Fritzsch, Simon; Scharner, Philipp; Weiß, Gregor

Estimating the relation between digitalization and the market value of insurers

Journal of Risk and Insurance

Provided in Cooperation with:
John Wiley & Sons

Suggested Citation: Fritzsch, Simon; Scharner, Philipp; Weiß, Gregor (2021) : Estimating the relation between digitalization and the market value of insurers, Journal of Risk and Insurance, ISSN 1539-6975, Wiley, Hoboken, NJ, Vol. 88, Iss. 3, pp. 529-567, https://doi.org/10.1111/jori.12346

Received: 15 September 2020 | Revised: 18 April 2021 | Accepted: 22 April 2021

DOI: 10.1111/jori.12346

ORIGINAL ARTICLE

Journal of Risk and Insurance

Estimating the relation between digitalization and the market value of insurers

Simon Fritzsch | **Philipp Scharner** | **Gregor Weiß**

Faculty of Economics and Management, Leipzig University, Leipzig, Germany

Correspondence
Gregor Weiß, Faculty of Economics and Management, Leipzig University, Grimmaische Str. 12, Leipzig 04109, Germany.
Email: weiss@wifa.uni-leipzig.de

Funding information
LBBW Asset Management Investmentgesellschaft mbH, Grant/Award Number: PhD scholarship

Abstract

We analyze the relation between digitalization and the market value of US insurance companies. To create a text-based measure that captures the extent to which insurers digitalize, we apply an unsupervised machine learning algorithm—Latent Dirichlet Allocation—to their annual reports. We show that an increase in digitalization is associated with an increase in market valuations in the insurance sector. In detail, capital market participants seem to reward digitalization efforts of an insurer in the form of higher absolute market capitalizations and market-to-book ratios. Additionally, we provide evidence that the positive relation between digitalization and market valuations is robust to sentiment in the annual reports and the choice of the reference document on digitalization, both being issues of particular importance in text-based analyses.

KEYWORDS
digitalization, Latent Dirichlet Allocation, machine learning, market valuation

JEL CLASSIFICATION
C33, C61, G22

1 | INTRODUCTION

Digitalization has already massively transformed many industries. The insurance industry, however, has yet to take advantage of the full potential of digital technologies. This becomes even more important as rising customer expectations, the effects of the financial markets crisis, and the zero interest rate policy lead to an increased competitive pressure. In general, there is no doubt on the strong impact digitalization will have on the insurance ecosystem (see, e.g., Cappiello, 2020). It is considered to affect the whole insurance value chain, from product development to pricing/underwriting, sales and distribution, policy and claims management, and asset and risk management (Eling & Lehmann, 2018). However, in contrast to other megatrends such as urbanization or aging societies, the precise scope of digitalization is difficult to grasp. Although we know that digitalization clearly manifests itself in cloud computing, Internet of Things, mobile communication, blockchain technology, artificial intelligence, and so forth (Schmidt, 2018), evidence on the question of how to measure digitalization and its relation to firm outcomes is still scarce (see, e.g., Bohnert et al., 2019; Hanelt et al., 2020; Scott et al., 2017).

In this paper, we fill this gap by proposing a new method to measure digitalization in the insurance sector. Our method exploits the prevalence of different topics in standard annual reports. Based on the assumption that digitally innovative insurers report their progress more extensively, Latent Dirichlet Allocation (LDA) helps to assess the importance of digitalization for the particular insurance company. At the same time, the method enables us to separate digitalization from mere firm innovation. In a second step, we use our text-based measure on digitalization to investigate its relation with the market valuation of a large set of publicly-listed US insurance companies. Finally, we account for potential confounding issues related to the construction of the digitalization measure, the reference document used for LDA, and the sentiment in which annual reports are written.

Our results provide first evidence for a positive association between digitalization efforts and market valuation in the US insurance sector. We find that an increase in digitalization is strongly related to an increase in market value and market-to-book value of US insurance companies. Put differently, market participants associate a more digitalized insurance company with higher future profitability and consequently a higher firm value. Although LDA is by design subject to some discretion, we show that our results are robust to several variations of our model parameters. Most importantly, our findings are robust to different numbers of topics used to structure the annual reports and to isolate digitalization from general innovation. Furthermore, the results do not depend on the discretionary choice of the reference document and are not confounded by annual reports' sentiment.

The topic model LDA by Blei et al. (2003) has only recently been introduced to the finance literature.[1] In general, topic models can be used to analyze large data sets of texts that are often unstructured (Roberts et al., 2016). These probabilistic models provide a finite set of common topics which optimally reflect a collection of documents. By applying a topic model to a specific document, we obtain a vector of topic loadings representing how intensively each topic is discussed in the respective document. One of the main advantages of LDA over simple word-list approaches is that the topics and corresponding word distributions arise endogenously from the data and do not have to be specified by the researcher. That is, the underlying machine

[1] One of the first applications of this approach in accounting and finance is due to Huang et al. (2018), who study topical differences between conference calls and subsequent analyst reports.

learning algorithm determines the terms that are most important to discriminate between documents and topics in an unsupervised fashion.

We then apply this powerful tool to the annual reports of 86 publicly listed US insurance companies available in Thomson Reuters Datastream from 2006 to 2015 and derive a distribution of topics for each of the annual reports. This yields a low-dimensional representation of the document (cf. Blei et al., 2003) that we exploit to construct our text-based measure of digitalization. For this purpose, we compare the extent to which each topic is discussed in the respective report to a reference document about digitalization in the insurance sector. Specifically, we use the paper by Bohnert et al. (2019) since it is closely related to our work. To the best of our knowledge, it is one of the few studies trying to establish an empirical relation between the expression of a digital agenda and the market valuation of insurance companies.[2] More exactly, based on these topic distributions we calculate a measure of similarity between the digitalization document and the insurers' annual reports using the Kullback–Leibler (KL) divergence (Kullback & Leibler, 1951). This measure is then used to proxy for the extent of digitalization in our sample insurance companies.

Our paper is related to a growing body of literature on textual analysis and machine learning in finance. Starting with Frazier et al. (1984), Antweiler and Frank (2004), and Tetlock (2007), researchers have studied the effect of qualitative information on equity valuations. More recent papers (e.g., Hanley & Hoberg, 2010; Hoberg & Maksimovic, 2015; Hoberg et al., 2014; Jegadeesh & Wu, 2013; Ke et al., 2017, 2019) conduct text-based analyses to examine a wide variety of finance research questions.[3] Intriguingly, within the field of text analysis and machine learning, LDA is becoming more and more popular (see, e.g., Ganglmair & Wardlaw, 2017; Goldsmith-Pinkham et al., 2016; Hoberg & Lewis, 2017; Huang et al., 2018; Lopez-Lira, 2019). Within this growing strand of literature, our paper is most closely related to Bellstam et al. (2020) and Lowry et al. (2019), who derive topics based on LDA and employ the KL divergence as a measure of similarity between probability distributions.

At the same time, our paper is related to a growing body of literature on the effect of digitalization in the insurance sector. This literature is basically centered around the impact of digitalization on the business model of insurers (see, e.g., Cappiello, 2020; Desyllas & Sako, 2013), new forms of online marketing and sales activities (Seitz, 2017), and the overall transformation of insurance companies (Barkur et al., 2007). The various facets of digitalization such as Big Data, artificial intelligence, predictive modeling, telematics, and Internet of Things are considered to have a tremendous impact on the whole insurance value chain. In detail, product design and development, underwriting/pricing, sales and distributions, as well as policy and claims management are all subject to fundamental change in the future (see, e.g., Cappiello, 2020; Meier & Stormer, 2012; Rayport & Sviokla, 1995; van Rossum et al., 2002). Numerous opportunities like a facilitated interaction with customers via mail, chatbots, and social media or cost reduction via automation and standardization of business processes are challenged by few risks like the depersonalization of the insurer-customer relationship (Cappiello, 2020). However, there is only little empirical evidence on the relation between digitalization and the market valuation of insurance companies.[4] Our work contributes to the current literature on digitalization and firm valuation in several ways. First, we propose a novel approach to quantify

[2]However, we also consider further reference documents in the robustness checks, for example, Cappiello (2020) and Nicoletti (2016) as well as Bohnert et al. (2019) with the empirical study being removed.

[3]An excellent review of this literature can be found in Lowry et al. (2016).

[4]One of the few empirical studies in this field, as mentioned above, is Bohnert et al. (2019).

the impact of digitalization on the insurance sector that can also be applied in any other empirical studies based on textual analyses. Second, by employing LDA, we overcome standard pitfalls that arise from text mining approaches. In fact, we do not rely on dictionary methods where the precise word lists depend on the researchers' discretion. Instead, the thematic structure within our collection of annual reports is identified via an unsupervised machine learning algorithm. Finally, we add to the new strand of literature on LDA by making use of the whole distribution of topics instead of just focusing on a particular topic. This allows us to differentiate between digitalization and innovation based on a medium number of topics. At the same time, our approach is less prone to topic splits (cf. Bellstam et al., 2020).

The remainder of the paper is structured as follows. Section 2 explains in detail the theoretical background of LDA and its implementation in the context of digitalization in the insurance sector. In Section 3, we describe and analyze our data and present the empirical strategy using multivariate ordinary least squares (OLS). Section 4 reports the estimation results including alternative specifications based on different topic distributions, sentiment subsamples, different calculations of our digitalization measure, and other reference documents on digitalization. Finally, Section 5 summarizes and gives concluding remarks.

2 | MEASURING DIGITALIZATION USING LDA

We use the LDA method due to Blei et al. (2003) as well as natural language processing techniques to automatically analyze annual reports of insurance companies in the United States. In LDA, unsupervised machine learning is used to obtain a finite set of topics frequently discussed in the annual reports along with the fraction of time each topic is covered in each of the reports. We use these information to derive a measure of digitalization. While Sections 2.1–2.4 are optional and can be skipped by the reader who is familiar with the techniques, the construction of our measure is described beginning from Section 2.5.

2.1 | Data preprocessing

To obtain meaningful topics it is very important to preprocess the raw text data before applying any model to them. This is done to reduce the vocabulary to a set of meaningful words that are likely to provide information about the topics and concepts of interest. This facilitates the derivation of meaningful topics that best fit the context of the annual reports. The preprocessing steps we conduct are standard (cf. Bellstam et al., 2020; Hansen et al., 2018; Lopez-Lira, 2019) and shortly outlined below.

After having extracted the plain text from the annual reports,[5] we start by lowercasing all letters. We then remove common stopwords, that is, words that are commonly used but do not bear a contextual meaning (e.g., "and," "or," "the," "of").[6] We continue by removing all one-letter words like "a" and "i" because these words are frequently used to itemize lists (cf. Lowry

[5]We retrieved the plain text via the Xpdf extraction engine.

[6]We thank Bill McDonald for providing lists of common stopwords on his website https://sraf.nd.edu/textual-analysis/resources/.

et al., 2016). We also exclude all numbers as our focus is on a qualitative analysis of the annual reports. Additionally, we remove all special characters and email addresses.

There are many different words that have the same meaning, for example, "technological" and "technology," but might be treated differently by the topic modeling algorithm. To avoid this we use a standard technique called *stemming* to derive groups of words with a similar meaning. We rely on Porter's algorithm (Porter, 1980) as the most common algorithm for stemming English documents that has proven to be empirically very effective (Manning et al., 2009). The algorithm essentially consists of five phases of word reduction which are applied sequentially to a text corpus. Word reduction is achieved by applying predefined rules, for example, removing "ing" at the end of words. In each of the phases there are different conventions on how to select rules. We use the implementation of Porter's algorithm readily available in the *tm* R-package (Feinerer & Hornik, 2020).

Finally, we exclude all words that appear less than 25 times in the whole data set. These words are on average used about once every 30 annual reports and are therefore very unlikely to contribute meaningfully to any of the discovered topics. Furthermore, these words might be due to errors in parsing the PDF documents, for example, a missing letter or whitespace character. By removing those rarely used words we are able to reduce noise in our derivation of topics and sparsity of the document term matrix, see Section 2.2.

We now have a list of meaningful words for every document. In the language of text mining these collections of words are referred to as a text corpus. For further analyses this text corpus is transformed into a document term matrix as outlined in the following subsection.

2.2 | Document term matrix

To apply the LDA methodology to the preprocessed text corpus we have to transform it into a structure that can be utilized by a statistical model. We therefore make use of the so-called *bag of words* approach. The underlying assumption is that the ordering of the words in a text is negligible so that it can be represented by a vector of word counts. That is, the bag of words approach is only concerned about how often specific words occur in a document while the place of occurrence is not considered. Furthermore, the specific ordering of the documents in a text corpus is assumed to be insignificant (Blei et al., 2003). At the cost of losing the word ordering we gain the possibility to apply powerful statistical models to our annual reports that are able to derive context not only within a document but even across documents (cf. Bellstam et al., 2020). The representation of documents as vectors of word counts is, however, not only fundamental to topic modeling but lies on the basis of a variety of information retrieval algorithms such as document scoring in a query, document classification, and document clustering (Manning et al., 2009).

By combining the individual vectors of word counts, we obtain a so-called document term matrix, where each row corresponds to a specific document and each column to a specific word. A simplified example can be found in Figure 1. For our sample of annual reports we obtain a matrix with 11,440 columns highlighting the importance of the preprocessing steps especially in reducing the dimensionality of the document term matrix. Document term matrices are typically very sparse because they contain the counts of words being used across the whole sample of annual reports (cf. Hansen et al., 2018). However, the whole vocabulary is not used in each individual report. Therefore, even after removing some of the most rarely used words in the preprocessing steps, our document term matrix only has 17% nonzero entries.

impact	digit	financi	insur	affect	valu	develop	manag	mobil	firm	communic	technolog
1	0	0	1	1	0	0	0	1	2	0	0
0	1	1	0	1	1	0	0	0	1	0	2
1	0	0	1	0	0	1	2	0	1	1	0
1	1	0	0	0	2	1	0	1	0	2	1

FIGURE 1 A simplified example of a document term matrix. This figure, adapted from Lopez-Lira (2019), shows a simplified example of a document term matrix with 4 documents and 12 terms, see Section 2.2 for details. Before construction of the matrix, stopwords are removed and words are stemmed. For more information on the preprocessing steps we refer to Section 2.1 [Color figure can be viewed at wileyonlinelibrary.com]

In this context, topic models such as LDA can be understood as very powerful dimensionality reduction techniques.

