and others that capture a specific group of companies (e.g., *application platform*). The proposed approach can help practitioners improve their DDD by explaining changes in the credit risk using qualitative data sources. In addition, this analysis highlights the potential of using topic models or related text representations to quantify analyst reports and integrate them into a quantitative risk management system. While earnings conference calls could also provide insights regarding credit risk, analyst reports can provide incremental information because analysts give independent weight to the topics that they consider to be important. If years of analyst experience and industry expertise are considered, this is a qualitative data source that should be integrated alongside quantitative metrics.

This study contributes to the following aspects of research. To our knowledge, this is the first study to link credit risk measured by CDS spreads with the sentiment from analyst reports. The results show an inverse relationship between textual sentiment and risk, which corresponds to results from financial news [11,27,29,30]. Given that the effect persists even when financial news sentiment is taken into account, the findings indicate an incremental value of analyst research. We are also the first to examine the statistical relationship between CDS spreads and topics identified in analyst reports to understand potential reasons for credit risk changes. This study's contribution lies in the empirical results, and it provides insights for the model component of a decision support system for credit risk. Our approach tackles the issue that arises with the rise of black-box machine learning models, which increasingly shift the focus from "why" to "what" [35]. This analysis can help to explore subjects associated with a changed risk situation. Finally, we contribute to research analyzing the relationship between credit risk and financial documents such as regulatory filings [27] or financial news [27,30].

Naturally, the conducted analysis has limitations and drawbacks that need to be addressed. Regarding the usage of text mining, it is important to consider that the predictions for sentiment or topics can be inaccurate. Aspects such as irony or complex sentence structures quickly stretch the limits of semantic richness that can be captured using techniques such as LDA. More powerful dense representations, for example, transformer models, can potentially represent text in a semantically

richer way. At the same time, the interpretability of the latent topics was of utmost importance for the use case at hand. Another limitation is that the analysis takes place at the quarterly level, as analyst reports tend to cluster around earnings announcements. Therefore, a granular analysis at the daily level would also be of interest. Furthermore, there is potential for mixed frequency analysis, as it is essential to consider accounting measures to avoid topics mistakenly being identified as relevant.

6. Conclusions

This study examined the value of analyst reports for DDD in the field of credit risk assessment. We link the latent topics in financial analyst reports to CDS spreads while accounting for established market-based, accounting, and macroeconomic measures. This scalable analysis approach was applied to 3386 analyst reports covering 22 companies. Regarding *RQ1*, the results suggest that the sentiment of analyst reports exhibits a statistical relationship with credit risk, as measured by CDS spreads, even if we account for news sentiment. The panel analysis shows that seven latent topics exhibit a significant statistical association, addressing *RQ2*. The topics exhibited distinct differences concerning how strongly they are focused on a company or industry. This paper's contribution lies in the empirical insights (association between CDS spread and sentiment/topics) and the fact that the analysis can be a

reference for studies addressing the intersection of credit risk and unstructured data. The presented approach can also be useful for risk managers to support their analysis of the risk situation.

Research gaps became apparent based on existing literature and throughout this study. First, there is a variety of design choices for the topic model. While in this case heuristics were used to remove unsuitable topics, the same logic could be extended using a specialized Bayesian topic model. Furthermore, mixed frequency and machine learning models should be utilized to consider macroeconomic and accounting variables while analyzing analyst reports in more detail. However, the tradeoff between forecast quality and interpretability must be considered. Therefore, future research should also investigate the prediction aspects in more depth. It would also be interesting to examine how key employees' decisions are impacted by incorporating a model prototype.

Funding

Declaration of Competing Interest

None.

Appendix A

Table A.1
Companies included in the analysis (only non-financial).

3M	AT&T	Boeing	Caterpillar	Chevron
Cisco	Coca-Cola	Exxon Mobil	General Electric	Home Depot
IBM	Intel	Johns. & Johns.	McDonald's	Merck
Microsoft	Nike	Procter & Gamble	Pfizer	Verizon
Walmart	Walt Disney			

Table A.2
Top 10 words per topic estimated via LDA for analyst reports. Some topics are not part of the regression analysis (see Section 4.2). Topic label assignment is based on the detailed topic analysis plots as seen in Section 4.4.

