Measuring the Effect of Digitalization Efforts on Bank Performance

Johannes Kriebel
University of Muenster*

Joern Debener
University of Muenster†

January 15, 2020

Abstract

There is an ongoing debate on whether and under which circumstances digitalization can increase company performance (IT profitability paradox). Digitalization is of particular importance to banks. The debate to date suffers from a lack of structured reported information on digitalization efforts. As our main contribution, we suggest a new measure of these efforts in banks by applying text mining methods to exploit unstructured information from annual reports. We confirm this approach by showing that it predicts a substitution of personnel expenses with non-personnel expenses within banks and further predicts banks' IT patenting activity. Assessing the sentiment of reporting on digitalization, banks that report in a positive context, in fact, have a higher subsequent performance. This hints towards explanations for the IT profitability paradox linked to organizational capabilities.

Keywords: Digital banking, IT profitability paradox, text mining

JEL Classification: G21

1 Introduction

The ongoing digital transformation fundamentally changes the way businesses operate and the way goods and services are produced and marketed. In the following, we use the term 'digitalization' to describe the increased use of information or digital technologies in organizations. The digitalization is highly relevant for the financial services sector including banks, which are considered particularly technology intensive,

*Corresponding author, Universitätsstraße 14-16, 48143 Münster, Germany, johannes.kriebel@wiwi.uni-muenster.de, phone +49-251-83-22692, fax +49-251-83-22882

†Universitätsstraße 14-16, 48143 Münster, Germany, joern.debener@wiwi.uni-muenster.de, phone +49-251-83-22879, fax +49-251-83-22882

as Beccalli (2007) points out. Maintaining existing digital infrastructures and implementing new digital business models require investments on a very large scale. This leads to the question how digitalization efforts affect banks' performance and what digital technologies banks should invest in.

The literature on this topic is rather sparse to date. This is arguably related to digitalization efforts generally being difficult to measure, as they are not disclosed in a structured and consistent way. Previous studies, therefore, mainly rely on self-reported survey data of fairly small sample sizes (e.g. Hitt and Brynjolfsson, 1996, Aral and Weill, 2007, Beccalli, 2007, and Mithas et al., 2012). The existing work in this field almost entirely uses data from the 1990s, which arguably does not provide much guidance to decision makers in a world where digital progress is proceeding fast. These surveys further use very broad categories for IT investments such as hardware, software, and outsourcing.

The existing literature is particularly indecisive on whether companies can increase profits with investments in IT. Hitt and Brynjolfsson (1996) argue that information technology can be considered to be a commodity that is available to all market participants. It is, therefore, not clear, whether companies can create competitive advantages out of it. Consequently, several important studies find no clear effect of investments in information technology on performance, which is referred to as the IT profitability paradox (Aral and Weill, 2007, Beccalli, 2007, and Ho and Mallick, 2010).

Other studies argue that certain conditions need to be met to make IT investments increase performance. Kohli and Devaraj (2003) and Aral and Weill (2007) argue that when considering the capabilities and actual usage, the link to performance is more clear. Beccalli (2007) regards investments in outsourcing as profitable while investments in software and hardware have no clear effect on the performance. Mithas et al. (2012) find that investments in IT particularly increase performance via increased sales while investments in efficient processes are less important. Shah and Shin (2007), in contrast, find investments in efficient processes to be crucial for increased performance. Stratopoulos and Dehning (2000) find that the effect of IT investments on profits is only temporary, which is in line with diminishing effects of competitive advantages. Wang (2010) does not find a clear instantaneous effect but mentions that IT investments can be profitable on longer horizons.

Our research approach is to solve the problem of measuring digitalization efforts by using text mining methods to extract related information from banks' annual reports. We therefore develop two measures: The first measure is a frequency based assessment of common digitalization keywords in these reports. The intuition behind this measure is that a bank will report digitalization aspects more extensively when it is more involved in implementing them. This is related to Bellstam et al. (2019) and Wang (2010) who show that information on innovation and digitalization efforts, respectively, can be extracted from analyst reports and newspapers. Developing this measure is the key contribution of our study. We validate the measure in two ways using a sample of all US banks that are listed on the New York Stock Exchange. The results indicate that a more frequent reporting of digitalization terms coincides with a substitution of personnel expenses with non-personnel expenses. This is in line with the study of Martín-Oliver and Salas-Fumás (2008) who find that information technology is a substitute for labor in banks. To further validate our approach, we show that banks which report more frequently about digitalization terms also file for more IT patents as measured by data available from the United States Patent and Trademark Office. This is in accordance Kleis et al. (2012) who show that IT investments lead to a higher innovation productivity and thus confirms the idea that the reporting frequency measures digitalization efforts. In line with many previous studies, the effect on the profitability is less clear. In a second step, we evaluate the sentiment of the context in which the digitalization terms are reported. The intuition behind this measure is that a bank will report digitalization aspects more positively when it succeeds in their implementation. When evaluating the context of the terms, a positive reporting of digitalization aspects coincides with a strong positive relation to performance. With regard to this second measure, we find evidence that links the IT profitability paradox to theoretical explanations related to the capabilities to succesfully implement IT (Aral and Weill, 2007, Stratopoulos and Dehning, 2000).

The remainder of the paper is structured as follows: Section 2 provides an overview of the existing literature on the IT productivity and profitability paradox and previous studies using text mining methods in finance and banking. Section 3 describes the applied text mining methods and the resulting digitalization measures. Section 4 introduces the data and our empirical model. Section 5 contains the empirical results and Section 6 presents robustness checks. Section 7 concludes.

2 Literature Review

2.1 Digitalization and Bank Performance

There has been some work on the link between IT and productivity. In his seminal article, Solow (1987) was one of the first to identify a so-called IT productivity paradox. The central observation is that, while the use of information technology in the form of personal computers in the United States increased significantly since 1970, measured aggregate productivity did not increase in the same way (Brynjolfsson, 1993). This unexpected observation resulted in a new strand of literature, aiming at a better understanding of the reasons for this paradox and the link between IT investments and productivity in general.