2.3 | LDA

Topic models can be used to analyze large data sets of texts that are often unstructured (Roberts et al., 2016). They are probabilistic models that produce a finite set of common topics that best represent a collection of documents whereby each topic itself can be represented as a distribution over words. Each document typically covers multiple topics. By applying a topic model to a specific document we obtain a vector of topic loadings representing how intensively each topic is discussed in the respective document. This yields essentially a probabilistic representation of the document.

In this paper we use LDA, a topic model developed by Blei et al. (2003).[7] The unsupervised machine learning algorithm LDA models each document in a text corpus as a finite mixture over an underlying set of latent topics. The topics are derived from the sets of words that group together in and across the annual reports. To derive the latent structure (the discussed topics) from the observed data (the words) generative models like LDA postulate a complex latent structure being responsible for the observed data. By employing statistical inference this latent structure can then be recovered (Griffiths & Steyvers, 2004). While the topics arise endogenously from the LDA algorithm, the number of topics has to be specified in advance. The selection of an optimal number of topics is discussed in Section 2.6.

Thinking of the documents as discrete distributions over topics which themselves are distributions over words can be seen as a matrix factorization of the document term matrix (Arun et al., 2010). That is, the document term matrix, containing per document counts of specific words, is factorized in a matrix mapping topics to documents and a matrix mapping words to topics. From this perspective, LDA can be understood as a type of principal component analysis

[7]Algorithms for topic modeling can be adapted to other kinds of discrete data and have been successfully applied to genetic data and social networks (cf. Blei, 2012).

(cf. Blei, 2012) that reduces the dimensionality of each document from a vector of thousands of words to a vector of the number of topics. However, as in the LDA model topics are understood as probability distributions over words, most of the information can be preserved in the factorized matrices.

More formally, LDA is a three-level hierarchical Bayesian model that relies on a generative process for each document D in a text corpus. First, the length N of the document is determined according to a Poisson distribution, then the parameter θ is chosen from a Dirichlet distribution. This parameter governs the distribution of topics in the document and is used to specify the particular topic z_n from which an individual word ω_n is generated. More exactly, z_n is chosen according to a multinomial distribution with parameter θ. Finally, a word ω_n is drawn from the multinomial distribution of the topic z_n. This procedure is repeated for all words in a document and for all documents in a text corpus.[8] This is essentially a two-stage process. First, generate a distribution over topics. Second, choose a specific topic from this distribution over topics and generate a word from the corresponding distribution over words. The intuition behind this is that a document usually covers multiple topics and that different topics use certain words in different frequencies.[9] A representation of the underlying generative process in the language of graphical models can be found in Figure 2.

The parameters underlying the model can be determined for instance by variational expectation-maximization (VEM) or Gibbs Sampling.[10] We rely on the R-package *topicmodels* that provides an interface to the C code by Blei et al. (2003) for estimating a LDA model based on the VEM algorithm.

For our digitalization measure we make use of the fact that each document has its individual distribution of topics providing a low-dimensional representation of each document. To infer topic loadings we are not restricted to the set of documents that we have applied the LDA to. Instead, we are able to first derive the topics that are common over all annual reports of the insurance companies in our sample by applying the LDA algorithm only to these documents. Afterwards, we can utilize this set of topics as well as the representation of the topics (probability distributions over words) to infer how intensively these topics are discussed in previously unseen documents. These might be for instance additional annual reports or in this paper the reference document on digitalization.

The LDA model has many advantages, especially over simple dictionary methods. First of all, there is no need to provide any lists of words as the underlying machine learning algorithm determines the terms that are most important to discriminate between documents and topics in an unsupervised fashion. That is, the topics and corresponding word distributions arise endogenously from the data. Additionally, in contrast to the word list approach each topic derived via LDA consists of the same set of words. However, the topics differ in the probabilities they assign to each word. The possibility that multiple topics can be responsible for the words in a single document gives a lot more flexibility in modeling textual data. It corresponds to the intuition that the same word might be used in different contexts and thus also in different topics. Hansen et al. (2018) give the example of the term "growth" being used to describe economic activity while at the same time "wage growth" might appear in a context of inflationary pressure. By modeling the interaction between documents and topics with the

[8] For all technical details see the original paper by Blei et al. (2003).

[9] Note that each topic consists of the same vocabulary. However, conditional on the topic, each word occurs in a different frequency.

[10] An introduction to VEM and Gibbs Sampling can be found in Wainwright and Jordan (2008) and Gelfand and Smith (1990), respectively.

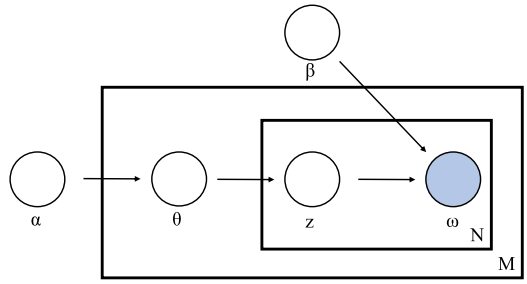

FIGURE 2 Graphical representation of LDA. This figure presents a graphical model of the LDA adapted from Blei et al. (2003). M is the set of all documents in a collection. The inner rectangle represents the ith document in the collection where N denotes the number of words of the specific document. The jth word in this document is generated from the random topic z where the topic is chosen from the specific topic distribution for that document specified by the parameter θ. The parameters α and β are priors that control the sparsity of topics within a document and the sparsity of words within a topic, respectively. Essentially, these parameters specify how much topics are needed to describe a document and how much words are needed to describe a topic. For more details on the LDA we refer to Section 2.3. LDA, Latent Dirichlet Allocation [Color figure can be viewed at wileyonlinelibrary.com]

probabilistic generative process underlying the LDA model, one can account for the usage of the same words in different contexts.

There are, however, also some disadvantages that are inherent to the LDA approach. Like most other topic models, the LDA model is based on the bag of words approach, that is, the order of words within a document is not considered but only the number of occurrences. Therefore, the context of specific words can only be derived based on the words that frequently occur within the same document while the distance between words is neglected. Nevertheless, although the exchangeability assumption behind the bag of words approach is unrealistic, it is reasonable in a context of assessing the coarse semantic structure of documents (see Blei, 2012). Another disadvantage is that, unlike the topics themselves, the number of topics does not arise endogenously from the data but has to be specified by the researcher in advance.[11] Other potential weaknesses are the exchangeability assumption on the documents within a collection and that relationships between topics are not considered (see Blei et al., 2010). In an overall view of the advantages and disadvantages, we are convinced that the LDA model as the "simplest" (Blei, 2012) and "most common topic model currently in use" (Lopez-Lira, 2019) is most appropriate in our context where we use the topics and corresponding probabilities as a low-dimensional representation of the documents.

2.4 | Other text modeling methods

In this section, we will shortly discuss other methods for the analysis of textual data in finance and insurance. Among the most common and simple approaches to automated content analysis are dictionary methods, where a list of words related to a specific topic is defined by the researcher. The prevalence of this topic in a document is then simply derived based on the

[11] The choice of the optimal number of topics is discussed in Section 2.6.

number of occurrences of the list entries. This approach can of course be extended to account for multiple topics. Word lists are an intriguingly simple approach to textual analysis. This comes, however, at the price of subjectivity as the words in the dictionary have to be specified by the researcher. Additionally, a broad concept like digitalization can hardly be captured by just counting words like "computer" or "IT" and providing a complete list of words poses a very complicated if not impossible task for the researcher. Furthermore, in dictionary methods all entries in a word list are assumed to be of equal importance.

A more sophisticated word list-based approach is provided by the key word in context (KWIC) concordances (see Gries & Newman, 2013) where not only the counts of specific words but also the words in their direct proximity are considered. The main advantage of this approach is that the immediate context in which words of interest are used is taken into account. For example, Bohnert et al. (2019) employ this method to analyze to which extent companies address digitalization in the context of external and internal stakeholders. However, the main issue with dictionary-based approaches that lists of words have to be specified by the researchers is not resolved.[12] Apart from that, we do not study the context in which digitalization is addressed in companies but rather the amount to which they digitalize making the LDA method more appropriate in our study.

Topic models, with LDA being the most common one, provide another approach to automated text analysis. An early and rather simple topic modeling method is Latent Semantic Analysis (LSA, Deerwester et al., 1990) that at its core is a principal component analysis performing a singular value decomposition on the document term matrix to extract the most informative dimensions. The probabilistic nature of LDA and the flexibility that comes from it is an important distinguishing feature from this dimensionality reduction technique. In fact, LDA was introduced to fix an issue with a probabilistic extension of the LSA method (probabilistic LSA, Hofmann, 1999).

LDA can be used as a module in more complicated models to relax some of its assumptions discussed in the previous section. For example, the topic model due to Wallach (2006) generates words inside a topic conditional on the previous word and Griffiths et al. (2005) propose a composite model combining LDA with a Hidden Markov Model (HMM) to account for short-range dependencies between words. Another extension is the hierarchical LDA (hLDA) by Blei et al. (2010). It introduces a hierarchy of topics by including the nested Chinese restaurant process in the generative model. However, the internal nodes of the resulting topic tree are not summaries of their children, that is, high probability words of a node do not necessarily coincide with high probability words of its children making interpretation difficult. Additionally, as in the LDA model, the number of topics still has to be specified by the researcher in advance.[13] In the structural topic model (STM) due to Roberts et al. (2016) covariates external to the respective document can enter into the model to allow for interactions between covariates and the topics. Grace (2019) uses this feature to analyze how topics in the 10-Ks vary with information about the company as well as over SIC code sector and time. However, as opposed to Grace (2019) we do not study how digitalization evolves depending on exogenous variables. Instead, we analyze how digitalization affects firm valuation. A summary of advantages and disadvantages of selected models can be found in Table 1. Of course, there exist more extensions of the LDA model,[14] see Blei (2012) for a review.

[12] Actually, by requiring not only to specify keywords but also additional lists of words to be analyzed in the direct proximity of the words of interest and the maximum distance between words to be considered, this method necessitates even more discretionary choices by the researcher.

[13] The hLDA model should not be confused with the Hierarchical Dirichlet Process (HDP, Teh et al., 2006) where the number of topics can be unbounded and learned from the data. However, in the HDP model the term "hierarchical" refers to the generative process and not to the topics which are flat clusterings.

[14] There even exist extensions to the computer vision field like the spatial LDA model by Wang and Grimson (2007) that are able to account for the spatial and temporal dependence of "visual words" in pictures and videos.

TABLE 1 Advantages and disadvantages of selected text modeling methods

Method	Advantages	Disadvantages
Word list approach	Simple	Context in which words appear is not considered
	Easy to interpret	Word lists depend on the researcher's discretion
KWIC concordances	Accounts for the context in which words of interest are used	Requires the researcher to specify additional words to be studied in the proximity of the words of interest
	Still simple and easy to interpret	Introduces more subjectivity than the word list approach
LDA	"Simplest" (Blei, 2012) and "most common topic model currently in use" (Lopez-Lira, 2019)	Accounts for context only via words that frequently occur in the same documents while the order of words within a document is neglected (bag of words approach)
	Dimensionality reduction technique that provides low-dimensional representations of documents	Number of topics must be specified by the researcher
	Accounts for the usage of the same words in different contexts by allowing multiple topics to be responsible for the same word in a single document	Interpretation is more complicated than in word list-based approaches but easier than in many other topic models
	Can be used to construct more complicated models	Topics are assumed to be flat clusterings
	Implementations available in many different programming languages	Higher computational effort, but feasible on modern computers
hLDA	Provides a hierarchy of topics	Most disadvantages of LDA remain valid
	Topics are organized in a tree with more general topics being located near the root and more specialized topics near the leaves	High probability words of a node do not coincide with high probability words of its children making interpretation difficult
		Generative process is more complex
		More hyperparameters have to be specified by the researcher
STM	Covariates external to the respective document can enter into the model	Most disadvantages of LDA remain valid
	Allows for interactions between covariates and topics	Expands the parameter space and the complexity of the generative process
		Introduces subjectivity in the choice of covariates

Note: Advantages and disadvantages of selected text modeling methods discussed in Sections 2.3 and 2.4 are summarized. Of course the strengths and weaknesses of the models have to be weighted against the background of the specific application.

Abbreviations: hLDA, hierarchical LDA; KWIC, key word in context; LDA, Latent Dirichlet Allocation; STM, structural topic model.

Overall, these extensions relax assumptions of the LDA model, allow for the inclusion of additional covariates or model relationships between the topics. However, although some of these methods exhibit improved language modeling performance especially (which is especially important in language generation), this comes at the price of an expanded parameter space and more complexity which can complicate interpretation. From a more practical perspective, there exist implementations of LDA in many different programming languages. Unfortunately, the same is not true for many of its extensions. Finally, Blei (2012) argue that the bag of words assumption underlying LDA is reasonable when uncovering the coarse thematic structure of documents on which our measure of digitalization is build.

2.5 | A text-based measure of digitalization

This section outlines how our measure of digitalization is derived from the annual reports of US insurance companies. The construction of our text-based measure is motivated by the approach of Bellstam et al. (2020).

By employing LDA we obtain 45 topics[15] that best describe the distribution of empirical word groupings across our sample of annual reports of US insurance companies. As has been outlined before, LDA is a dimensionality reduction technique that essentially reduces the dimension of a document from the number of different words to the number of topics. The topics arise endogenously from the data based on an unsupervised machine learning algorithm. As a result, we obtain a discrete probability distribution over 45 topics for each of the reports. This distribution corresponds to how intensively each topic is covered in the respective annual report. According to Blei et al. (2003) "the topic probabilities provide an explicit representation of a document." We build on this low-dimensional representation of our annual reports to derive a text-based measure of digitalization.