Topic	Top 10 words with highest relevance per topic
1 Application platform	solution, appliions, platform, appliion, infrastructure, partner, analytics, workload, iot, public_cloud
2 Broker disclaimer	research, subject, compensation, specific_recommendation, received_compensation, view_expressed, analyst, reflect_personal, report_accurately, affiliate
3 Nike	dtc, adidas, basketball, demand_creation, digital, athletic, jordan, footwear_apparel, western_europe, woman
4 Smartphone	iphone, net_add, arpu, smartphone, churn, upgrade, prepaid, subsidy, post_paid, verse
5 Uptrend	upside, solid, momentum, improving, cycle, incremental, gain, healthy, near_term, strength
6 Walmart ecommerce	commerce, merchane, flipkart, ecommerce, food, online, sam, assortment, format, fuel
7 Financial statement analysis	charge, income, gaap, adjusted, related, item, loss, accounting, excluding, approximately
8 Sector trade	rank, sector, pair_trading, fed, best_rank, similarity_index, closest_competitor, attractiveness, quality, ecb
9 Earnings call	forward_looking, today, slide, statement, analyst, investor_relation, good_morning, qtr, grew, financial
10 Debt to cash flow	debt, free_cash, flow, repurchase, balance_sheet, shareholder, buyback, liquidity, fund, ratio
11 Orthopedic surgery	synthes, surgery, worldwide, ous, operational, spine, knee, hip, surgical, otc
12 Mobile service provider	spectrum, lte, carrier, sprint, verizons, fiber, unlimited, fcc, tower, telecom
13 Natural gas	permian, natural_gas, gas, lng, exploration, refining, bbl, liquid, xto, crude_oil
14 Caterpillar machinery	machinery, dealer_inventory, energy_transportation, erpillars, engine, machine, construction_equipment, oil_gas, mining_equipment, aftermarket
15 Scenario-based prediction	scenario, base_case, upside, assume, assumption, case, model, downside, analysis, assuming
16 Issues	issue, dont, number, fact, make, doe, big, long, problem, clear
17 Internet provider	support, solution, offer, ethernet, managed, access, feature, carrier, vpn, unified_communiions
18 Miscellaneous	thing, lot, kind, people, talk, great, little_bit, sort, good, question
19 Soft drinks	bottler, bottling, sparkling, refranchising, drink, pep, cce, concentrate, csd, sparkling_beverage
20 Regulatory disclaimer	research_analyst, exchange_regulated, regulation_authority, authority, security_plc, prudential_regulation, taiwan_security, registration_number, exchange, future_commission
21 Buy-hold-sell recomm.	annual_quarterly, key_statistic, reflects_previous, sell_hold, hold_sell, buy, forecast, argus_rating, buy_rated, adjusted
22 Intel chips	dcg, memory, cpu, ccg, altera, nand, amd, foundry, tsmc, mobileye
23 P&G body care	fabric_care, grooming, diaper, care, commodity, hair_care, family_care, gillette, developed, olay
24 Microsoft cloud & personal comp.	lidin, gaming, aws, intelligent_cloud, microsofts, personal_computing, github, window_oem, mpc, pbp

(continued on next page)

Table A.2 (*continued*)