In a review of the respective literature, Brynjolfsson (1993) discusses explanations for this paradox. A major issue of early studies is the neglection of a learning curve: Since the use of new IT may require experience before the investment can unfold its benefits, the inclusion of lagged variables may be necessary. Also, outputs created by IT like quality and variety of the products may be difficult to measure, especially in service industries such as banking. Furthermore, a firm-level analysis instead of an aggregate view may yield different results. In a theoretical approach, Soh and Markus (1995) develop a process theory synthesis to understand how IT investments create business value. They conclude that IT investments need to be supported by IT management and by the appropriate use of IT through personnel to ultimately lead to a higher productivity and a better financial performance. This theoretical framework and the arguments in Brynjolfsson (1993) have been further investigated by several studies. Lichtenberg (1995) analyzes a short panel of firm-level data based on two company surveys and finds significant output impacts of both, IT investments and the number of IT employees. Bresnahan et al. (2002) and Brynjolfsson et al. (2002) show a positive impact of IT investments in combination with complementary organizational structures on the productivity of firms. Kleis et al. (2012) find that IT investments lead to a higher innovation productivity, which supports Brynjolfsson (1993) in his argument that IT increases outputs in a way that can be difficult to measure. Dedrick et al. (2003) conduct a literature review and conclude that the impact of IT on productivity is highly industry-dependent with no clear evidence for banks. As banks' outputs may be particularly difficult to measure, the authors suggest that banks need to be analyzed separately with respect to this characteristic.

Literature on the impact of IT on the productivity of banks is relatively sparse. Parsons et al. (1993) investigate the effects of IT investments in Canadian banking and find a moderately increasing productivity over the next five years. For a small panel of 12 US banks, Shu and Strassmann (2005) find a positive impact of IT investments on productivity as well. Casolaro and Gobbi (2007) examine the relationship between the IT capital stock and productivity for a panel of Italian banks and provide evidence of a positive correlation. In a study on Spanish banks, Martín-Oliver and Salas-Fumás (2008) find that banks mainly use IT as a substitute for labor. Furthermore, they show that IT investments contribute substantially to banks' output. Koetter and Noth (2013) confirm these results for a panel of German savings banks, they find a positive output contribution of IT investments as well.

While the case is more clear for the link between IT and productivity, Hitt and Brynjolfsson (1996) point out that an increase in productivity does not necessarily lead to an increase in profitability. Productivity analyses are mainly concerned with the question whether more output can be generated with the same input, a higher productivity might not lead to competitive advantages, when a production technology is available to all market participants.

In fact, Carr (2003) argues that IT should be seen as an ordinary input that does not create competitive advantages, as the major technologies are freely available. Beccalli (2007) empirically supports this line of argumentation. In an econometric analysis comprising European banks, Beccalli (2007) does not generally find positive impacts of IT investments on the profitability. Yet, IT services from external providers seem to have a positive significant impact on banks' profits. Stratopoulos and Dehning (2000) find that IT increases performance for a short time, but this advantage diminishes due to competitors copying the IT efforts, which supports Carr (2003). Ho and Mallick (2010) argue that IT investments can sometimes even have a negative impact on the financial performance of banks. Wang (2010) does not find empirical evidence of an improved financial performance by chasing IT trends in the short-term, while there might be positive effects in the long-term. However, a higher reputation and executive compensation goes hand in hand with the implementation of fashionable IT, thereby giving an explanation for how decision makers choose investments in IT. In line with these studies, Aral and Weill (2007) do not find an impact of IT investments alone on the profitability of companies in the US. However, they find a significant impact of IT investments in combination with the organizational capability to use IT on companies' profits, thus backing the theoretical framework of Soh and Markus (1995). In contrast, Mithas et al. (2012) find that mere investments in IT increase the profitability of companies. They identify increased sales as the main profit driver while investments in processes do not seem to increase profitability. Shah and Shin (2007), however, find improved processes to be the missing link between IT investments and increased profitability. Hernando and Nieto (2007) further find the adoption of internet banking as a delivery channel to positively affect the profitability of Spanish banks.

The effect of IT investments on performance remains disputed to date. There is in particular no consensus which technologies increase performance and under which circumstances this is the case. We contribute to this literature by suggesting a new measure for digitalization efforts and validating it using changes in the cost structure of banks and banks' patenting activity. We further find evidence of the IT profitability paradox and, using a sentiment-based measure, provide evidence linking the paradox to organizational capabilities.

2.2 Text Mining Methods in Finance and Banking

Text mining describes the process of retrieving information from unstructured text data. In recent finance and banking literature, this approach increasingly gains popularity.

There are several applications to analyzing text data to forecast stock returns. While Hagenau et al. (2013) use machine learning methods to classify company news, Tetlock et al. (2008) find that the sentiment of stock news can predict fundamental financial performance and stock returns. Jegadeesh and Wu (2013) go in the same direction and quantify the tone of IPO prospectuses and relate it to IPO underpricing.

Other studies focus on topic modeling techniques. Hendry and Madeley (2010) analyze communication statements from the Bank of Canada and determine topics that are significant drivers of interest rate markets. Jegadeesh and Wu (2017) break down Federal Reserve communications into topics and connect different addressed topics with stock and bond market reactions. Bellstam et al. (2019) use Latent Dirichlet Allocation to model topics in analyst reports in order to measure innovation efforts of companies. The innovation measure allows to predict the financial performance of companies. This approach is similar to our approach and supports its applicability.

Considering banking and risk management purposes, Ganglmair and Wardlaw (2017) measure the complexity of loan contracts and show that more complex loan contracts need to be renegotiated more often. Saha et al. (2016) make an effort to improve the loan processing of banks by automatically classifying loan applications in order to detect fraudulent activities. Cecchini et al. (2010) and Hoberg and Lewis (2017) use the Management Discussion and Analysis section of 10-K filings to predict financial events like fraud and bankruptcy.

Previous studies show that text data can provide useful information that goes beyond reported structured data. Furthermore, there appears to be a variety of promising methods appropriate to analyze text data. We make use of these applications to develop new digitalization measures in order to better understand the link between digitalization and bank performance.

3 Text Mining Methods

3.1 Term Frequencies

A major issue in investigating the link between digitalization efforts and bank performance is the unavailability of structured data on these efforts. Previous studies mostly relied on self-reported survey data of often relatively small sample sizes (e.g. Hitt and Brynjolfsson, 1996, Aral and Weill, 2007, Beccalli, 2007, and Mithas et al., 2012). This restricts the widespread analysis of this issue over many firms. It further introduces selection issues. Our research approach addresses this problem and suggests a convenient solution by utilizing text mining methods.