Therefore, we compare the topic distribution of each annual report, that is, the extent to which each topic is covered in the respective report, to a reference document about digitalization in the insurance sector (Bohnert et al., 2019).[16] Although the reference document about digitalization has not been presented to the LDA algorithm, the structure of the underlying generative process allows us to derive the distribution of the previously identified 45 topics in the digitalization document. Based on the distribution of topics in the reference document and in the annual reports we can calculate a measure of similarity between the digitalization document and the reports. The intuition behind this is that an annual report covering similar topics as the digitalization document extensively is more likely about digitalization than an annual report that discusses these topics only marginally.

We take a similar approach to Lowry et al. (2019) and Bellstam et al. (2020) and quantify the similarity between different topic distributions based on the KL divergence (Kullback & Leibler, 1951). The KL divergence measures how much information is lost when one uses the distribution of topics in a particular report to proxy for the distribution of topics in the digitalization document. The reports with a low KL divergence are therefore most likely discussing topics related to digitalization whereas the ones with a large divergence are not. This reasoning is supported by Figure 3 where we present the KL divergence between our reference document

[15]The number of topics is a hyperparameter that has to be specified by the researcher in advance. The optimal choice of the number of topics is discussed in Section 2.6.

[16]In Section 4.4 we also consider different documents about digitalization. The results remain qualitatively unchanged.

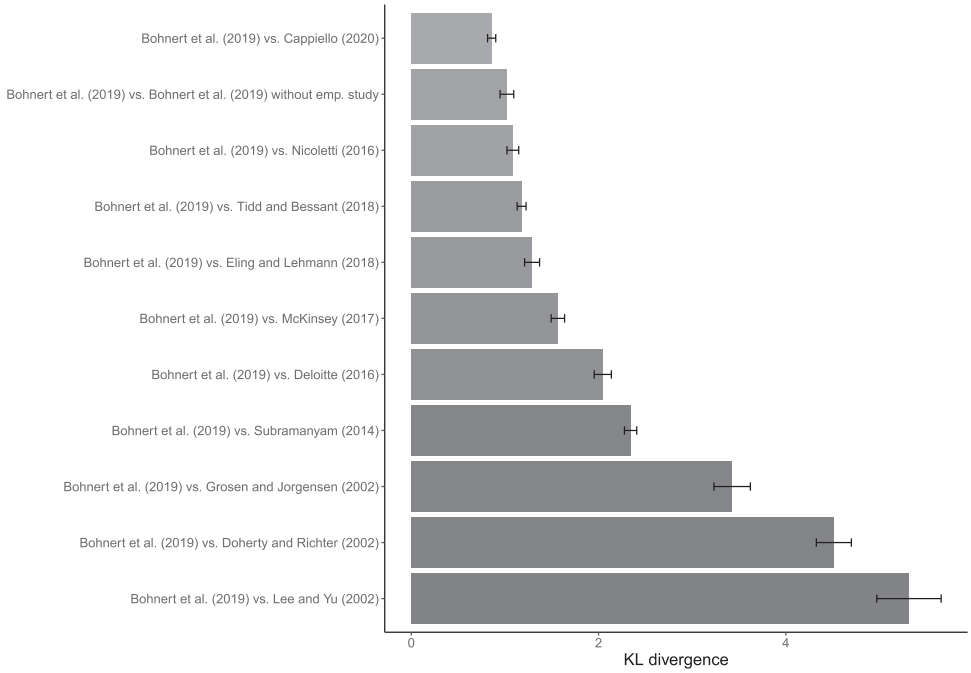

FIGURE 3 The topic distribution as a representation of a document. This figure compares the distribution of topics in our reference document on digitalization (Bohnert et al., 2019) to various other documents on digitalization (Bohnert et al., 2019; Cappiello, 2020 with the empirical section being excluded; Eling & Lehmann, 2018; Nicoletti, 2016 and the white papers McKinsey, 2017 and Deloitte, 2016), on innovation (Tidd & Bessant, 2018), on financial statement analysis (Subramanyam, 2014) as well as randomly chosen papers from the oldest available issue in the public archive of *The Journal of Risk and Insurance* (Doherty & Richter, 2002; Grosen & Jørgensen, 2002; Lee & Yu, 2002) covering unrelated subjects. The dissimilarity between the respective topic distributions is measured by the KL divergence where more similar documents exhibit a lower KL divergence value. We provide the means of 100 bootstrap samples that we obtain by repeatedly sampling 90% of the annual reports at random and applying the LDA method with 45 topics. Based on the topics obtained in each bootstrap sample, we derive the distribution of topics in the various reference documents that have not entered into the estimation process of the LDA. The bands provide 95% confidence intervals for the mean computed based on the bootstrap samples. KL, Kullback–Leibler; LDA, Latent Dirichlet Allocation

on digitalization and various other documents on digitalization (Bohnert et al., 2019; Cappiello, 2020 with the empirical section being excluded; Eling & Lehmann, 2018; Nicoletti, 2016 and the industry white papers by McKinsey, 2017 and Deloitte, 2016), on innovation (Tidd & Bessant, 2018), and on financial statement analysis (Subramanyam, 2014). To provide additional robustness we also include three randomly chosen papers from the oldest available issue in the public archive of *The Journal of Risk and Insurance* (Doherty & Richter, 2002; Grosen & Jørgensen, 2002; Lee & Yu, 2002) that are completely unrelated to the topic at hand. The analysis is based on 45 topics.

The mean KL divergences between the topic distributions of our reference document on digitalization and some other documents on digitalization (Cappiello, 2020; the paper by Bohnert et al., 2019 with the empirical section being removed; and the book by Nicoletti, 2016) are the lowest. The KL divergence between our reference document on digitalization and the document on innovation (Tidd & Bessant, 2018) is also relatively low. This is not surprising as

digitalization and innovation are related concepts. However, while the higher level of dissimilarity between Deloitte (2016) and our reference document on digitalization can be explained by the fact that this white paper also covers the more general subjects disruption and innovation, the higher mean dissimilarity between our reference document on digitalization and McKinsey (2017) and Eling and Lehmann (2018), respectively, illustrates that discriminating between digitalization and innovation is difficult and has limitations.[17] The KL divergence between the topic distributions in our reference document on digitalization and the document on financial statement analysis by Subramanyam (2014) is quite large. That is, even though the topics are derived from annual reports, the topic distributions are well suited to differentiate between digitalization on the one and general finance language on the other hand. To provide a more complete picture, we added three papers from The Journal of Risk and Insurance covering completely unrelated topics. As expected, we obtain very high KL divergences signaling a large dissimilarity.

For a specific annual report the text-based digitalization measure is calculated as the reciprocal of the KL divergence between the topic distribution in the digitalization document and the topic distribution in the particular annual report. We calculate the reciprocal to obtain a measure that is high for more digitalized and low for less digitalized companies.[18] For convenience, the measure is then linearly scaled to the interval [0, 1], that is, the observation with the lowest value of the digitalization value assumes the value 0 while the observation with the highest digitalization value takes on the value 1.[19] A summary of the whole process for deriving the measure of digitalization can be found in Figure 4.

We depart from Bellstam et al. (2020) particularly by deriving our text-based measure of digitalization from the *distribution of topics* within each document while the measure of innovation in Bellstam et al. (2020) is based on the *loadings on a specific topic* (the "innovation topic") across all documents. Bellstam et al. (2020) identify the "innovation topic" by calculating the KL divergence between the distribution of *words* in a reference document on innovation (Tidd & Bessant, 2018) and the distribution of words in the topics derived via LDA. The topic with the lowest KL divergence is assumed to be the "innovation topic" and the innovation measure is defined as the loadings on the innovation topic across documents. For illustrative purposes, we provide a word cloud for the "digitalization topic"[20] derived

[17]When we base the analysis on the five most pronounced topics (out of 45) in Bohnert et al. (2019) accounting for nearly 90% of the paper, the mean KL divergences between our reference document on digitalization and the other documents on digitalization are the lowest, followed by the document on innovation and the remaining documents. However, to avoid the introduction of an additional parameter (the number of the most prominent topics to consider), our main measure is based on the whole topic distribution. Nevertheless, we include regression results for a digitalization measure based on the five most prominent topics in Bohnert et al. (2019) in Table 8.

[18]The main measure in Bellstam et al. (2020) is based on a fourth-root transformation. For parsimony, we do not further transform our measure. However, the empirical results remain qualitatively unchanged when we replace our original measure by its fourth root.

[19]In the following we provide some exemplary excerpts from the annual reports within the top decile of firm-year observations according to our text-based digitalization measure: "Responding to higher customer expectations, we recently introduced our mobile app, which makes it easier for our customers to find their nearest agent or repair shop from their mobile device, and a new online environment for policyholders to manage their policies. We have also made improvements to our online quoting interface to make the process of buying insurance easier" (Infinity Property and Casualty Corporation, annual report 2012). "Expanding use of third-party data and analytics will identify profitable growth opportunities. With the amount of consumer data available, tapping in to 'Big Data' supports targeted marketing efforts based on educator household characteristics. This is the initial step in a multi-year strategy to identify the most efficient ways to access preferred segments of the educator market" (Horace Mann Educators Corporation, annual report 2014). "More than 23 million consumers have access to Rally, our online digital health portal. Users are steadily advancing in selecting primary care physicians, making better use of urgent care over emergency care and more readily adopting personal health and condition management programs" (UnitedHealth Group, annual report 2015)

[20]Note that our proposed digitalization measure is not based on a specific topic (the "digitalization topic") and that the word cloud is provided for illustrative purposes only. Additionally, we want to point out that when deriving topics via LDA, the same word appears in multiple topics. More exactly, all topics obtained by a LDA consist of the same set of words. However, the topics differ in the probabilities they assign to each word. It is therefore not surprising that (stemmed) words like "compani," "insur," or "financi" appear frequently in the "digitalization topic." This does not mean that these words are more influential in the "digitalization topic" but merely reflects the overall commonality of these terms in the annual reports of US insurance companies. The relative importance of single words can be better assessed based on the ratio between the frequency of a particular word in the "digitalization topic" and the overall

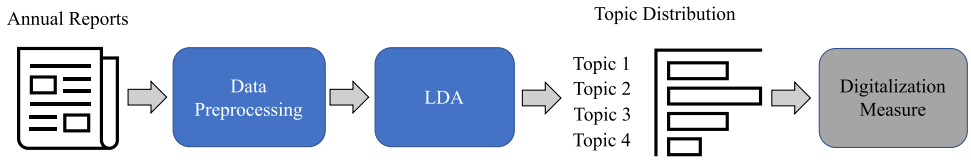

FIGURE 4 Construction of the text-based measure of digitalization *Notes*: This figure, freely adapted from Lopez-Lira (2019), shows the main steps to derive our text-based measure of digitalization (see Section 2 for details). First, we preprocess the annual reports of US insurance companies (removing stopwords, stemming, etc.) and construct a document term matrix. Then, we employ LDA to derive a distribution over topics for each of the reports. These topic distributions are subsequently used to calculate the measure of digitalization [Color figure can be viewed at wileyonlinelibrary.com]

analogously to this approach but based on the reference document on digitalization (Bohnert et al., 2019) in Figure 5.

We can, however, not simply apply this approach analogously to measure digitalization. When we replace the document on innovation by our reference document on digitalization we identify the same topic. That is, when adapting the approach by Bellstam et al. (2020), the "innovation topic" and the "digitalization topic" are identical. Again, this is not surprising as innovation and digitalization are related topics. However, choosing the same topic as

FIGURE 5 Word cloud of the "digitalization topic." This figure shows the 50 most frequent (stemmed) terms in the "digitalization topic." This topic is determined according to Bellstam et al. (2020). That is, the word distributions in each topic obtained via a 45 topic LDA are compared with the word distribution in the reference document on digitalization by Bohnert et al. (2019). We use the KL divergence as a measure of similarity between these distributions and choose the topic with the lowest KL divergence as the "digitalization topic." Note that our proposed text-based digitalization measure is not based on a specific topic (the "digitalization topic"). Instead, we derive our measure by exploiting the whole topic distribution, see Section 2.5 for details. The word cloud is provided for illustrative purposes only. KL, Kullback–Leibler; LDA, Latent Dirichlet Allocation

digitalization and innovation topic would imply both measures to be defined as the loadings across documents on the *same* topic. Consequently, the digitalization and the innovation measure would be identical. Of course, one could simply increase the number of topics until the innovation and the digitalization topic are different.[21] However, that causes another problem apart from adding more complexity to the model. The vector of loadings of all documents on a specific topic obtained via LDA is approximately orthogonal to the vector of loadings on any other topic. This stays true when the number of topics is increased (cf. Arun et al., 2010). This means that even if one could identify different innovation and digitalization topics, the innovation and the digitalization measures according to Bellstam et al. (2020) would exhibit a correlation of approximately 0. This would, however, contradict the intuition that innovation and digitalization are related concepts and would essentially represent a measure of digitalization while controlling for innovation.

Another problem that arises from measuring a concept of interest based on a single topic is the following: When increasing the number of topics, at some point a particular topic might split into two aspects of the same concept. This is, for instance, observed in Bellstam et al. (2020) in the case of 50 topics. The measures derived from the split topics would again have a correlation of nearly 0. Consequently, a measure based on a single topic might completely miss a part of the concept of interest. This problem does not emerge with our measure based on the distribution of topics rather than words.

2.6 | Optimal number of topics

While the topics and the topic loadings arise endogenously from the LDA algorithm, the number of topics plays the role of a hyperparameter. In machine learning, this is a parameter that controls the learning process and has to be specified in advance. In contrast, the topic distributions within each document and the word distributions within each topic are parameters that are derived by the unsupervised machine learning algorithm via the training process. Therefore, before employing the LDA method, the researcher has to provide the number of topics to be extracted from the documents by the algorithm. On the one hand, the number of topics has to be sufficient to distinguish between different themes in the document, but on the other hand should not be too high to ensure interpretability of the topics (Lopez-Lira, 2019).