Topic	Top 10 words with highest relevance per topic
25 Boeing airplanes	max, airline, airplane, airbus, plane, commercial_airplane, flight, aerospace, fleet, supplier
26 Intel mobile products	tablet, chip, notebook, atom, shipment, sandy_bridge, amd, microprocessor, arm, asp
27 Efficacy pharmaceuticals	dose, placebo, efficacy, fda, tofacitinib, phase_iii, clinical, dos, xarelto, therapy
28 Patent lawsuit	investigation, patent, court, sec, lawsuit, litigation, settlement, doj, legal, claim
29 Vaccines	vaccine, pfizers, prevnar, animal_health, xeljanz, wyeth, lipitor, generic, januvia, innovative
30 Survey data	survey, march, respondent, week, ubs_evidence, proprietary, basket, surveyed, monthly, indie
31 Executive managers	ceo, president, board, executive, officer, director, role, chairman, chief, proposal
32 Network infrastructure	ucs, routing, nexus, public_sector, switch, juniper, router, collaboration, recurring, switching_routing
33 Competitor analysis	firm, division, economy, moat, economic_moat, giant, competition, account, rise, growing
34 M&A deal	deal, transaction, synergy, close, merger, accretive, acquire, announced, accretion, stake
35 Healthcare input cost	rose, health_care, electronics_energy, safety_graphic, local_currency, raw_material, fell, acelity, lcd, non_recurring
36 Macroeconomic weakness	pressure, weakness, weak, macro, negative, lowering, near_term, spending, cut, environment
37 Johnson & Johnson products	zytiga, remicade, imbruvica, darzalex, invokana, ims, daratumumab, xarelto, olysio, stelara
38 COVID-19	covid, fy20, fy19, fy21, 4q19, important_closure, pandemic, 1q20, 2q20, 1q19
39 IBM product suite	strategic_imperative, signing, mainframe, constant_currency, watson, gts, red_hat, cognitive, outsourcing, analytics
40 Satellite TV	video, directv, dtv, time_warner, entertainment, ott, twx, warnermedia, bundle, verse
41 Option metrics	option, implied_volatility, volatility, considered_overvalued, spread, straddle, etf, strike, considered_undervalued, reuters_com
42 Revenue estimate	rev, street, guide, est., beat, mgmt, guided, model, miss, flat
43 Change prediction	grew, declined, sequentially, flat, slightly, prior, decreased, compared, expects, noted
44 Disney media and parks	star_war, box_office, cable_network, programming, theme_park, movie, park_resort, abc, cable, fox
45 Macroeconomic exposure	fiscal, basis_point, fourth_quarter, ago, compared, rose, period, foreign_currency, declined, totaled
46 Valuation	valuation, premium, relative, trade, count, historical, trading, yield, sector, dividend_yield
47 Strategic direction	strategy, opportunity, capability, focused, process, organization, strategic, scale, initiative, effort
48 Microsoft gaming and search	xbox, microsofts, search, nokia, window_phone, bing, server_tool, skype, user, surface
49 Cancer treatment	keytruda, nsclc, lung_cancer, cancer, tumor, keynote, chemotherapy, pfs, combo, melanoma
50 Home Depot	pro, hds, lowes, big_ticket, appliance, weather, online, lumber, supply_chain, diy
51 Meeting analysis	meeting, belief, noted, opportunity, highlighted, expects, update, analyst, reiterated, strategy
52 McDonald's revenue sources	breakfast, apmea, qsr, food, franchisees, sandwich, eotf, franchisee, chicken, burger
53 Sovereign energy demand	country, india, government, energy, plant, japan, chinese, local, world, facility
54 Energy production	alstom, oil_gas, gas_turbine, energy, wind, renewables, nbcu, finance, bhge, restructuring
55 Forecast evaluate	forecast, beat, slightly, previously, ahead, raised, raising, versus, upside, unchanged

References

[1] T. Van Gestel, B. Baesens, Credit Risk Management: Basic Concepts: Financial Risk Components, Rating Analysis, Models, Economic and Regulatory Capital, Oxford University Press, Oxford, United Kingdom, 2008.

[2] J. Hull, M. Predescu, A. White, The relationship between credit default swap spreads, bond yields, and credit rating announcements, J. Bank. Financ. 28 (11) (2004) 2789–2811, https://doi.org/10.1016/j.jbankfin.2004.06.010.

[3] F.A. Longstaff, S. Mithal, E. Neis, Corporate yield spreads: default risk or liquidity? New evidence from the credit default swap market, J. Financ. 60 (5) (2005) 2213–2253, https://doi.org/10.1111/j.1540-6261.2005.00797.x.

[4] J.O. Soares, J. Pina, M. Ribeiro, M.C. Lopes, Quantitative vs. qualitative criteria for credit risk assessment, Front. Finance Econ. 8 (1) (2011) 69–87.

[5] R. Agarwal, V. Dhar, Big data, data science, and analytics: the opportunity and challenge for IS research, Inf. Syst. Res. 25 (3) (2014) 443–448.

[6] E. Brynjolfsson, L.M. Hitt, H.H. Kim, Strength in Numbers: How Does Data-Driven Decisionmaking Affect Firm Performance?, Available at SSRN 1819486, 2011.

[7] T.H. Davenport, J. Kirby, Beyond automation, Harv. Bus. Rev. 93 (6) (2015) 58–65.