We address two aspects with two different measures: The first measure is based on the frequencies of digitalization terms, so-called keywords. For this purpose, we select a list of particularly relevant keywords from digitalization dictionaries.[1] We count the frequency of these digitalization-related terms in banks'

[1] The most frequently appearing digitalization keywords are displayed in Figure 1. The full list of keywords is available from

annual reports.[2] The measured term frequency (Tf) is then weighted by the inverse document frequency (Idf), which counts in how many documents of all documents in the Corpus (N) the specific term is used. This is a common weighting-procedure that puts extra weight on terms that are often used in a specific document but rarely in other documents. Terms with a high Tf and a low document frequency (Df) are assumed to be distinctive for a document. Using this weighting, we compute the so-called term frequency-inverse document frequency (Tf-Idf) for every term j:

$$Tf\text{-}Idf(j) = Tf(j) \times Idf(j) \;\; with \;\; Idf(j) = log(\frac{N}{Df(j)}) \;\; for \;\; j = 1, ..., J \tag{1}$$

We compute the term frequency-inverse document frequency for every term and every annual report which results in a structured term-document matrix. For every document, we obtain the overall measure TF.IDF.SCORE by the following equation:

$$TF.IDF.SCORE = \sum_{j \in J} Tf\text{-}Idf(j) \tag{2}$$

The rationale behind this measure is that digitalization efforts by banks will be reported in their annual reports. We expect that banks which are more active in digitalization efforts report more frequently on these aspects.

3.2 Keywords in Context Sentiment

The second digitalization measure goes beyond the frequency based approach. We now specifically consider the context in which the digitalization terms, the so-called keywords, are used. For every keyword, we find all corresponding locations in the annual reports. For the contexts, we perform a sentiment analysis comparable to Tetlock et al. (2008). We use ten words before and after the respective digitalization term and use punctuation to reduce the context to words of the same sentence. We use the sentiment lexicon of Loughran and McDonald (2011), which is specifically designed for financial reporting and include valence shifters like negators, adversative conjunctions, amplifiers and de-amplifiers. To measure the sentiment, we count all positive words in the contexts of our keywords and subtract the number of all negative words. Dividing by the count of locations, we calculate an average sentiment for the context of the keywords in every annual report:

the authors.

2 As annual reports, we use 10-K filings as they have been shown to be particularly informative. The 10-K filings are available over https://www.sec.gov/edgar.shtml.

$$KWIC.SENTI = \frac{\sum_{l \in locations}(w^l_{positive} - w^l_{negative})}{\#locations} \tag{3}$$

The rationale behind this second measure is that banks will report more positively when they are successful in their efforts. This could help in identifying banks that are more capable in implementing new digital technologies.

4 Empirical Methods and Data Description

4.1 Empirical Methods

The main aim of our study is to develop a measure for banks' digitalization efforts. This is of particular importance in order to achieve a better understanding of the IT profitability paradox and of the ways how banks manage digitalization successfully. There are, therefore, two central steps: First, we assess whether the frequency-based approach measures digitalization efforts well. Second, we study how this measure relates to bank profitability.

In order to validate the measurement of digitalization efforts, we conduct an analysis with two steps. We first assess whether the frequency of digitalization terms in banks' annual reports explains changes in the structure of expenditures of banks. In addition, we collect US IT patent data to test whether the frequency of digitalization terms is related to the patenting activity of banks.

Starting with the structure of expenditures and following Martín-Oliver and Salas-Fumás (2008), banks which are more active in digitalization should use this to substitute labor with IT systems. We, therefore, hypothesize that more digitalization efforts should coincide with a substitution of personnel expenses with non-personnel non-interest expenses. This is measured by dividing personnel expenses by total non-interest expenses (PERS.EXP.NON.INT.EXP) and other non-interest expenses by total non-interest expenses (OTHER.EXP.NON.INT.EXP). To assess, whether the overall cost to maintain a certain volume of business changes, we calculate the ratio of total non-interest expenditures by total assets (NON.INT.EXP.TA). In terms of our digitalization measures, a more frequent use of digitalization keywords should coincide with higher OTHER.EXP.NON.INT.EXP and lower PERS.EXP.NON.INT.EXP. Furthermore, if digitalization generally reduces operational costs, a more frequent use of digitalization keywords should coincide with a reduced NON.INT.EXP.TA.

With respect to the patenting activity, we follow Kleis et al. (2012) and expect that banks which report more frequently about digitalization are more innovative. This should result in more more IT patent applications. We use new patent applications per employees (NEW.PAT.EMPL) to measure the patenting activity of banks. The model for the first step is presented in Equation 4.

$$Validation\_Variable_{it} = \alpha_i + \gamma_t + \beta_{dig} X_{it}^{dig} + \beta_{contr} X_{it}^{contr} + \varepsilon_{it} \tag{4}$$

In the notation of Equation 4 (and Equation 5), α_i and γ_t are bank and time fixed-effects. β_{dig} and β_{contr} are vectors of regression coefficients. X_{it}^{dig} and X_{it}^{contr} are the vectors of independent variables and control variables for bank i and time t. ε_{it} is an error term.

In the second step, we assess the relation of the digitalization measures to performance measures (Equation 5). We include three measures of performance: Return on average equity (ROAE), return on average assets (ROAA) and cost income ratio (CIR). These performance measures are frequently used in related studies (e.g. Beccalli, 2007, Aral and Weill, 2007, and Hernando and Nieto, 2007).

$$Performance\_Variable_{it} = \alpha_i + \gamma_t + \beta_{dig} X_{it}^{dig} + \beta_{contr} X_{it}^{contr} + \varepsilon_{it} \tag{5}$$

The ROAE and ROAA measure how well a certain business volume and capital is used to generate income. The CIR measures how much income is generated compared to the non-interest expenses as the cost of the organization. This is also to some degree a measure of how efficient an organization works. In relation to the digitalization measures, when digitalization efforts generally increase performance, a higher TF.IDF.SCORE should coincide with a higher ROAA, ROAE, and lower CIR. However, given the results from earlier IT profitability paradox studies, it is unclear, whether there is in fact a positive link. We furthermore hypothesize that successful digitalization efforts have a positive link to performance. The KWIC.SENTI should therefore coincide with a higher ROAA, ROAE, and a lower CIR.