There exist several methods for determining the number of topics in a data-driven manner, for example, approaches introduced by Griffiths and Steyvers (2004), Cao et al. (2009), Arun et al. (2010), and Deveaud et al. (2014). These technical methods rely on different objective functions to be minimized or maximized to best describe the underlying text corpus. Not surprisingly, these methods do not agree on the optimal number of topics in our application. Furthermore, some of them suggest even more than 100 topics whereby interpretability of the topics would be lost and overfitting issues might arise. According to Blei (2012), "develop[ing] evaluation methods that match how the algorithms are used," particularly for determining the optimal number of topics, is still an open direction for research in topic modeling. In contrast to

frequency across all reports. To provide some examples, these ratios are smaller than 1 for "compani" (0.96), "insur" (0.90), and "financi" (0.90). This illustrates that even though these terms appear frequently in the "digitalization topic," they are less frequently used than in the average annual report. On the contrary, terms like "digit" (7.74), "mobil" (5.44), "internet" (5.06), and "autom" (5.13) occur clearly more often in the "digitalization topic" than in the average annual report.

[21] Even for 30 and 45 topics the digitalization and the innovation topic stay the same. The issue remains when we identify the digitalization topic based on another reference document on digitalization (Cappiello, 2020).

the technical methods for topic number selection, we do not intend to fully capture every aspect of the documents. Instead, we employ LDA to measure the extent to which digitalization is discussed in the reports. Although there is no clear guidance in the literature on how to select the optimal number of topics as most researchers apply different rules based on the task at hand, in many applications a number of topics between 10 and 50 seems to be appropriate (see, e.g., Bao & Datta, 2014; Bellstam et al., 2020; Ganglmair & Wardlaw, 2017; Weiss Hanley & Hoberg, 2019; Israelsen, 2014; Lopez-Lira, 2019).

Bellstam et al. (2020) use 15 topics to derive a text-based measure of innovation. Digitalization can be understood as a more specific concept than the broader notion of innovation. We therefore also consider 30 and 45 topics to resemble the thematic structure of the collection of annual reports more granularly. As innovation and digitalization are related topics we choose the number of topics such that we are best able to differentiate between these two concepts. Our measure of digitalization is based on the distribution of topics within each annual report. This distribution of each particular report is compared to the topic distribution in a reference document on digitalization (Bohnert et al., 2019). To differentiate between digitalization and innovation, we choose the number of topics such that the mean dissimilarity between the topic distribution in the reference document on digitalization and the document on innovation (Tidd & Bessant, 2018) is maximized. This leads to a number of 45 topics being most appropriate for measuring digitalization even when we additionally analyze 60 topics.[22] While Bellstam et al. (2020) rely on 15 topics to measure innovation, this number is too low to sufficiently capture the thematic structure of the annual reports in our sample to measure digitalization. This is manifested in the fact that for 15 topics the mean KL divergence between the documents on digitalization is not statistically significantly lower than the mean KL divergence between the document on digitalization and innovation. It is therefore not surprising that we partly yield insignificant results when including this measure based on 15 topics in our regression framework. However, additional to our baseline measure derived from 45 topics we also consider a measure obtained by a more parsimonious LDA based on 30 and even 15 topics in our regression framework, see Section 4.2.

2.7 | Sentiment

Our text-based measure of digitalization as introduced in the previous sections does not take sentiment into account. For a report covering digitalization in a rather negative tone, our procedure might nevertheless assign a high value to the digitalization proxy although this report is less likely to represent more digitalization efforts by the firm (cf. Bellstam et al., 2020). This might induce measurement errors to our measure of digitalization.

Sentiment is most frequently measured by counting the occurrence of specific "positive" or "negative" words see (see Henry & Leone, 2016). Of course, the drawbacks related to word list-based approaches (see Section 2.4) also apply in this case. First of all, lists of positive and negative words have to be provided by the researcher. Second, the sentiment of specific words can depend on the context in which it is used. For example, Loughran and McDonald (2011)

[22]To ensure robust results, we calculate the similarity measure (KL divergence) based on 100 bootstrap samples which are computed by a LDA with 15, 30, 45, and 60 topics based on 100 subsamples obtained by repeatedly choosing 90% of the reports at random. Based on the topics we derive for each of these samples we can calculate the distribution of topics in the documents on digitalization and innovation. We then compute the similarity measure between these topic distributions. The number of topics is then selected based on the mean over the 100 bootstrap samples. These calculations are computationally very expensive and were performed on the Big-Data-Cluster Galaxy provided by the University Computing Center at Leipzig University.

find that more than 70% of the negative words in the Harvard Psychosociological Dictionary[23] are typically not negative in a financial context (e.g., board, capital, or liability). We therefore use a dictionary of positive and negative words provided by Loughran and McDonald (2011) and Bodnaruk et al. (2015) that has been specifically adjusted for financial language.[24] We follow Bellstam et al. (2020) and measure sentiment as the difference between positive and negative words divided by the total number of words. Of course, there are other dictionaries as well as other methods to measure sentiment. For more details we refer to Loughran and McDonald (2011) and Lowry et al. (2016).

To overcome potential measurement errors introduced by not considering sentiment, we exclude all firm-year observations with sentiment below the 25% quantile in a robustness check (see Section 4.3). Note that the LDA is performed based on all annual reports in our sample and firm-year observations are removed according to sentiment *after* the construction of the digitalization measure. Consequently, topics still arise endogenously from the data and are not affected by the word lists on which measurement of sentiment is based on.

3 | FINANCIAL DATA AND EMPIRICAL STRATEGY

3.1 | Sample construction

We start the construction of our sample by selecting all insurance companies in the US with stock market data available in Thomson Reuters Datastream. We use market value-based measures due to their ability to capture short-term performance and long-term prospects (Allen, 1993; Lubatkin & Shrieves, 1986). The focus on publicly-listed US insurance companies is motivated by the strong distortionary effect different regulatory and accounting standards in different countries would have in our setting. Thus, we collect annual reports for as many publicly-listed US insurers as possible. Annual reports have already been subject to textual analysis (e.g., Gatzert & Heidinger, 2020; Li, 2008; Yekini et al., 2016). In contrast to Bellstam et al. (2020), we prefer annual reports over analyst reports since the former provide first-hand information about a company's status quo, current projects, and upcoming trends. Furthermore, we consider digitalization issues to be less obvious to externals because digitalization efforts are often aimed at improving internal procedures rather than developing new insurance products.

We complement our data set with accounting data (i.e., total assets, ROA, total investment, solvency ratio, and current liquidity ratio) from Orbis Insurance Focus covering the period from 2006 to 2015. As will be described in the following Section 3.3, we lag all explanatory variables by 1 year. Thus, our initial sample consists of 86 insurance companies from 2006 to 2015 resulting in 748 observations in an unbalanced panel.

3.2 | Summary statistics

Table 2 presents the summary statistics of the sample. Our first main measures of interest, *market value*, exhibits mean and median values of $US 7.69 bn. and $US 2.33 bn. indicating a

[23]The Harvard Psychosociological Dictionary, more specifically the Harvard-IV-4 TagNeg file, is a commonly used source for word classifications.
[24]See Loughran and McDonald Sentiment Word Lists at https://sraf.nd.edu/textual-analysis/resources.

TABLE 2 Summary statistics

	Obs.	Mean	Median	*SD*	Min	Max
Market Value (bn.)	748	7.69	2.33	15.45	0.00	186.30
Market-to-Book	735	1.25	1.10	0.76	−0.79	8.52
$\text{Digitalization}_{15,t-1}$	748	0.15	0.12	0.15	0.00	1.00
$\text{Digitalization}_{30,t-1}$	748	0.16	0.07	0.18	0.00	1.00
$\text{Digitalization}_{45,t-1}$	748	0.15	0.05	0.20	0.00	1.00
$\text{Digitalization}_{60,t-1}$	748	0.19	0.08	0.22	0.00	1.00
Total Assets_{t-1} (bn.)	748	54.60	8.95	140.84	0.01	1048.36
ROA_{t-1}	748	3.42	2.96	4.60	−36.37	20.23
Investment_{t-1} (bn.)	748	32.78	5.84	87.85	0.00	806.04
Solvency R._{t-1}	748	25.69	25.73	13.08	1.92	69.46
Current R._{t-1}	748	93.40	92.12	34.09	3.42	289.77
Foreign Assets R._{t-1}	527	2.81	0.00	12.35	0.00	87.38

Note: Summary statistics on all variables used in the multivariate ordinary least squares analyses are presented. The panel spans from 2007 to 2016. The following columns present the number of observations, mean, median, standard deviation, as well as minimum and maximum value. The variable definitions and data sources are given in Appendix A.

right-skewed distribution. We account for the skewness by taking the log of *market value*. The mean and median of the second dependent variable, *market-to-book value*, are 1.25 and 1.10, respectively. Our main explanatory variable, *digitalization* based on a 45 topic distribution, ranges between 0 and 1 with mean and median of 0.15 and 0.05, respectively. Values of the digitalization measure close to 1 indicate a very digitalized insurance company whereas values close to 0 indicate the opposite. Consequently, a mean value of 0.15 emphasizes that the insurance industry on average has yet to take advantage of the full potential of digital technologies. Companies that exhibit particularly high values of digitalization over time are Horace Mann Educators, Argo, and Aflac, for instance. Their digitalization values continuously rank among the top 10%. In contrast, companies that digitalize comparatively little over time are MetLife, Enstar, or Alleghany. Here, digitalization appears to be less pronounced in annual reports with values below 0.025. Statistical moments of the three alternative digitalization measures based on 15, 30, and 60 topic distributions are similar to those of our main digitalization measure.

The time dimension of our digitalization variable is depicted in Figure 6 along with the 10th and 90th percentiles indicating a high cross-sectional variation. At first sight, the fairly stable annual mean values just below 0.2 with a peak above 0.2 in 2006 appear to be counterintuitive since one might expect digitalization to become more important over time, especially in recent years. However, taking into consideration the time span of our sample from 2006 to 2015 it rather covers what could be described as the first surge of digitalization. This first surge is characterized by improvements like e-mail alerts or investments in online resources.[25]

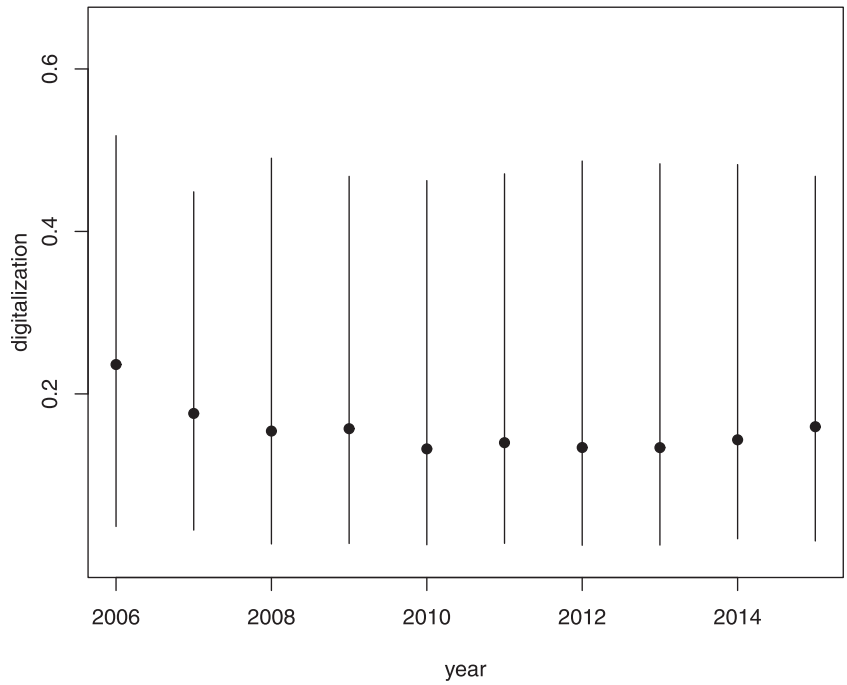

FIGURE 6 Digitalization measure over time. This figure shows the mean of the digitalization measure (based on 45 topics) between 2006 and 2015 along with the 10th and 90th percentiles. For details on the construction of our text-based measure of digitalization we refer to Section 2.5

To get a better idea of what is causing this rather constant digitalization trend on average, we next examine the average relative frequency of particular words related to digitalization in the annual reports over time.[26] The results presented in Figure 7 show that basic digital issues often expressed through terms stemming from "internet," "online," and "web" are more prominent in the first years of our sample period. In the last years of our sample period, one can recognize the beginning of a second surge of digitalization expressed through terms stemming from "data," "mobile," or "digital." In between, the average value of the digitalization proxy decreases (see Figure 6) which might be due to consequences of the financial crisis in 2007/2008 for the insurance industry including regulatory issues. Taken together, we can infer that the rather constant average trend of our measure is probably caused by the balancing effects of different digital issues across time.

In addition to the variance of the digitalization measure across time, we also investigate its cross-sectional variation with respect to business line, size, profitability, and market orientation. Table 3 presents the results of mean comparisons for life and nonlife, small and large, profitable and less profitable, and (inter-)national insurance firms.

[25] In the following we provide exemplary excerpts from the 2006 annual reports of Aflac Inc. and Principal Financial Group: "In 2006, we also tested AflacAnywhere[SM], a new technology that greatly improves communications between headquarters and our sales force. AflacAnywhere enables sales associates and coordinators to receive notification on important information via **e-mail alert**, text message on a cell phone or PDA, or computer-generated voice message to any phone number" (Aflac Inc., annual report 2006). "For customers, we've **invested in online resources**, improved technology in our contact centers and simplified communication materials" (Principal Financial Group, annual report 2006).

[26] Note that our text-based measure of digitalization is not based on particular words but on the distribution of topics in the annual reports. Figure 7 is provided for illustrative purposes only.