[8] T. Loughran, B. McDonald, Textual analysis in accounting and finance: a survey, J. Account. Res. 54 (4) (2016) 1187–1230.

[9] D. Bradley, J. Clarke, S. Lee, C. Ornthanalai, Are Analysts' recommendations informative? Intraday evidence on the impact of time stamp delays, J. Financ. 69 (2) (2014) 645–673.

[10] E.X. Li, K. Ramesh, M. Shen, J.S. Wu, Do analyst stock recommendations piggyback on recent corporate news? An analysis of regular-hour and after-hours revisions, J. Account. Res. 53 (4) (2015) 821–861.

[11] M. Liebmann, A.G. Orlov, D. Neumann, The tone of financial news and the perceptions of stock and CDS traders, Int. Rev. Financ. Anal. 46 (2016) 159–175, https://doi.org/10.1016/j.irfa.2016.05.001.

[12] H.-M. Lu, F.-T. Tsai, H. Chen, M.-W. Hung, S.-H. Li, Credit rating change modeling using news and financial ratios, ACM Trans. Manag. Inf. Syst. 3 (3) (2012) 1–30.

[13] K. Galil, G. Soffer, Good news, bad news and rating announcements: an empirical investigation, J. Bank. Financ. 35 (11) (2011) 3101–3119, https://doi.org/10.1016/j.jbankfin.2011.04.010.

[14] A.J. McNeil, R. Frey, P. Embrechts, Quantitative Risk Management: Concepts, Techniques and Tools - Revised Edition, 2nd ed., Princeton University Press, Princeton, New Jersey, 2015.

[15] N. Chen, B. Ribeiro, A. Chen, Financial credit risk assessment: a recent review, Artif. Intell. Rev. 45 (1) (2016) 1–23, https://doi.org/10.1007/s10462-015-9434-x.

[16] J. Bai, S.-J. Wei, When is there a strong transfer risk from the sovereigns to the corporates? Property rights gaps and CDS spreads, in: NBER Working Paper Series, (National Bureau of Economic Research), 2012.

[17] A. Berndt, R.A. Jarrow, C. Kang, Restructuring risk in credit default swaps: an empirical analysis, Stoch. Process. Appl. 117 (11) (2007) 1724–1749.

[18] K.L. Womack, Do brokerage Analysts' recommendations have investment value? J. Financ. 51 (1) (1996) 137–167.

[19] X. Chen, Q. Cheng, K. Lo, On the relationship between analyst reports and corporate disclosures: exploring the roles of information discovery and interpretation, J. Account. Econ. 49 (3) (2010) 206–226.

[20] E. Soltes, Private interaction between firm management and sell-side analysts, J. Account. Res. 52 (1) (2014) 245–272.

[21] A.H. Huang, R. Lehavy, A.Y. Zang, R. Zheng, Analyst information discovery and interpretation roles: a topic modeling approach, Manag. Sci. 64 (6) (2017) 1–23.

[22] P.T. Elgers, M.H. Lo, R.J. Pfeiffer Jr., Delayed security Price adjustments to financial Analysts' forecasts of annual earnings, Account. Rev. 76 (4) (2001) 613–632.

[23] R. Barniv, O.-K. Hope, M.J. Myring, W.B. Thomas, Do analysts practice what they preach and should investors listen? Effects of recent regulations, Account. Rev. 84 (4) (2009) 1015–1039.

[24] A.H. Huang, A.Y. Zang, R. Zheng, Evidence on the information content of text in analyst reports, Account. Rev. 89 (6) (2014) 2151–2180.

[25] J. Roeder, Alternative data for credit risk management: an analysis of the current state of research, in: Proceedings of the 34th Bled eConference, Bled, Slovenia, 2021, pp. 167–179.

[26] Y. Bao, A. Datta, Simultaneously discovering and quantifying risk types from textual risk disclosures, Manag. Sci. 60 (6) (2014) 1371–1391.

[27] F.-T. Tsai, H.-M. Lu, M.-W. Hung, The impact of news articles and corporate disclosure on credit risk valuation, J. Bank. Financ. 68 (2016) 100–116, https://doi.org/10.1016/j.jbankfin.2016.03.018.