In our models, we control for several other possible influences on the cost structure and the performance. First, we use bank and year fixed-effects as stated above to control for unobserved entity and time characteristics. Similar to Berger and Bouwman (2013), we then control for differences in size by including the logarithm of the total assets (logTA). The ratio of equity over total assets (TETA) controls for the capital structure as well as bank risk (see Berger and Bouwman, 2013). The ratio of liquid assets by total assets (LIQU.ASSETS.TA) controls for liquidity. As in Hernando and Nieto (2007), we use the ratio of non-performing loans by total loans (LOANS.DEP.ST.FUND) as a second control for bank-risk. The control variables are the same for the models in Equation 4 and Equation 5. All tables state standard errors that are clustered by bank and year.

4.2 Data Description

The data we use is obtained from three sources. All financial data is taken from Fitch Connect. The annual reports include 10-K and 10-K405 filings and are downloaded from the SEC EDGAR database for the years 1993 to 2018. We use the annual reports of all US banks listed on the New York Stock Exchange in January 2019. For calculating the digitalization measures, we use 73 keywords that were collected from digitalization dictionaries.[3] In order to analyze the banks' patenting activities, we use patent data from the United States Patent and Trademark Office (USPTO). All variables are winsorized at 0.01 and 0.99.

As a first step, we convert the banks' annual reports into a form that can be easily analyzed. This step is commonly called preprocessing and is a crucial part of text mining analyses (see Uysal and Gunal, 2014). In order to reduce the variety of inflections and manifestations of the digitalization keywords, we transform the annual reports into lower case, remove all punctuation and lemmatize all words.[4] Furthermore, we remove numbers, special characters and redundant whitespaces as to enhance the computability of our digitalization measures. We apply the same treatment to the list of digitalization keywords we look for.

Table 1 lists summary statistics for the dependent, independent, and control variables in three sections. The first section presents the dependent variables containing the three performance variables and the three variables for the cost structure. Considering the performance variables, the banks in the sample have a mean ROAE of about 9.8% and a mean ROAA of about 0.9%. It is further notable that the minimum values are strongly negative. This is due to the years of the financial crisis in which some of the sample banks incurred severe losses.[5] Considering the cost structure variables, the non-interest expenses are on average about equally divided between personnel expenses and non-personnel expenses. Most of the banks have a comparable level of these cost variables. Half of the bank-year observations lie within 49% and 56% of personnel expenses. The non-interest expenses normalized by the total assets make up about 3%. This is also relatively similar for many of the bank-year observations. Half of the sample has a value between 2.5% and 3.6%. The mean value for the number of new patent applications per employees (in thousand) is 0.053%.

The second section of Table 1 displays summary statistics for the digitalization measures. The table lists the unweighted frequency of digitalization keywords in the annual reports (TF), the frequency of digitalization keywords relative to the length of the report, REL.TF, and the TF.IDF.SCORE. The annual reports

[3] The most frequently appearing keywords are reported in Figure 1. The full list of keywords is available from the authors.

[4] Lemmatization is a common procedure in text mining applications, it groups inflected forms of words in order to reduce the variety of word inflections. For the keywords in context sentiment, we remove all punctuation except full stops as this allows us to determine the beginning and ending of sentences.

[5] We include a robustness check for this characteristic in Section 6

Table 1: Summary statistics

The three sections of this table report summary statistics for the dependent variables, the digitalization measures, and the control variables. "N" is the number of non-missing values, "Mean" the mean, "St. Dev." the standard deviation, "Min" the minimum, "25 Pctl.", "Median", and "75 Pctl." the first, second and third quartile, and "Max" the maximum.

Statistic	N	Mean	St. Dev.	Min	25 Pctl.	75 Pctl.	Max
Dependent variables:							
ROAE	770	9.845	8.979	−38.150	6.630	15.240	25.915
ROAA	770	0.948	0.751	−2.706	0.732	1.350	2.661
CIR	776	63.693	12.416	32.427	56.905	69.242	110.757
PERS.EXP.NON.INT.EXP	768	0.511	0.084	0.144	0.488	0.560	0.728
OTHER.EXP.NON.INT.EXP	768	0.489	0.084	0.272	0.440	0.512	0.856
NON.INT.EXP.TA	768	0.032	0.013	0.009	0.025	0.036	0.144
NEW.PAT.EMPL	613	0.053	0.169	0.000	0.000	0.000	1.061
Digitalization measures:							
TF	954	28.471	26.412	0.000	9.000	41.000	159.000
REL.TF	954	0.001	0.0004	0.000	0.0003	0.001	0.004
TF.IDF.SCORE	954	2.499	1.589	0.000	1.403	3.493	11.560
KWIC.SENTI	954	−0.063	0.087	−0.358	−0.119	0.000	0.258
Control variables:							
logTA	776	24.405	1.942	19.819	23.118	25.784	28.463
TETA	776	0.099	0.030	0.017	0.078	0.113	0.296
LIQU.ASSETS.TA	775	9.240	11.593	0.832	2.655	9.020	56.762
NPL.RATIO	760	1.423	1.576	0.006	0.458	1.812	8.338

contain about 28 keywords for each report on average. 75% of the annual reports use digitalization terms at least nine times. This indicates that digitalization is treated quite regularly in the annual reports. Some of the annual reports use digitalization terms more than 150 times. The last row in this section of the table displays summary statistics for the keywords in context sentiment. It is notable that this measure is slightly negative on average. Even the 75% percentile is not yet positive. This indicates that digitalization terms are often used in a negative context. Many of the annual reports might therefore consider digitalization efforts something challenging rather than a competitive advantage.

Considering the control variables, the logarithm of the total assets lies within 19.819 (Provident Financial Services, Inc., about $404 million) and 28.463 (JPMorgan Chase & Co, about $2.298 trillion). The banks fund themselves with 9.9% equity on average. The mean ratio of liquid assets over total assets is at about 9.2% and the non-performing loans amount to an average of about 1.4% of the loan volume in the sample banks.

4.3 Digitalization in Annual Reports

This section presents several figures to provide a visual impression of what keywords are mainly important in the annual reports. It further provides some impression of how the frequency of keywords in annual reports changes over time.