TABLE 3 Descriptive statistics (subsamples)

	Life		Nonlife			
Sector	**Mean**	***SD***	**Mean**	***SD***	**Difference**	***p* Value**
$\text{Digitalization}_{45,t-1}$	0.156	0.196	0.140	0.204	0.016	(.374)
Observations	159		494		653	
	Small		**Large**			
Size	**Mean**	***SD***	**Mean**	***SD***	**Difference**	***p* Value**
$\text{Digitalization}_{45,t-1}$	0.093	0.109	0.163	0.200	−0.070***	(.000)
Observations	187		187		374	
	Small		**Large**			
Profitability	**Mean**	***SD***	**Mean**	***SD***	**Difference**	***p* Value**
$\text{Digitalization}_{45,t-1}$	0.132	0.202	0.162	0.179	−0.030	(.132)
Observations	187		186		373	
	International		**National**			
Orientation	**Mean**	***SD***	**Mean**	***SD***	**Difference**	***p* Value**
$\text{Digitalization}_{45,t-1}$	0.123	0.167	0.163	0.215	0.031*	(.080)
Observations	197		551		677	

Note: Summary statistics on four characteristic variables of our sample: Sector/Business Line, Size, Profitability, and Market Orientation are presented. The panel spans from 2007 to 2016. The following columns present the number of observations, mean, standard deviation, as well as the difference in means including the *p* value. The division into life and nonlife stems from Thomson Reuters Datastream. The subsample statistics for size and profitability only consider the first (small) and fourth (large) quartile of the respective variable. *, **, and *** denote statistical significance at the 10, 5, and 1 percent levels, respectively.

Classification into life and nonlife stems from Thomson Reuters Datastream. Since some insurers cover both business lines we excluded them from this comparison. The mean values do not differ significantly which also applies to the median values (see Figure 8a). There seems to be no significant difference in the extent of digitalization across business lines.

In contrast, size seems to be of higher importance. A comparison of the first and the fourth quartile of insurance companies with respect to size shows a significant difference (see Table 3). On average, large firms exhibit higher values of digitalization (mean = 0.163) than small firms (mean = 0.093). However, comparing the median values in Figure 8b) gives a different picture since median digitalization in the smallest firms lies above the median digitalization of large firms. Hence, this evidence does not support a clear positive relation between size and the extent of digitalization in an insurance company. Furthermore, we do not find clear evidence that profitability affects the extent of digitalization in a company. Although from Figure 8c one might get the impression that the most profitable firms (top 25%) exhibit higher values of digitalization on average, there is no significant difference in means (see Table 3).

Digitalization might also be related to the market orientation of a company (national vs. international). However, using the ratio of foreign assets to total assets from Thomson Reuters Datastream as a proxy for the market orientation of an insurer, we do not find strong empirical evidence for such a relation. According to Table 3, the difference in means between the

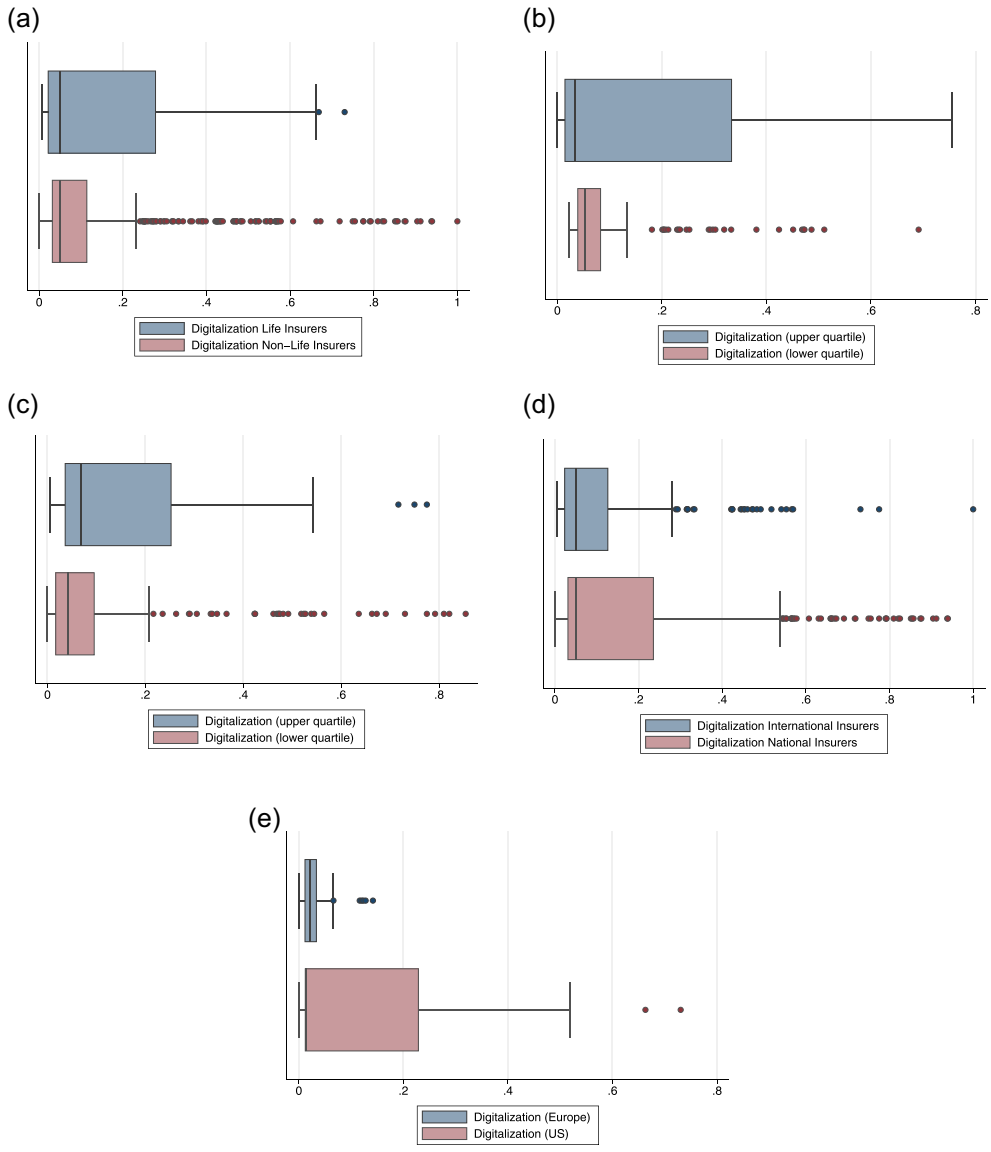

FIGURE 8 Digitalization distributions according to different criteria *Notes*: This figure presents the digitalization measures of insurance companies included in our sample with respect to five categories: business line (a), size (b), profitability (c), market orientation (d), and geographical origin (e). Classification into business line, that is, life and nonlife, stems from Thomson Reuters Datastream. We excluded insurers covering both business lines from the comparison. Classification according to size (large vs. small) and profitability (high vs. low) is made using the upper and lower quartile of companies with respect to total assets and ROA, respectively. Market orientation (international vs. national) is assessed based on the ratio of foreign assets to total assets. A positive ratio is considered to indicate an international orientation of the firm. For the comparison of geographical origin we contrast the digitalization measures of eight large European insurance companies, namely Allianz, Aviva, AXA, CNP, Generali, Mapfre, Prudential, and Zurich with those of eight similarly large US insurance companies, namely AIG, MetLife, Hartford, Lincoln National, Principal, Allstate, Ameriprise, and Aflac [Color figure can be viewed at wileyonlinelibrary.com]

TABLE 4 Summary statistics european versus US subsample

	Obs.	Mean	Median	*SD*	Min	Max
$\text{Digitalization}_{\text{European}}$	74	0.030	0.021	0.031	0.000	0.142
$\text{Digitalization}_{\text{US}}$	80	0.129	0.014	0.195	0.000	0.730

Note: Summary statistics on the digitalization measure based on a Latent Dirichlet Allocation (45 topic distribution) applied to the annual reports of eight large US and European insurance companies, respectively, are presented. The annual reports cover the period from 2006 to 2015. The European insurance companies are Allianz, Aviva, AXA, CNP, Generali, Mapfre, Prudential, and Zurich. The US insurance companies are AIG, MetLife, Hartford, Lincoln National, Principal, Allstate, Ameriprise, and Aflac. The companies are chosen on the basis of size and international focus (for the European insurers). For further explanation see Section 3.2. The following columns present the number of observations, mean, median, standard deviation, as well as minimum and maximum value. Data from 2006 to 2015. European Insurance Companies: Allianz, Aviva, AXA, CNP, Generali, Mapfre, Prudential, Zurich. US Insurance Companies: AIG, MetLife, Hartford, Lincoln National, Principal, Allstate, Ameriprise, Aflac.

international and national insurance companies in our sample is only weakly significant at the 10 percent level with national insurers exhibiting a higher mean (0.163). However, the median values appear to be fairly equal (see Figure 8d).

In a further illustrative example, we compare digitalization in the US insurance sector with the European insurance market. Therefore, we apply our LDA to the annual reports of eight large European insurance companies, namely Allianz, Aviva, AXA, CNP, Generali, Mapfre, Prudential, and Zurich. The choice of the European insurance companies is motivated by their international focus as well as their size and hence relevance for the European insurance sector. To maintain comparability, we also choose a subsample of large US insurance companies for the descriptive analysis of the digitalization measures. These are AIG, MetLife, Hartford, Lincoln National, Principal, Allstate, Ameriprise, and Aflac. The results are presented in Figure 8e and Table 4.

Although median values of the two groups are relatively similar, mean values differ substantially. On average, large US insurance companies exhibit much higher values of digitalization than their European counterparts. In detail, the mean digitalization value of the eight large US insurance companies (0.129) is four times higher than the one of the European insurance companies (0.030). This leads to the conclusion that digitalization might be of higher importance for American insurance companies whereas European insurers might lag behind.

3.3 | Empirical strategy

We analyze how digitalization of the business model is associated with the firm value of an insurance company using panel data regressions. However, estimating the relation is not straightforward due to issues of endogeneity. In fact, one could imagine that large and profitable insurance companies have more capacities to invest in digitalization than small insurers. This would lead to a situation in which digitalization not just affects market valuation, but vice versa. As a result, our estimated coefficients would be biased because of reverse causality. In addition, there is also a problem of omitted variable bias. More specifically, other independent variables omitted in the regression are likely to be correlated with both our proxies for market valuation and the main explanatory variable digitalization. If they become part of the error term, the OLS assumption of conditional mean independence will be violated resulting in

biased estimates for the effect of digitalization on market valuation. We address the potential endogeneity by making use of two basic econometric means. First, we lag all our explanatory variables by one year to make sure that they will not be affected by the current firm value and hence not be subject to estimation biases due to reverse causality. Second, we add company and year fixed effects to the regressions to account for unobserved heterogeneity across firms and time.[27] In particular, we estimate the model:

$$FirmValue_{i,t} = \beta_0 + \beta_1 Digitalization_{i,t-1} + \beta_2 X_{i,t-1} + \delta_t + u_i + \epsilon_{i,t}.$$

The index i represents a particular insurance company whereas index t denotes the year. Firm value is approximated by the *market value* or the *market-to-book value*, respectively. $Digitalization_{i,t-1}$ represents the digitalization measure derived from our LDA. As already mentioned, we find 45 topics to maximize the mean dissimilarity between the topic distribution in the reference document on digitalization and the document on innovation, thus reflecting the digitalization approach most accurately. However, in subsequent robustness checks we also run the regression with digitalization measures derived from 15 to 30 topic distributions. $X_{i,t-1}$ is a vector of control variables that are commonly used. In detail, we add the natural logarithm of total assets to account for the size of an insurer, return on assets (ROA) to account for its profitability, the natural logarithm of total investment, and the solvency as well as the current liquidity ratio to account for the short- and long-term obligations of an insurance company.

4 | ESTIMATION RESULTS

4.1 | Baseline estimation

Having laid out the estimation strategy to identify the relation between digitalization and firm outcome, Table 5 presents the results of our main estimation using our whole sample of insurance companies and a distribution over 45 topics to calculate our digitalization measure. As already mentioned, we use two proxies for firm value, that is, *market value* and *market-to-book value*. Furthermore, we control for unobserved heterogeneity using firm fixed effects (columns 1–4) and also time fixed effects (columns 2 and 4).[28] In all specifications, the relation between digitalization and market value and market-to-book value, respectively, is positive and statistically significant. However, as already mentioned in Section 3.3, our estimated coefficients might be subject to endogeneity. Consequently, establishing a causal link between digitalization and firm value is not possible unequivocally. The estimated coefficient for digitalization in the OLS regression on market value including company fixed effects is 0.374. Hence, an increase in digitalization by one standard deviation is associated with an increase in market value by about 7.48% (0.20×0.374) in the subsequent year. Similarly, an increase in digitalization by one standard deviation is related to an increase in the market-to-book value by

[27] Another method to address endogeneity is the use of instrumental variables. However, since empirical evidence on the relation between firm value and digitalization is scarce, we were not able to find a commonly accepted instrument in the literature. In separate regressions, we tried to instrument our digitalization measure using dummy variables for the presence of a digital officer and a CEO change, respectively. Unfortunately, both of them turned out to be invalid. Therefore, we leave the issue of instrumentation up for future research.

[28] Pointing toward the R^2 of each regression, it can be seen that the explanatory power of our estimation equation is fairly high which is probably due to firm fixed effects. They obviously capture a lot of variation in the data. Consequently, adding time fixed effects does not add much to the explanatory power leading to the conclusion that time trends do not play an important role in our setting.

TABLE 5 The relation between digitalization and firm valuation

	(1)	(2)	(3)	(4)
Dependent	**Market value**	**Market value**	**MtB**	**MtB**
$\text{Digitalization}_{45,t-1}$	0.374***	0.238***	0.402***	0.188*
	(0.091)	(0.084)	(0.135)	(0.110)
$\text{Total Assets}_{t-1}$	0.506**	0.207	0.901***	0.726***
	(0.231)	(0.171)	(0.255)	(0.278)
ROA_{t-1}	0.029***	0.007	0.044***	0.014*
	(0.010)	(0.008)	(0.008)	(0.007)
Investment_{t-1}	0.430**	0.590***	−0.597***	−0.348*
	(0.208)	(0.134)	(0.201)	(0.201)
$\text{Solvency Ratio}_{t-1}$	0.037***	0.035***	−0.028**	−0.017*
	(0.008)	(0.005)	(0.012)	(0.010)
$\text{Current Ratio}_{t-1}$	−0.009***	−0.009***	0.012**	0.011**
	(0.003)	(0.002)	(0.005)	(0.005)
Company FE	Yes	Yes	Yes	Yes
Year FE	No	Yes	No	Yes
Observations	748	748	735	735
R^2	0.97	0.98	0.60	0.71
F test	53.04	39.79	8.26	3.95

Note: The results of the panel regressions that examine the relation between digitalization based on a 45 topic distribution and firm value proxied by the log of *market value* and *market-to-book value* (MtB), respectively, are presented. The topic distribution in each annual report has been compared to the one in Bohnert et al. (2019). Columns (1) and (2) show the estimation results for the OLS regressions using the log of *market value* as the dependent variable. Columns (3) and (4) report the estimation results of *market-to-book value* as the dependent variable. The panel has one observation for each company-year combination, and spans the time period 2007–2016. We include company fixed effects in all specifications. Columns (2) and (4) additionally include year fixed effects. Standard errors are reported in parentheses. In all specifications standard errors are robust to heteroskedasticity. Furthermore, the model fit (R^2) and test statistics for the joint significance of regressors (F test) are reported at the bottom of the table. *, **, and *** denote statistical significance at the 10%, 5%, and 1% levels, respectively.

about 8.04% (0.20×0.402) in the next year.[29] The relation becomes slightly weaker when we add time fixed effects but remains significant. Our results provide evidence for a strong positive relation between digitalization and firm valuation. Put differently, market participants might expect insurance companies making progress in digitalization to attain higher future cashflows and become more profitable. As a result, the market value and the market-to-book value increase.