[28] L. Wei, G. Li, X. Zhu, J. Li, Discovering bank risk factors from financial statements based on a new semi-supervised text mining algorithm, Account. Finance 59 (3) (2019) 1519–1552, https://doi.org/10.1111/acfi.12453.

[29] L.A. Smales, News sentiment and Bank credit risk, J. Empir. Financ. 38 (2016) 37–61, https://doi.org/10.1016/j.jempfin.2016.05.002.

[30] S. Yang, Z. Liu, X. Wang, News sentiment, credit spreads, and information asymmetry, N. Am. J. Econ. Financ. 52 (2020), https://doi.org/10.1016/j.najef.2020.101179.

[31] C.A. Greatrex, The credit default swap Market's reaction to earnings announcements, J. Appl. Financ. 19 (1/2) (2009) 193–216.

[32] D.J. Power, Decision support systems: a historical overview, Handbook Decis. Support. Syst. 1 (2008) 121–140.

[33] D.J. Power, Using 'big data' for analytics and decision support, J. Decis. Syst. 23 (2) (2014) 222–228, https://doi.org/10.1080/12460125.2014.888848.

[34] A. Bhimani, Exploring big data's strategic consequences, J. Inf. Technol. 30 (1) (2015) 66–69.

[35] A. Abbasi, S. Sarker, R.H.L. Chiang, Big data research in information systems: toward an inclusive research agenda, J. Assoc. Inf. Syst. 17 (223) (2016) i–xxxii, https://doi.org/10.17705/1jais.00423.

[36] A. McAfee, E. Brynjolfsson, Big data: the management revolution, Harv. Bus. Rev. 90 (10) (2012) 60–68.

[37] E. Brynjolfsson, K. McElheran, Data in Action: Data-Driven Decision Making in US Manufacturing, Available at SSRN 2722502, 2016.

[38] V. Grover, R.H. Chiang, T.-P. Liang, D. Zhang, Creating strategic business value from big data analytics: a research framework, J. Manag. Inf. Syst. 35 (2) (2018) 388–423.

[39] T.H. Davenport, Competing on analytics, Harv. Bus. Rev. 84 (1) (2006) 98–107.

[40] U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, From data mining to knowledge discovery in databases, AI Mag. 17 (3) (1996) 37–54.

[41] R. Feldman, I. Dagan, Knowledge discovery in textual databases (KDT), KDD 95 (1995) 112–117.

[42] National Bureau of Economic Research, US Business Cycle Expansions and Contractions. https://www.nber.org/cycles.html, 2020 (accessed 09/22/2020).

[43] K. Galil, O.M. Shapir, D. Amiram, U. Ben-Zion, The determinants of CDS spreads, J. Bank. Financ. 41 (2014) 271–282, https://doi.org/10.1016/j.jbankfin.2013.12.005.

[44] S. Mayordomo, J.I. Pena, E.S. Schwartz, Are all credit default swap databases equal? Eur. Financ. Manag. 20 (4) (2014) 677–713.

[45] M. Greenwood-Nimmo, J. Huang, V.H. Nguyen, Financial sector bailouts, sovereign bailouts and the transfer of credit risk, J. Financ. Mark. 42 (2016) 121–142, https://doi.org/10.1016/j.finmar.2018.11.001.

[46] S.R. Das, P. Hanouna, A. Sarin, Accounting-based versus market-based cross-sectional models of CDS spreads, J. Bank. Financ. 33 (4) (2009) 719–730, https://doi.org/10.1016/j.jbankfin.2008.11.003.

[47] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, in: Advances in Neural Information Processing Systems 26, Lake Tahoe, USA, 2013, pp. 3111–3119.

[48] R. Řehůřek, P. Sojka, Software framework for topic modelling with large corpora, in: Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks, ELRA, Valetta, Malta, 2010.

[49] E. Cambria, S. Poria, A. Gelbukh, M. Thelwall, Sentiment analysis is a big suitcase, IEEE Intell. Syst. 32 (6) (2017) 74–80.

[50] T. Loughran, B. McDonald, When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks, J. Financ. 66 (1) (2011) 35–65.

[51] E. Henry, Are investors influenced by how earnings press releases are written? J. Bus. Commun. 45 (4) (2008) 363–407.

[52] Y. Yang, M.C.S. Uy, A. Huang, FinBERT: a pretrained language model for financial communications, arXiv (2020) preprint arXiv:2006.08097.