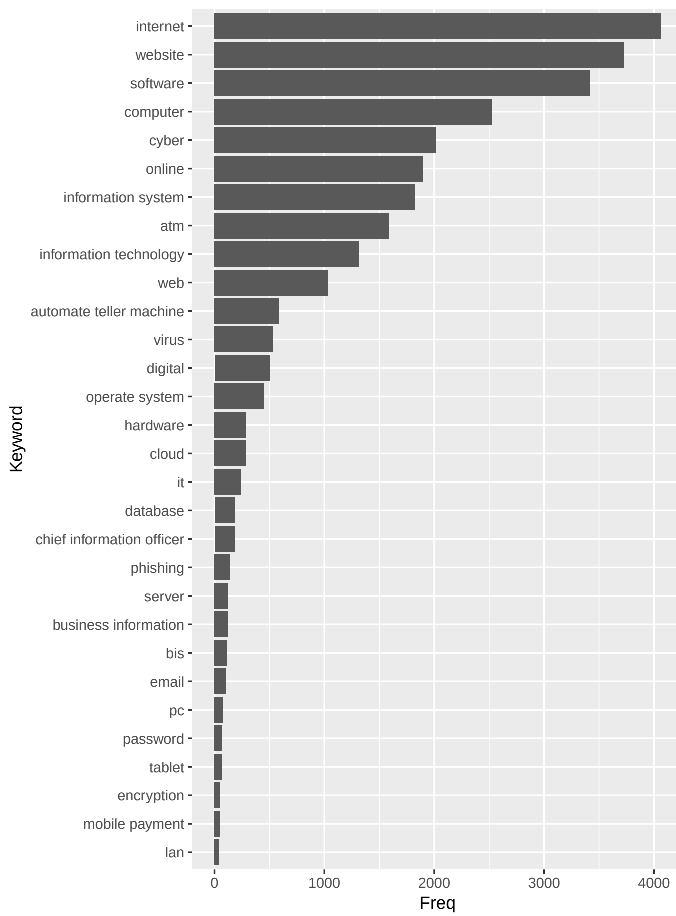
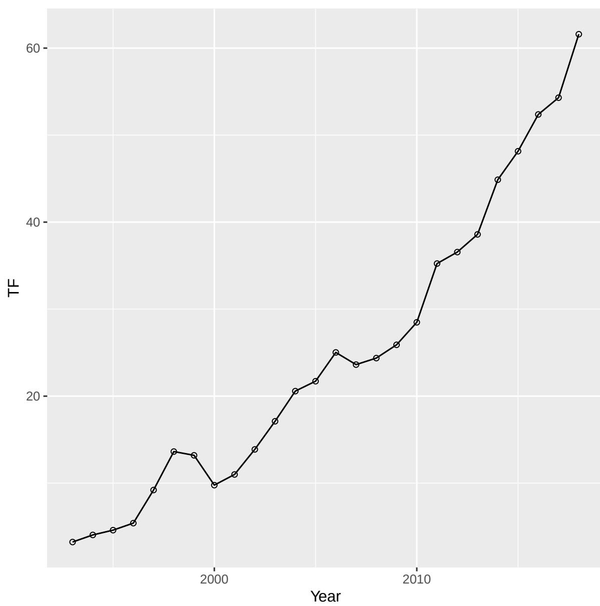


Figure 1: 30 most frequent keywords in the annual reports

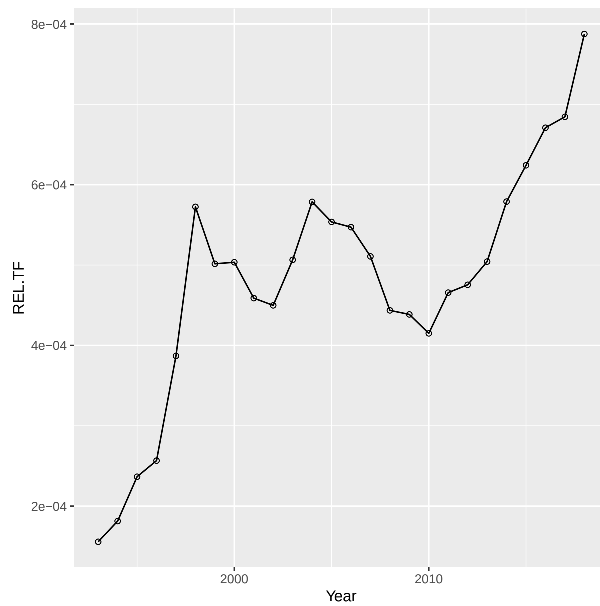
This figure displays the 30 most frequent keywords in the annual reports. The frequency is stated as the unweighted term frequency.

Figure 1 displays the 30 most frequently used digitalization terms. Among the ten most frequent, terms related to the internet and the world wide web such as ”internet” and ”website” appear particularly frequent. Then ”computer”, ”software”, and ”information system” appear quite regularly. Automated teller machines further make up a considerable part of the terms used. Another relatively important topic seems to be cyber security as reflected in words such as ”cyber” or ”virus”.

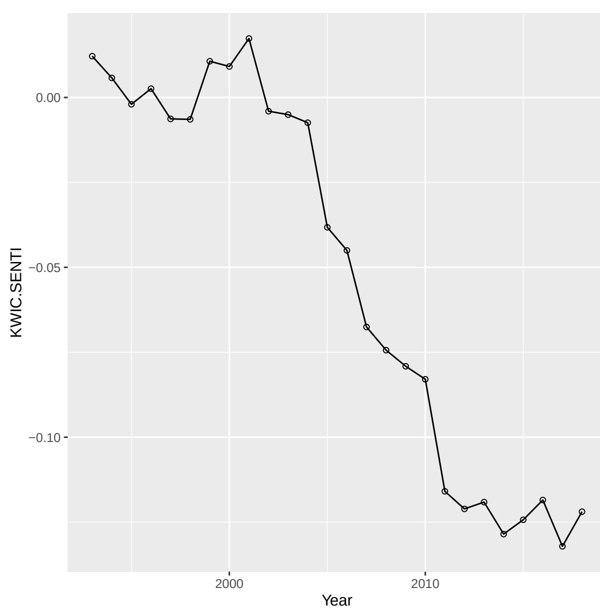
Figure 2a to 2c plot the average values of the digitalization measures by year. These figures should provide some information on how the extent to which the annual reports treat digitalization topics changes over time. Figure 2a shows the average unweighted frequency of keywords used in the annual reports. It is noteworthy that the keywords are used much more frequently over time. While digitalization terms appear on average only less than ten times in annual reports in 1993, they appear more than 60 times on average in 2018. However, as the length of the annual reports generally tends to increase over time, we also plot the



(a) Term frequency by year



(b) Relative term frequency by year



(c) Keywords in context sentiment by year

Figure 2: Digitalization measures over time

This figure displays the average values of the digitalization measures by year.

average relative frequency (REL.TF) of the keywords by year in Figure 2b. Given this measure, the extent of annual reports addressing digitalization topics strongly increased before the year 2000 and then remained on a more or less similar level. During the financial crisis these topics decreased in importance but subsequently strongly increased again. The extent of digitalization topics in 2018 is considerably above the level before the financial crisis. This implies that digitalization topics appear to become more important to banks over time. The development of the keywords in context sentiment in Figure 2c is also quite noteworthy. While the sentiment of the context was almost neutral or positive until the beginning of the new century, it became more and more negative afterwards. From 2012 on this development consolidated but the sentiment remained clearly negative on average. This hints towards banks often considering digitalization as a challenge rather than a competitive advantage.