[29]Instead of considering an increase of the digitalization measure by one standard deviation, one could also examine the effect of an increase from the first to the third quartile. Such an increase of about 0.1771 is associated with an increase in market value of approximately 6.62% (R^2). For the market-to-book value the increase is about 7.12% (R^2).

4.2 | Alternative number of topics

To check for the robustness of our main estimation results, we replace our main explanatory variable by digitalization measures derived from LDA with different numbers of topics. The results are presented in Table 6. In detail, we use a 15 topic, a 30 topic, and a 60 topic LDA model, respectively. The 15 topic model has already been employed by Bellstam et al. (2020). However, in our context it rather serves as a reference since a lower number of topics might blur the line we want to establish between innovation in general and digitalization in particular.[30] Taking this into account, it is not surprising that the statistical significance of the estimated coefficients for the digitalization variable based on 15 topics (line 1) is not as strong as in Table 5. Therefore, we can infer that a more granular topic distribution is necessary to capture digitalization instead of general firm innovation. However, the estimated coefficient in the model using firm fixed effects (column 1) is of the same magnitude.

In contrast to the 15 topic model specification, the estimated coefficients for our models based on 30 and 60 topic distributions (lines 2 and 3) are predominantly statistically significant. Especially OLS estimation results using firm fixed effects exhibit estimation coefficients for the main explanatory variable that are similar to those in our main estimations (Table 5). Hence, we can conclude that our main estimations are robust to alternative calculations of the digitalization measure using different topic distributions as long as the number of topics exceeds those used to capture general innovation (see, e.g., Bellstam et al., 2020).

4.3 | Sentiment analysis

As we argue in Section 2.7, a particular report with a high value of the digitalization value is less likely to represent more digitalization by the firm when the report is written in a negative or neutral tone. In a subsample analysis, we therefore only consider firm-year observations with a sentiment above the 25% quantile.

The results presented in Table 7 for market value and the market-to-book value, respectively, are in line with those presented in Tables 5 and 6. Again, there is a positive and (highly) statistically significant relation between digitalization based on a 45 topic distribution and market value and market-to-book value, respectively. The estimated coefficients are of the same magnitude. According to this specification, an increase in digitalization by one standard deviation is associated with an increase in market value by about 7.62% (0.20×0.381) in the model with firm fixed effects (column 3). The estimated coefficients for the digitalization proxies based on 30 and 60 topic distributions are also similar to those reported in Table 6.

4.4 | Alternative specifications and reference documents

In the last part of our empirical investigation, we control for potential confounding effects due to the construction of our digitalization measure or the choice of our reference document. According to Bellstam et al. (2020), we consider a measure based on only 98% of the annual

[30] See Section 2.6 for details.

TABLE 6 The relation between digitalization and firm valuation (altered topic distribution)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Dependent	**Market value**	**Market value**	**Market value**	**Market value**	**Market value**	**Market value**	**MtB**	**MtB**	**MtB**	**MtB**	**MtB**	**MtB**
$Digitalization_{15,t-1}$	0.480***	0.325***					0.497**	0.108				
	(0.127)	(0.121)					(0.220)	(0.167)				
$Digitalization_{30,t-1}$			0.426***	0.297***					0.535***	0.267*		
			(0.104)	(0.082)					(0.190)	(0.142)		
$Digitalization_{60,t-1}$					0.371***	0.246***					0.477***	0.238**
					(0.091)	(0.070)					(0.144)	(0.106)
Total $Assets_{t-1}$	0.514**	0.207	0.524**	0.214	0.502**	0.200	0.910***	0.732***	0.921***	0.731***	0.894***	0.717**
	(0.232)	(0.173)	(0.231)	(0.171)	(0.232)	(0.172)	(0.257)	(0.280)	(0.256)	(0.279)	(0.255)	(0.279)
ROA_{t-1}	0.029***	0.007	0.028***	0.007	0.028***	0.007	0.045***	0.014*	0.044***	0.014*	0.044***	0.014*
	(0.011)	(0.008)	(0.011)	(0.008)	(0.010)	(0.008)	(0.008)	(0.007)	(0.008)	(0.007)	(0.008)	(0.007)
$Investment_{t-1}$	0.422**	0.584***	0.423**	0.583***	0.437**	0.593***	−0.606***	−0.350*	−0.604***	−0.353*	−0.585***	−0.344*
	(0.208)	(0.134)	(0.208)	(0.133)	(0.209)	(0.134)	(0.203)	(0.202)	(0.201)	(0.202)	(0.201)	(0.202)
Solvency $Ratio_{t-1}$	0.037***	0.035***	0.037***	0.035***	0.037***	0.035***	−0.028**	−0.017*	−0.027**	−0.017*	−0.027**	−0.017*
	(0.008)	(0.005)	(0.008)	(0.005)	(0.008)	(0.005)	(0.012)	(0.010)	(0.012)	(0.010)	(0.012)	(0.010)
Current $Ratio_{t-1}$	−0.009***	−0.009***	−0.009***	−0.009***	−0.009***	−0.009***	0.013**	0.011**	0.012**	0.011**	0.012**	0.011**
	(0.003)	(0.002)	(0.003)	(0.002)	(0.003)	(0.002)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)
Company FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes

TABLE 6 (Continued)

Dependent	(1) Market value	(2) Market value	(3) Market value	(4) Market value	(5) Market value	(6) Market value	(7) MtB	(8) MtB	(9) MtB	(10) MtB	(11) MtB	(12) MtB
Observations	748	748	748	748	748	748	735	735	735	735	735	735
R^2	0.97	0.98	0.97	0.98	0.97	0.98	0.60	0.71	0.60	0.71	0.60	0.71
F Test	49.47	39.48	52.69	40.44	53.09	41.40	8.16	3.98	8.17	4.01	8.62	4.24

Note: The results of the panel regressions that examine the relation between digitalization based on alternative topic distributions on firm value proxied by the log of market value and *market-to-book value* (MtB), respectively, are presented. The topic distribution in each annual report has been compared to the one in Bohnert et al. (2019). Columns (1)–(6) show estimation results for market value as dependent variable. In Columns (7)–(12), regression results for *market-to-book value* are reported. In columns (1) and (2) as well as in columns (7) and (8) the digitalization measure is based on a 15 topic distribution, whereas columns (3) and (4) as well as columns (9) and (10) were calculated on the basis of a 30 topic distribution LDA. Columns (5) and (6) as well as columns (11) and (12) report the results based on a 60 topic distribution LDA. The panel has one observation for each company-year combination, and spans the time period 2007–2016. We include company fixed effects in all specifications. Even columns additionally include year fixed effects. Standard errors are reported in parentheses. In all specifications standard errors are robust to heteroskedasticity. Furthermore, the model fit (R^2) and test statistics for the joint significance of regressors (*F* test) are reported at the bottom of the table.
*, **, and *** denote statistical significance at the 10%, 5%, and 1% levels, respectively.

TABLE 7 The relation between digitalization and firm valuation above the 25% sentiment

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Dependent	**Market value**	**Market value**	**Market value**	**Market value**	**Market value**	**Market value**	**MtB**	**MtB**	**MtB**	**MtB**	**MtB**	**MtB**
$Digitalization_{30,t-1}$	0.446***	0.318***					0.443**	0.232				
	(0.113)	(0.090)					(0.196)	(0.159)				
$Digitalization_{45,t-1}$			0.381***	0.265***					0.329**	0.169		
			(0.096)	(0.091)					(0.128)	(0.110)		
$Digitalization_{60,t-1}$					0.345***	0.222***					0.409***	0.216*
					(0.092)	(0.069)					(0.144)	(0.115)
Total $Assets_{t-1}$	0.522**	0.277*	0.502**	0.268	0.492**	0.257	0.749***	0.571**	0.730***	0.565**	0.717***	0.552**
	(0.227)	(0.164)	(0.226)	(0.163)	(0.228)	(0.166)	(0.232)	(0.248)	(0.232)	(0.248)	(0.232)	(0.249)
ROA_{t-1}	0.024**	0.002	0.024**	0.002	0.023**	0.002	0.046***	0.015*	0.046***	0.015*	0.046***	0.015*
	(0.012)	(0.009)	(0.012)	(0.008)	(0.011)	(0.008)	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)
$Investment_{t-1}$	0.372*	0.543***	0.379*	0.551***	0.391*	0.557***	−0.513***	−0.270	−0.508***	−0.264	−0.490***	−0.258
	(0.205)	(0.130)	(0.205)	(0.130)	(0.207)	(0.131)	(0.176)	(0.181)	(0.176)	(0.181)	(0.176)	(0.182)
Solvency $Ratio_{t-1}$	0.030***	0.032***	0.029***	0.032***	0.030***	0.032***	−0.025***	−0.014*	−0.025***	−0.014*	−0.024**	−0.014*
	(0.007)	(0.005)	(0.007)	(0.005)	(0.007)	(0.005)	(0.009)	(0.009)	(0.010)	(0.009)	(0.009)	(0.009)
Current $Ratio_{t-1}$	−0.008**	−0.008***	−0.008**	−0.008***	−0.008**	−0.008***	0.010**	0.008*	0.010**	0.008*	0.010**	0.008*
	(0.003)	(0.002)	(0.003)	(0.002)	(0.003)	(0.002)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
Company FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes

TABLE 7 (Continued)

Dependent	(1) Market value	(2) Market value	(3) Market value	(4) Market value	(5) Market value	(6) Market value	(7) MtB	(8) MtB	(9) MtB	(10) MtB	(11) MtB	(12) MtB
Observations	572	572	572	572	572	572	559	559	559	559	559	559
R^2	0.97	0.98	0.97	0.98	0.97	0.98	0.69	0.79	0.69	0.79	0.70	0.79
F test	38.98	38.80	39.29	38.12	39.16	39.17	8.26	4.37	8.50	4.68	8.57	4.83

Note: The results of the panel regressions that examine the relation between digitalization and *market value* (columns 1–6) and *market-to-book value* (columns 7–12), respectively, in sentiment subsamples. Only those company-year observations that exhibit a sentiment above the 25%-percentile have been selected. Columns (1) and (2) as well as columns (7) and (8) show the estimation results using a 30 topic distribution to construct the digitalization regressor. Columns (3) and (4) as well as columns (9) and (10) report the estimation results including a digitalization regressor based on a 45 topic distribution. Columns (5) and (6) as well as columns (11) and (12) report the results based on a 60 topic distribution LDA. The panel has one observation for each company-year combination, and spans the time period 2007–2016. We include company fixed effects in all specifications. Even columns additionally include year fixed effects. Standard errors are reported in parentheses. In all specifications standard errors are robust to heteroskedasticity. Furthermore, the model fit (R^2) an test statistics for the joint significance of regressors (*F* test) are reported at the bottom of the table. *, **, and *** denote statistical significance at the 10%, 5%, and 1% levels, respectively.

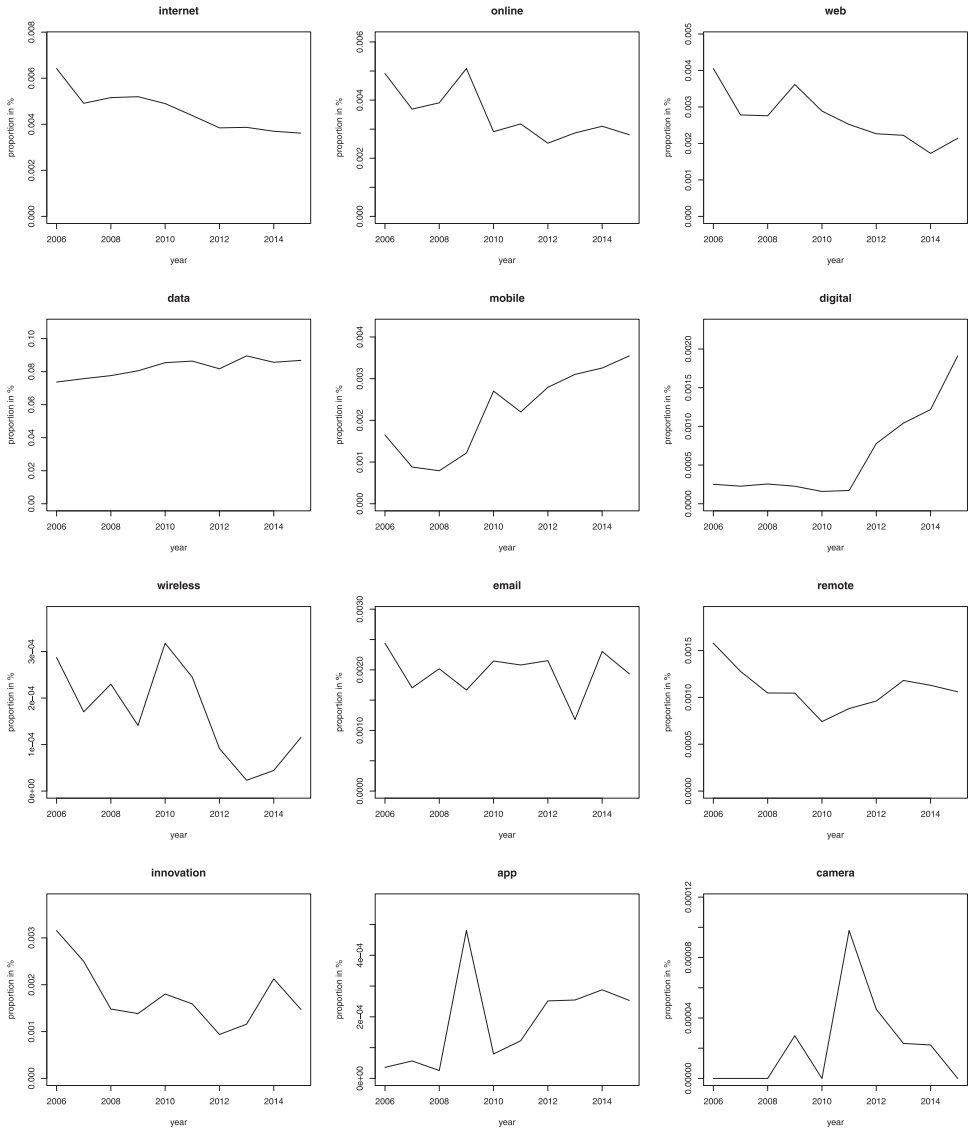

FIGURE 7 Relative frequency of words related to digitalization. This figure presents the relative frequency of words related to digitalization in the annual reports in our sample between 2006 and 2015. Note that our measure of digitalization is derived from the distribution of topics within the annual reports (see Section 2.5)

reports with the longest 2% of the reports being excluded.[31] Furthermore, we construct a measure based on the five most prominent topics (out of 45) in Bohnert et al. (2019) representing nearly 90% of the paper to account for potential confounding effects by the less pronounced topics. Finally, we consider a fourth-root transformation of the original measure.