[53] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, in: Proceedings of the Proceedings of the 2019 Conference of the NAACL, Minneapolis, Minnesota, 2019, pp. 4171–4186.

[54] M. Palmer, J. Roeder, J. Muntermann, Induction of a sentiment dictionary for financial analyst communication: a data-driven approach balancing machine learning and human intuition, J. Business Analyt. (2021) 1–21, https://doi.org/10.1080/2573234X.2021.1955022.

[55] D.M. Blei, A.Y. Ng, M.I. Jordan, Latent Dirichlet allocation, J. Mach. Learn. Res. 3 (2003) 993–1022.

[56] D.M. Blei, Probabilistic topic models, Commun. ACM 55 (4) (2012) 77–84, https://doi.org/10.1145/2133806.2133826.

[57] L. Yao, D. Mimno, A. McCallum, Efficient methods for topic model inference on streaming document collections, in: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2009, pp. 937–946.

[58] D. Aunon-Nerin, D. Cossin, T. Hricko, Z. Huang, Exploring for the determinants of credit risk in credit default swap transaction data: is fixed-income markets' information sufficient to evaluate credit risk? FAME Res. Paper 65 (2002) https://doi.org/10.2139/ssrn.375563.

[59] R.C. Merton, On the pricing of corporate debt: the risk structure of interest rates, J. Financ. 29 (2) (1974) 449–470, https://doi.org/10.2307/2978814.

[60] S.T. Bharath, T. Shumway, Forecasting default with the Merton distance to default model, Rev. Financ. Stud. 21 (3) (2008) 1339–1369.

[61] J. Kleinow, Systemrelevante Finanzinstitute - Systemrisiko und Regulierung im europäischen Kontext, Springer Gabler, Wiesbaden, 2016.

[62] M. Röder, A. Both, A. Hinneburg, Exploring the space of topic coherence measures, in: Proceedings of the Eighth ACM International Conference on Web Search and Data Mining, ACM, Shanghai, China, 2015, pp. 399–408.

[63] C. Sievert, K. Shirley, LDAvis: a method for visualizing and interpreting topics, in: Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces, 2014, pp. 63–70.

[64] F. Kiesel, J. Spohnholtz, CDS spreads as an independent measure of credit risk, J. Risk Financ. 18 (2) (2017) 122–144, https://doi.org/10.1108/JRF-09-2016-0119.

[65] J. Donovan, J. Jennings, K. Koharki, J. Lee, Determining Credit Risk Using Qualitative Disclosure, Available at SSRN 3149945, 2018.

[66] J. Lee, S. Kim, Y.J. Park, Investor sentiment and credit default swap spreads during the global financial crisis, J. Futur. Mark. 37 (7) (2017) 660–688, https://doi.org/10.1002/fut.21828.

Jan Roeder is a doctoral student and research associate at the University of Goettingen. He obtained his master's degree in business information systems at the Georg-August-University of Goettingen in 2017 and his bachelor's degree in 2014. In his research, he is particularly concerned with how heterogeneous data sources, especially text data, can be made usable and how they can be analyzed. This is being done specifically in the context of credit risk management.
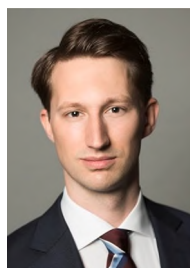
Matthias Palmer is a doctoral student and research associate at the University of Goettingen. Prior to that, he obtained a bachelor's degree in business administration from the University of Hamburg in 2013 and a master's degree in finance, accounting and taxes from the University of Goettingen in 2016. His research interests focus primarily on leveraging text data for finance-specific applications and, in particular, on the work of financial analysts.
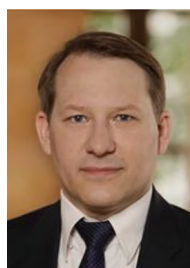
Jan Muntermann is a Professor of Financial Data Analytics at the University of Augsburg. His research interests include (big) data analytics and managerial decision support, digital business strategy development and execution in the financial services sector, and the conceptual and methodological foundations of theory development in information systems research. He has published in outlets such as *Information Systems Research, Journal of the Association for Information Systems, Journal of Information Technology, Decision Support Systems* and the *European Journal of Information Systems*.