Figure 3 then provides a representation of what topics are most important over time. It shows the frequency of the ten most common keywords divided by the overall frequency of these keywords by year. When assessing the most important topics in 1993, it is interesting to notice that pre-internet era technologies and in particular automated teller machines dominate. There are two important new topics that enter the plot over the years. This is first the appearance and increasing extent of internet related terms from 1995 on and then the appearance and increasing extent of the word ”cyber” as in ”cyber security” or ”cyber crime” from 2011 on. These are interesting aspects as they reflect the growing importance of online services from the end of the 90s on and an increased actual or perceived threat of cyber crime activities from 2011 on.[6]

5 Empirical Analysis

This section presents the empirical results. It uses the empirical approach outlined in Section 4.1 starting with the results for the frequency based measure and then proceeds with the results for the context sentiment based measure. The frequency based measure aims at approximating the digitalization efforts of banks, whereas the sentiment based measure aims at approximating the capability to successfully implement digitalization. In a first step of our analysis, we evaluate, whether our text mining approach truly captures the digitalization efforts of banks. In a second step, we analyse the relation of our digitalization measures with performance measures.

The results for the frequency based measure are displayed in Table 2. The table lists the three models with the TF.IDF.SCORE as an independent variable and the performance variables as dependent variables presented in the left three columns. There is no significant relation of the TF.IDF.SCORE to the performance

[6] The use of cyber security related terms might be less informative of efforts related to digitalization or might affect the bank performance and cost structure in different ways than usual efforts. We discuss this in the robustness checks in Section 6.

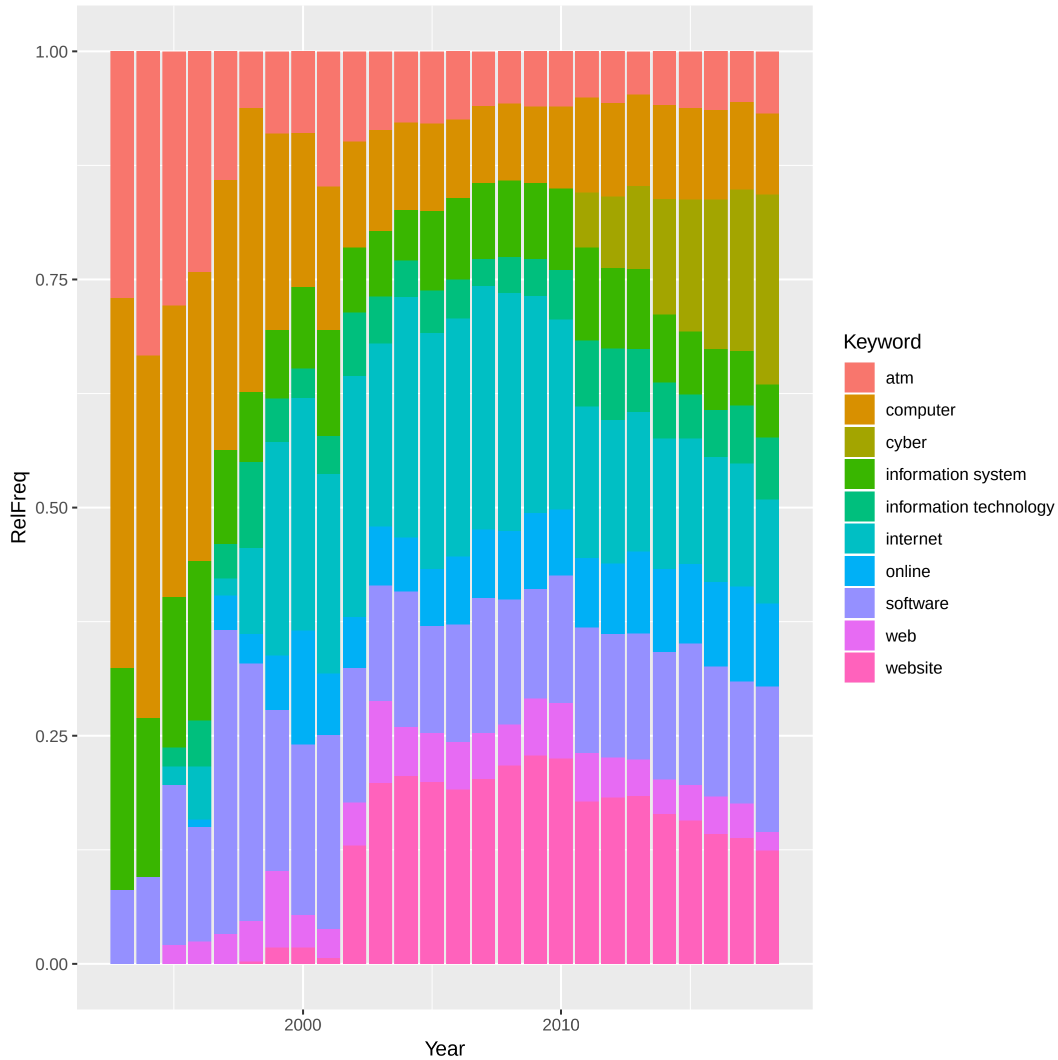


Figure 3: Ten most frequent keywords by year

This figure displays the relative frequency of the ten most frequent keywords by year.