The choice of Bohnert et al. (2019) as our main reference paper is motivated by its particular focus on digitalization in the insurance sector. However, the paper contains an empirical

[31] In Bellstam et al. (2020) all analyst reports with less than 100 words are dropped as well. However, as all annual reports in our sample consist of more than 100 words (after preprocessing), we do not exclude any short reports.

TABLE 8 The relation between digitalization and firm market value (alternative measure construction and reference documents)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
Dependent	**MV**	**MV**	**MV**	**MV**	**MV**	**MV**	**MV**	**MV**	**MV**	**MV**	**MV**	**MV**	**MV**	**MV**
$Digitalization_{BNT98,t-1}$	0.352***	0.272***												
	(0.105)	(0.095)												
$Digitalization_{BNT5,t-1}$			0.404***	0.255***										
			(0.101)	(0.096)										
$Digitalization_{B4R,t-1}$					0.365***	0.228***								
					(0.083)	(0.074)								
$Digitalization_{BNTtxt,t-1}$							0.436***	0.271***						
							(0.107)	(0.097)						
$Digitalization_{CPLO,t-1}$									0.305***	0.207***				
									(0.080)	(0.072)				
$Digitalization_{NCLT,t-1}$											0.357***	0.235***		
											(0.090)	(0.075)		
$Digitalization_{McK,t-1}$													0.310***	0.213***
													(0.085)	(0.074)
Total $Assets_{t-1}$	0.526**	0.237	0.506**	0.209	0.502**	0.202	0.510**	0.209	0.510**	0.207	0.510**	0.206	0.512**	0.208
	(0.230)	(0.172)	(0.231)	(0.171)	(0.230)	(0.171)	(0.231)	(0.171)	(0.231)	(0.171)	(0.230)	(0.171)	(0.230)	(0.171)
ROA_{t-1}	0.030***	0.009	0.029***	0.007	0.028***	0.007	0.029***	0.008	0.029***	0.007	0.028***	0.007	0.028***	0.007
	(0.012)	(0.009)	(0.011)	(0.008)	(0.010)	(0.008)	(0.011)	(0.008)	(0.010)	(0.008)	(0.010)	(0.008)	(0.010)	(0.008)
$Investment_{t-1}$	0.413**	0.577***	0.429**	0.590***	0.438**	0.592***	0.427**	0.588***	0.426**	0.587***	0.427**	0.587***	0.424**	0.586***
	(0.208)	(0.135)	(0.208)	(0.134)	(0.208)	(0.134)	(0.208)	(0.133)	(0.208)	(0.133)	(0.208)	(0.133)	(0.208)	(0.133)

(Continues)

TABLE 8 (Continued)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
Dependent	**MV**	**MV**	**MV**	**MV**	**MV**	**MV**	**MV**	**MV**	**MV**	**MV**	**MV**	**MV**	**MV**	**MV**
Solvency Ratio$_{t-1}$	0.036***	0.036***	0.037***	0.035***	0.037***	0.035***	0.037***	0.035***	0.037***	0.035***	0.037***	0.035***	0.037***	0.035***
	(0.007)	(0.005)	(0.008)	(0.005)	(0.007)	(0.005)	(0.008)	(0.005)	(0.008)	(0.005)	(0.008)	(0.005)	(0.008)	(0.005)
Current Ratio$_{t-1}$	−0.010***	−0.009***	−0.009***	−0.009***	−0.009***	−0.009***	−0.009***	−0.009***	−0.009***	−0.009***	−0.009***	−0.009***	−0.009***	−0.009***
	(0.003)	(0.002)	(0.003)	(0.002)	(0.003)	(0.002)	(0.003)	(0.002)	(0.003)	(0.002)	(0.003)	(0.002)	(0.003)	(0.002)
Company FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes
Observations	732	732	748	748	748	748	748	748	748	748	748	748	748	748
R^2	0.97	0.98	0.97	0.98	0.97	0.98	0.97	0.98	0.97	0.98	0.97	0.98	0.97	0.98
F test	53.44	44.99	52.46	39.58	54.79	40.15	52.33	39.62	52.56	40.19	53.49	40.58	52.46	40.34

Note: The results of the panel regressions that examine the relation between digitalization based on a 45 topic distribution and firm value proxied by the log of *market value* are presented. For robustness, we consider digitalization measures that have been constructed in an alternative way or are based on a different reference document on digitalization: A measure based on only 98% of the annual reports with the longest 2% of the reports being excluded (columns 1 and 2), a measure based on the five most prominent topics (out of 45) in Bohnert et al. (2019) (columns 3 and 4), a fourth-root transformation of the original measure (columns 5 and 6), and measures based on the reference documents Bohnert et al. (2019) without the empirical section (columns 7 and 8), Cappiello (2020) (columns 9 and 10), Nicoletti (2016) (columns 11 and 12), and McKinsey (2017) (columns 13 and 14). The panel has one observation for each company-year combination, and spans the time period 2007-2016. We include company fixed effects in all specifications. Even columns additionally include year fixed effects. Standard errors are reported in parentheses. In all specifications standard errors are robust to heteroskedasticity. Furthermore, the model fit (R^2) and test statistics for the joint significance of regressors (F test) are reported at the bottom of the table. *, **, and *** denote statistical significance at the 10%, 5%, and 1% levels, respectively.

TABLE 9 The relation between digitalization and firm market-to-book value (alternative measure construction and reference documents)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
Dependent	**MtB**	**MtB**	**MtB**	**MtB**	**MtB**	**MtB**	**MtB**	**MtB**	**MtB**	**MtB**	**MtB**	**MtB**	**MtB**	**MtB**
$Digitalization_{BNT98,t-1}$	0.487***	0.254*												
	(0.173)	(0.135)												
$Digitalization_{BNT5,t-1}$			0.422***	0.206*										
			(0.146)	(0.121)										
$Digitalization_{B4R,t-1}$					0.428***	0.197*								
					(0.133)	(0.108)								
$Digitalization_{BNTtxt,t-1}$							0.520***	0.256*						
							(0.178)	(0.142)						
$Digitalization_{CPLO,t-1}$									0.321***	0.144				
									(0.117)	(0.094)				
$Digitalization_{NCLT,t-1}$											0.426***	0.199*		
											(0.143)	(0.110)		
$Digitalization_{McK,t-1}$													0.308**	0.119
													(0.120)	(0.096)
Total $Assets_{t-1}$	0.921***	0.744**	0.902***	0.728***	0.896***	0.721***	0.905***	0.726***	0.906***	0.728***	0.905***	0.725***	0.908***	0.730***
	(0.261)	(0.290)	(0.255)	(0.278)	(0.254)	(0.277)	(0.255)	(0.278)	(0.255)	(0.278)	(0.255)	(0.278)	(0.256)	(0.279)
ROA_{t-1}	0.046***	0.015**	0.045***	0.014*	0.044***	0.014*	0.045***	0.014**	0.044***	0.014*	0.044***	0.014*	0.044***	0.014*
	(0.009)	(0.008)	(0.008)	(0.007)	(0.008)	(0.007)	(0.008)	(0.007)	(0.008)	(0.007)	(0.008)	(0.007)	(0.008)	(0.007)
$Investment_{t-1}$	−0.618***	−0.363*	−0.600***	−0.348*	−0.587***	−0.346*	−0.599***	−0.350*	−0.603***	−0.350*	−0.599***	−0.350*	−0.605***	−0.350*
	(0.205)	(0.205)	(0.202)	(0.201)	(0.200)	(0.201)	(0.201)	(0.202)	(0.202)	(0.202)	(0.202)	(0.202)	(0.202)	(0.202)
Solvency $Ratio_{t-1}$	−0.029**	−0.017*	−0.028**	−0.017*	−0.027**	−0.017*	−0.028**	−0.017*	−0.028**	−0.017*	−0.027**	−0.017*	−0.028**	−0.017*
	(0.012)	(0.010)	(0.012)	(0.010)	(0.012)	(0.010)	(0.012)	(0.010)	(0.012)	(0.010)	(0.012)	(0.010)	(0.012)	(0.010)
Current $Ratio_{t-1}$	0.012**	0.011**	0.012**	0.011**	0.012**	0.011**	0.012**	0.011**	0.012**	0.011**	0.012**	0.011**	0.012**	0.011**
	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)

(Continues)

TABLE 9 (Continued)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
Dependent	**MtB**	**MtB**	**MtB**	**MtB**	**MtB**	**MtB**	**MtB**	**MtB**	**MtB**	**MtB**	**MtB**	**MtB**	**MtB**	**MtB**
Company FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes
Observations	719	719	735	735	735	735	735	735	735	735	735	735	735	735
R^2	0.61	0.71	0.60	0.71	0.60	0.71	0.60	0.71	0.60	0.71	0.60	0.71	0.60	0.71
F test	7.81	4.19	8.24	3.96	8.33	3.89	8.21	3.94	8.12	3.90	8.20	3.97	8.08	3.88

Note: The results of the panel regressions that examine the relation between digitalization based on a 45 topic distribution and firm value proxied by *market-to-book value* (MtB) are presented. For robustness, we consider digitalization measures that have been constructed in an alternative way or are based on a different reference document on digitalization: A measure based on only 98% of the annual reports with the longest 2% of the reports being excluded (columns 1 and 2), a measure based on the five most prominent topics (out of 45) in Bohnert et al. (2019) (columns 3 and 4), a fourth-root transformation of the original measure (columns 5 and 6), and measures based on the reference documents Bohnert et al. (2019) without the empirical section (columns 7 and 8), Cappiello (2020) (columns 9 and 10), Nicoletti (2016) (columns 11 and 12), and McKinsey (2017) (columns 13 and 14). The panel has one observation for each company-year combination, and spans the time period 2007–2016. We include company fixed effects in all specifications. Even columns additionally include year fixed effects. Standard errors are reported in parentheses. In all specifications standard errors are robust to heteroskedasticity. Furthermore, the model fit (R^2) and test statistics for the joint significance of regressors (*F* test) are reported at the bottom of the table. *, **, and *** denote statistical significance at the 10, 5, and 1 percent levels, respectively.

section which might complicate a proper analysis of the topic distribution. As a robustness check, we therefore compare the topic distributions in our annual reports to those in Bohnert et al. (2019) without the empirical section. Furthermore, we choose another insurance-related paper (Cappiello, 2020) as well as a textbook (Nicoletti, 2016) and a consultancy report (McKinsey, 2017) on digitalization in the insurance sector to justify that our estimation results do not depend on the particular choice of the reference document on digitalization.[32]

The results are presented in Table 8 for the market value as the dependent variable and in Table 9 for the market-to-book ratio as the dependent variable. The estimated coefficients (rows 1–7) are in line with those resulting from our baseline regression. The relation between digitalization and firm value remains positive and statistically significant. Furthermore, the results are in most cases robust to the inclusion of time fixed effects (even columns).

5 | CONCLUSION

Although there is ample evidence on the fundamental impact digitalization will have on the business model of insurance companies, it is not yet clear whether it represents "only" a megatrend insurance companies need to follow or powerful means to increase their profitability. Unfortunately, the answer to that question is not straightforward since digitalization is difficult both to quantify and to differentiate from general innovation. In this study, we present a novel approach to tackle these two problems using unsupervised machine learning algorithms.

In detail, we exploit the prevalence of different topics in insurers' annual reports to construct a text-based measure of digitalization. By employing LDA, we determine the distribution over topics in each report and compare it to a reference document on digitalization based on the KL divergence. We then use this measure of similarity between the reports and the reference document to proxy for the extent of digitalization in an insurance company.

The digitalization proxy is then used as main explanatory variable to investigate the relation between digitalization and firm market valuation in a multivariate OLS model including firm and time fixed effects. Our results show that digitalization efforts are positively rewarded by market participants. An increase in digitalization is related to an increase in the market value of the insurers in our sample. Put differently, market participants expect efforts in digitalization to result in higher future profits. The estimation results are robust to different specifications of the LDA model employing other topic distributions or sentiment analysis. Furthermore, they neither depend on the particular calculation of the digitalization measure nor the choice of the digitalization reference document.

Of course, there are also limitations to our approach. A first limitation of this study concerns the assumptions of the LDA model, most notably the bag of words assumption. There exist several extensions of the LDA model that address some of its shortcomings. However, the question which topic model to use when being confronted with a new set of texts and a new task is still an open direction for topic modeling (Blei, 2012). Additionally, while we are doing our best in constructing a measure to discriminate between digitalization and innovation, our approach has limitations in disentangling both concepts because they are closely related to each other. A third limitation concerns the question of causation between digitalization and firm market valuation. The estimated effect of our digitalization measure might be subject to

[32]We also compared our results to other insurance-related papers, for example, Eling and Lehmann (2018). The results are in line with our findings.

endogeneity issues if larger or more profitable insurers implement digital systems quicker and more extensively than their competitors simply because of greater capacities. Whereas descriptive analyses of our sample do not provide clear evidence in favor of these channels, we refrain from interpreting our estimation results causally and instead consider them as profound confirmation of a positive relation between digitalization and firm market valuation. Establishing a causal link is an interesting avenue for future research.

Digitalization is a complex concept that cannot easily be captured empirically. However, with the rise of machine learning algorithms in the field of textual analysis and massive gains in computational power, the researcher is provided with a new set of powerful tools to analyze large amounts of textual data and retrieve the underlying thematic structure. In this sense, our approach can be considered a first step toward a new empirical analysis of the impact of digitalization not just in the insurance sector, but also in any other sector that will be disrupted by digitalization. Moreover, it could also be applied in various other settings where it is hard to retrieve empirical data due to the indefinite subject of research, for example, corporate sustainability/social responsibility and its impact on corporate performance.

ACKNOWLEDGMENTS

We thank Joan Schmit (the editor), Antonella Cappiello, and two anonymous referees for very useful comments that led to a significant improvement of this paper. Simon Fritzsch gratefully acknowledges financial support through a PhD scholarship from LBBW Asset Management Investmentgesellschaft mbH, Stuttgart. The source code of the *R* programs used in this paper is available from https://github.com/GNFWeiss/digitalization_insurance.

REFERENCES

Allen, F. (1993). Strategic management and financial markets. *Strategic Management Journal*, *14*, 11–22.

Antweiler, W., & Frank, M. Z. (2004). Is all that talk just noise? The information content of internet stock message boards. *Journal of Finance*, *59*, 1259–1294.

Arun, R., Suresh, V., Veni Madhavan, C. E., & Narasimha Murthy, M. N. (2010). On finding the natural number of topics with latent Dirichlet allocation: Some observations, In M. J. Zaki, J. X. Yu, B. Ravindran, V. Pudi, *Advances in knowledge discovery* (Vol. 6118, pp. 391–402). https://link.springer.com/chapter/10.1007/978-3-642-13657-3_43

Bao, Y., & Datta, A. (2014). Simultaneously discovering and quantifying risk types from textual risk disclosures. *Management Science*, *60*, 1371–1391.

Barkur, G., Varambally, K., & Rodrigues, L. L. (2007). Insurance sector dynamics: Towards transformation into learning organization. *The Learning Organization*, *14*, 510–523.

Bellstam, G., Bhagat, S., & Cookson, J. A. (2020). A text-based analysis of corporate innovation. *Management Science*, 1–28. https://doi.org/10.1287/mnsc.2020.3682

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, *55*, 77–84.

Blei, D. M., Griffiths, T. L., & Jordan, M. I. (2010). The nested Chinese restaurant process and bayesian nonparametric inference of topic hierarchies. *Journal of the ACM*, *57*, 1–30.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993–1022.

Bodnaruk, A., Loughran, T., & McDonald, B. (2015). Using 10-K text to gauge financial constraints. *Journal of Financial and Quantitative Analysis*, *50*, 623–646.

Bohnert, A., Fritzsche, A., & Gregor, S. (2019). Digital agendas in the insurance industry: The importance of comprehensive approaches. *Geneva Papers on Risk and Insurance—Issues and Practice*, *44*, 1–19.

Cao, J., Xia, T., Li, J., Zhang, Y., & Tang, S. (2009). A density-based method for adaptive LDA model selection. *Neurocomputing*, *72*, 1775–1781.

Cappiello, A. (2020). The digital (R)evolution of insurance business models. *American Journal of Economics and Business Administration*, *12*, 1–13.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, *41*, 391–407.

Deloitte. (2016). *Insurers on the brink: Disrupt or be disrupted*. White paper.

Desyllas, P., & Sako, M. (2013). Profiting from business model innovation: Evidence from Pay-As-You-Drive auto insurance. *Research Policy*, *42*, 101–116.

Deveaud, R., Sanjuan, E., & Bellot, P. (2014). Accurate and effective latent concept modeling for ad hoc information retrieval. *Document numérique*, *17*, 61–84.

Doherty, N. A., & Richter, A. (2002). Moral hazard, basis risk, and gap insurance. *Journal of Risk and Insurance*, *69*, 9–24.

Eling, M., & Lehmann, M. (2018). The impact of digitalization on the insurance value chain and the insurability of risks. *Geneva Papers on Risk and Insurance—Issues and Practice*, *43*, 359–396.

Feinerer, I., & Hornik, K. (2020). tm: Text Mining Package.

Frazier, K. B., Ingram, R. W., & Tennyson, B. M. (1984). A methodology for the analysis of narrative accounting disclosures. *Journal of Accounting Research*, *22*, 318–331.

Ganglmair, B., & Wardlaw, M. (2017). Complexity, standardization, and the design of loan agreements. *SSRN Electronic Journal*.

Gatzert, N., & Heidinger, D. (2020). An empirical analysis of market reactions to the first solvency and financial condition reports in the European insurance industry. *Journal Risk and Insurance*, *87*, 407–436.

Gelfand, A. E., and, Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, *85*, 398–409.

Goldsmith-Pinkham, P., Hirtle, B., & Lucca, D. O. (2016). *Parsing the content of bank supervision*. Federal Reserve Bank of New York Staff Reports, no. 770.

Grace, M. (2019). *Risk identification: What is in the 10-K?*, 2019 ARIA Annual Meeting Paper.

Gries, S. T., & Newman, J. (2013). Creating and using corpora. In R. J. Podesva, & D. Sharma (Eds.), *Research methods in linguistics* (pp. 257–287). Cambridge University Press.

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(Suppl. 1), 5228–5235.

Griffiths, T. L., Steyvers, M., Blei, D. M., & Tenenbaum, J. B. (2005). Integrating topics and syntax. Advances in Neural Information Processing Systems 17 - Proceedings of the 2004 Conference, NIPS 2004.

Grosen, A., & Jørgensen, P. L. (2002). Life insurance liabilities at market value: An analysis of insolvency risk, bonus policy, and regulatory intervention rules in a barrier option framework. *Journal of Risk and Insurance*, *69*, 63–91.

Hanelt, A., Firk, S., Hildebrandt, B., & Kolbe, L. M. (2020). Digital M&A, digital innovation, and firm performance: an empirical investigation. *European Journal of Information Systems*, *24*, 1–24.

Hanley, K. W., & Hoberg, G. (2010). The information content of IPO prospectuses. *Review of Financial Studies*, *23*, 2821–2864.

Hansen, S., McMahon, M., & Prat, A. (2018). Transparency and deliberation within the FOMC: A computational linguistics approach. *Quarterly Journal of Economics*, *133*, 801–870.

Henry, E., & Leone, A. J. (2016). Measuring qualitative information in capital markets research: Comparison of alternative methodologies to measure disclosure tone. *Accounting Review*, *91*, 153–178.

Hoberg, G., & Lewis, C. (2017). Do fraudulent firms produce abnormal disclosure? *Journal of Corporate Finance*, *43*, 58–85.

Hoberg, G., & Maksimovic, V. (2015). Redefining financial constraints: A text-based analysis. *Review of Financial Studies*, *28*, 1312–1352.

Hoberg, G., Phillips, G., & Prabhala, N. (2014). Product market threats, payouts, and financial flexibility. *Journal of Finance*, *69*, 293–324.

Hofmann, T. (1999). Probabilistic latent semantic indexing. In F. Gey (Ed.), *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 50–57). ACM.

Huang, A. H., Lehavy, R., Zang, A. Y., & Zheng, R. (2018). Analyst information discovery and interpretation roles: A topic modeling approach. *Management Science*, *64*, 2833–2855.

Israelsen, R. D. (2014). Tell it like it is: Disclosed risks and factor portfolios. *SSRN Electronic Journal*.

Jegadeesh, N., & Wu, Di. (2013). Word power: A new approach for content analysis. *Journal of Financial Economics*, *110*, 712–729.

Jegadeesh, N., & Wu, D. (2017). Deciphering fedspeak: The information content of FOMC meetings. *SSRN Electronic Journal*.

Ke, Z. T., Kelly, B. T., & Xiu, D. (2019). *Predicting returns with text data*. National Bureau of Economic Research.

Kullback, S., & Leibler, R. A. (1951). On Information and sufficiency. *The Annals of Mathematical Statistics*, *22*, 79–86.

Lee, J. -P., & Yu, M. -T. (2002). Pricing default-risky CAT bonds with moral hazard and basis risk. *Journal of Risk and Insurance*, *69*, 25–44.

Li, F. (2008). Annual report readability, current earnings, and earnings persistence. *Journal of Accounting Research*, *45*, 221–247.

Lopez-Lira, A. (2019). *Risk factors that matter: Textual analysis of risk disclosures for the cross-section of returns*. Jacobs Levy Equity Management Center for Quantitative Financial Research Paper.

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

Loughran, T., & McDonald, B. (2016). Textual analysis in accounting and finance: A survey. *Journal of Accounting Research*, *54*(4), 1187–1230.

Lowry, M., Michaely, R., Volkova, E. (2020). Information Revealed through the Regulatory Process: Interactions between the SEC and Companies ahead of Their IPO. *The Review of Financial Studies*, *33* (12), 5510–5554.

Lubatkin, M., & Shrieves, R. E. (1986). Towards reconciliation of market performance measures to strategic management research. *Academy of Management Review*, *11*, 497–512.

Manning, C. D., Raghavan, P., & H. Schütze (2009). *Introduction to information retrieval*. Cambridge Univ. Press, reprinted. ed.

McKinsey. (2017). Digital disruption in insurance: Cutting through the noise. White paper.

Meier, A. and Stormer, H. (2012): *eBusiness & eCommerce: Management der digitalen Wertschöpfungskette*. (3rd ed.). Springer.

Nicoletti, B. (2016): *Digital insurance: Business innovation in the post-crisis era, Palgrave Studies in Financial Services Technology*. Palgrave MacMillan.

Porter, M. F. (1980). An algorithm for suffix stripping. *Program*, *14*, 130–137.

Rayport, J. F., & Sviokla, J. F. (1995). Exploiting the virtual value chain. *Harvard Business Review*, *73*, 75–85.

Roberts, M. E., Stewart, B. M., & Airoldi, E. M. (2016). A model of text for experimentation in the social sciences. *Journal of the American Statistical Association*, *111*, 988–1003.

Schmidt, C. (2018): *Insurance in the digital age: A view on key implications for the economy and society*, The Geneva Association—International Association for the Study of Insurance Economics.

Scott, S. V., van Reenen, J., & Zachariadis, M. (2017). The long-term effect of digital innovation on bank performance: An empirical study of SWIFT adoption in financial services. *Research Policy*, *46*, 984–1004.

Seitz, M. (2017). Online Insurance Management among German Farmers. The Proceedings of the 17th International Joint Conference Central and Eastern Europe in the Changing Business Environment, pp. 212–220.

Subramanyam, K. R. (2014). *Financial statement analysis*. (11th ed.). McGraw Hill Education.

Teh, Y. W., Jordan, M. I., Beal, M. J., & Blei, D. M. (2006). Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, *101*, 1566–1581.

Tetlock, P. C. (2007). Giving content to investor sentiment: The role of media in the stock market. *Journal of Finance*, *62*, 1139–1168.

Tidd, J., & Bessant, J. R. (2018): *Managing innovation: Integrating technological, market and organizational change* (6th ed.). Wiley.

van Rossum, A., de Castries, H., & Mendelsohn, R. (2002). The debate on the insurance value chain. *Geneva Papers on Risk and Insurance—Issues and Practice*, *27*, 89–101.

Wainwright, M. J., & Jordan, M. I. (2008). Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, *1*, 1–305.

Wallach, H. M. (2006). Topic modeling, In W. Cohen (Ed.), *Proceedings of the 23rd international conference on Machine learning* (pp. 977–984). ACM.

Wang, X., & Grimson, E. (2007). Spatial latent Dirichlet allocation. *Advances in Neural Information Processing Systems*, *20*, 1577–1584.

Weiss Hanley, K., & Hoberg, G. (2019). Dynamic interpretation of emerging risks in the financial sector. *Review of Financial Studies*, *32*, 4543–4603.

Yekini, L. S., Wisniewski, T. P., & Millo, Y. (2016). Market reaction to the positiveness of annual report narratives. *British Accounting Review*, *48*, 415–430.

How to cite this article: Fritzsch, S., Scharner, P., & Weiß, G. (2021). Estimating the relation between digitalization and the market value of insurers. *J Risk Insur*. 88, 529–567. https://doi.org/10.1111/jori.12346

APPENDIX A: VARIABLE DEFINITIONS AND DATA SOURCES

The appendix presents definitions for the dependent and independent variables that are used in the empirical study and that have not been calculated using LDA. Capital market data are retrieved from Thomson Reuters Datastream and accounting data are retrieved from Orbis Insurance Focus. All accounting data are collected in US dollar.

Variable	Description	Source
Capital market data		
Market value	Natural logarithm of the share price multiplied by the number of ordinary shares in issue	Thomson Reuters Datastream
Market-to-book value	Market value of common equity divided by the balance sheet value of common equity in the company	Thomson Reuters Datastream
Accounting data		
Total assets	Natural logarithm of an insurer's total assets at fiscal year end	Orbis Insurance Focus
ROA	Return on Assets defined as net income over total assets (in %)	Orbis Insurance Focus
Total investment	Natural logarithm of an insurer's total amount of money invested into capital	Orbis Insurance Focus
Solvency ratio	Net assets divided by net premiums written (in %)	Orbis Insurance Focus
Current ratio	Current assets divided by current liabilities (in %)	Orbis Insurance Focus
Foreign assets ratio	Foreign assets divided by total assets (in %)	Thomson Reuters Datastream