in all three performance variables. This is in line with results from the IT profitability paradox literature that do not indicate a clear relation of IT investments on the performance. The table lists the models for the cost structure variables in the three columns on the right of the table. In line with the expected substitution of personnel expenses by other operating expenses, the coefficient of the TF.IDF.SCORE is negative and significant for the PERS.EXP.NON.INT.EXP and positive and significant for the OTHER.EXP.NON.INT.EXP. Therefore, our results are in accordance with the findings of Martín-Oliver and Salas-Fumás (2008). It is noteworthy that the coefficient in the sixth column using the NON.INT.EXP.TA as a dependent variable is not significant. This points towards a sole substitution effect while the overall cost that is necessary to maintain a certain level of business remains unchanged. Furthermore, a more frequent reporting of digitalization

Table 2: Regression results - term frequency-inverse document frequency

This table displays the regression results using the term frequency-inverse document frequency. The table reports standard errors in parentheses. The standard errors are clustered by year and bank. *p<0.1; **p<0.05; ***p<0.01

	DV						
	ROAE	ROAA	CIR	PERS.EXP.NON.INT.EXP	OTHER.EXP.NON.INT.EXP	NON.INT.EXP.TA	NEW.PAT.EMPL
TF.IDF.SCORE	-0.093	-0.004	-0.013	-0.003^{*}	0.003^{*}	-0.0002	0.012^{*}
	(0.163)	(0.015)	(0.328)	(0.002)	(0.002)	(0.001)	(0.007)
logTA	-0.218	0.060	-1.960	-0.012	0.012	-0.001	0.015
	(1.267)	(0.094)	(2.023)	(0.009)	(0.009)	(0.002)	(0.018)
TETA	50.033	8.391^{***}	-65.711^{**}	-0.097	0.097	0.011	0.085
	(38.091)	(2.567)	(33.218)	(0.209)	(0.209)	(0.042)	(0.432)
LIQU.ASSETS.TA	-0.147	-0.016^{*}	0.351^{***}	0.002^{**}	-0.002^{**}	-0.00005	0.0002
	(0.133)	(0.009)	(0.108)	(0.001)	(0.001)	(0.0001)	(0.001)
NPL.RATIO	-1.973^{***}	-0.156^{***}	0.791	-0.011^{***}	0.011^{***}	0.002^{*}	0.012
	(0.527)	(0.038)	(0.870)	(0.003)	(0.003)	(0.001)	(0.008)
Observations	754	754	760	752	752	752	602
Adjusted R^2	0.530	0.519	0.573	0.736	0.736	0.679	0.552
F-test (p-value)	0.000	0.000	0.000	0.000	0.000	0.000	0.000

terms coincides with a higher number of IT patent filings, which further validates our text mining approach.

Regarding the control variables, banks' total assets do not appear to have a clear relationship with performance measures or the cost structure of banks. Banks with a higher ratio of equity to total assets have a higher ROAA and a lower CIR. Banks that hold a higher share of their assets in liquid assets perform worse on the performance measures and tend to have more personnel expenses in their cost structure. A higher NPL.RATIO is related to a weaker bank performance.

Table 3 displays the results for the keywords in context measure. When assessing the coefficients of the KWIC.SENTI with the performance measures as dependent variables, it is apparent that there is a strong relation of the sentiment to the performance measures. In the case of ROAE and ROAA, both coefficients are positive and significant. In the case of CIR, the coefficient is negative and significant. Overall, the performance of a bank is stronger, when it reports digitalization related terms in a more positive context. We interpret this in a way that banks which are more capable of successfully implementing digitalization have a higher bank performance. While the success of digitalization efforts is clearly related to the performance, it is not clearly related to the cost structure. This points out that the cost structure develops independent of whether digitalization efforts are successful. This seems to indicate that the increased bank performance is not due to the cost structure but due to a stronger capacity to generate income. The coefficients of the

Table 3: Regression results - keywords in context sentiment

This table displays the regression results using the keywords in context sentiment. The table reports standard errors in parentheses. The standard errors are clustered by year and bank. *p<0.1; **p<0.05; ***p<0.01

	DV						
	ROAE	ROAA	CIR	PERS.EXP.NON.INT.EXP	OTHER.EXP.NON.INT.EXP	NON.INT.EXP.TA	NEW.PAT.EMPL
KWIC.SENTI	7.499^{**}	0.552^{*}	-19.107^{***}	0.028	-0.028	-0.011	0.002
	(3.735)	(0.290)	(6.837)	(0.042)	(0.042)	(0.009)	(0.093)
logTA	-0.334	0.051	-1.589	-0.011	0.011	-0.001	0.014
	(1.290)	(0.093)	(1.877)	(0.009)	(0.009)	(0.002)	(0.017)
TETA	51.018	8.467^{***}	-69.352^{**}	-0.094	0.094	0.009	0.043
	(38.712)	(2.603)	(33.280)	(0.209)	(0.209)	(0.042)	(0.417)
LIQU.ASSETS.TA	-0.155	-0.017^{**}	0.367^{***}	0.002^{**}	-0.002^{**}	-0.00004	0.0001
	(0.135)	(0.009)	(0.108)	(0.001)	(0.001)	(0.0001)	(0.002)
NPL.RATIO	-1.953^{***}	-0.155^{***}	0.759	-0.011^{***}	0.011^{***}	0.002^{*}	0.013
	(0.523)	(0.038)	(0.852)	(0.003)	(0.003)	(0.001)	(0.008)
Observations	754	754	760	752	752	752	602
Adjusted R^2	0.533	0.521	0.580	0.734	0.734	0.682	0.546
F-test (p-value)	0.000	0.000	0.000	0.000	0.000	0.000	0.000

control variables are qualitatively similar to the ones in Table 2.

Table 4 lists the coefficients of the independent variables for models with a lag of one and a lag of two years. This is done to assess over how many periods the relation of the independent variables to the performance and the cost structure is lasting. The first two sections of the table list the results for the frequency based measures. The second two sections of the table list the results for the keywords in context sentiment. The results with a lag of one are qualitatively similar to the results with no lags. When considering two lags, the frequency based measure is not related to the cost structure or patents and the sentiment based measure still relates to bank performance. This indicates that, while digitization efforts seem to be implemented in a timely manner after reporting, successfully implemented digitalization efforts have a long-lasting impact on bank performance.

6 Robustness Checks

In this section, we aim at discussing possible objections a reader might have considering the results from Section 5. We account for the overall sentiment in annual reports, financial crisis years, and the inclusion of cyber security terms.

Table 4: Regression results - term frequency-inverse document frequency and keywords in context sentiment with lags

This table displays the regression results using the term frequency-inverse document frequency and the keywords in context sentiment with lags of one and two. The controls are the same as in Table 2 and Table 3. The table reports standard errors in brackets. The standard errors are clustered by year and bank. *p<0.1; **p<0.05; ***p<0.01

	DV						
	ROAE	ROAA	CIR	PERS.EXP.NON.INT.EXP	OTHER.EXP.NON.INT.EXP	NON.INT.EXP.TA	NEW.PAT.EMPL
TF.IDF.SCORE_lag1	-0.186	-0.011	0.234	-0.004^{*}	0.004^{*}	0.0003	0.010^{*}
	(0.222)	(0.019)	(0.381)	(0.002)	(0.002)	(0.0005)	(0.006)
Controls	Yes	Yes	Yes	Yes	Yes	Yes	
Observations	726	726	732	724	724	724	591
Adjusted R^2	0.531	0.526	0.577	0.738	0.738	0.681	0.550
F-test (p-value)	0.000	0.000	0.000	0.000	0.000	0.000	0.000
TF.IDF.SCORE_lag2	-0.158	-0.013	0.392	-0.004	0.004	0.001	0.005
	(0.284)	(0.022)	(0.342)	(0.002)	(0.002)	(0.0004)	(0.004)
Controls	Yes	Yes	Yes	Yes	Yes	Yes	
Observations	697	697	703	695	695	695	579
Adjusted R^2	0.534	0.534	0.581	0.750	0.750	0.690	0.548
F-test (p-value)	0.000	0.000	0.000	0.000	0.000	0.000	0.000
KWIC.SENTI_lag1	14.014^{**}	0.979^{**}	-18.553^{***}	0.015	-0.015	-0.012	-0.072
	(5.534)	(0.404)	(5.444)	(0.042)	(0.042)	(0.009)	(0.075)
Controls	Yes	Yes	Yes	Yes	Yes	Yes	
Observations	726	726	732	724	724	724	591
Adjusted R^2	0.537	0.531	0.583	0.736	0.736	0.683	0.546
F-test (p-value)	0.000	0.000	0.000	0.000	0.000	0.000	0.000
KWIC.SENTI_lag2	20.064^{***}	1.440^{***}	-14.350^{**}	0.036	-0.036	-0.015^{*}	-0.043
	(6.593)	(0.432)	(6.002)	(0.046)	(0.046)	(0.009)	(0.075)
Controls	Yes	Yes	Yes	Yes	Yes	Yes	
Observations	697	697	703	695	695	695	579
Adjusted R^2	0.548	0.544	0.583	0.749	0.749	0.691	0.547
F-test (p-value)	0.000	0.000	0.000	0.000	0.000	0.000	0.000

Our results for the keywords in context sentiment depend on the idea that we measure how successful banks' digitalization efforts are. This is a crucial characteristic of our empirical methodology. However, a more positive reporting on digitalization aspects might in fact proxy for the general positivity in the annual report and thus relate to the bank performance. In order to test the robustness of our results towards the general positivity of annual reports, we compute the overall sentiment for each annual report and include it as an additional control variable. The respective results are presented in Table A.1. The results remain qualitatively unchanged.

Our sample contains the years of the financial crisis. Without any specific expected direction, this might bias our results, as banks' performance was affected substantially. We therefore also estimate our models with exclusion of the years 2008 to 2010. The results are reported in Table A.1. All results remain qualitatively unchanged.

One could further argue that cyber security terms do not fit in the purpose of measuring digitalization efforts. The use of cyber security terms might as well result from reactions to external pressure such as increased cyber attacks or a generally different attention towards this topic. Therefore, the use of these terms might not result from strategic decisions that aim at creating a more competitive organization. Table A.1, therefore, reports results when excluding the cyber security terms in our analysis. The results remain qualitatively unchanged. Interestingly, the coefficient for the KWIC.SENTI is now negative and significant in the model with the NON.INT.EXP.TA as a dependent variable. This could indicate that successful digitalization is also effective via reducing costs besides being effective in generating better profits.

7 Conclusion

Digitalization plays a crucial role in the business models of banks. Yet, in the literature on information technology and company performance it is unclear whether and how information technology can increase performance. One central theoretical consideration is that information technology can be considered a commodity that is available to all market participants. Although companies invest vast amounts of money into IT, some authors therefore argue that IT does not provide a lasting advantage over competitors. This is called the IT profitability paradox. As the digital transformation fundamentally changes the way companies operate, it is a highly relevant question if and under what circumstances information technology can increase bank performance.

The existing empirical literature to date is rather small and indecisive on whether and how IT increases performance. The sparseness of this literature, despite the strong practical relevance, is arguably related to the difficulty to measure digitalization efforts as these are not reported in a structured way.

To overcome this problem, we suggest text mining based digitalization measures to extract information directly from banks' annual reports. The basic idea is that a frequent use of digitalization keywords in these reports is indicative of more digitalization efforts in the respective organization. In line with this argument, we find that a more frequent use of digitalization terms is related to a substitution of personnel expenses by other expenses and furthermore to a higher number of IT patent filings. At the same time and in line with the IT profitability paradox, the measure is not clearly related to increases in performance.

We, additionally, suggest a second measure that accounts for the sentiment of the context in which the digitalization keywords are used. The rationale here is that a more positive reporting of these terms is indicative of more successful digitalization efforts and a better capability to use IT. In fact, this measure is clearly related to an increased performance using several performance measures. Our results, therefore, provide evidence for explanations of the IT profitability paradox that link the paradox to organizational capabilities.

Appendix

Table A.1: Regression results: Robustness overall sentiment, financial crisis and cyber security terms

This table displays regression results using the overall sentiment of annual reports as an additional control variable, excluding the years 2008 to 2010 and excluding terms related to cyber security. The table reports standard errors in brackets. The standard errors are clustered by year and bank. *p<0.1; **p<0.05; ***p<0.01

		DV						
		ROAE	ROAA	CIR	PERS.EXP.NON.INT.EXP	OTHER.EXP.NON.INT.EXP	NON.INT.EXP.TA	NEW.PAT.EMPL
Overall Sent.	KWIC.SENTI	6.331*	0.454	-16.584**	0.016	-0.016	-0.009	0.043
		(3.537)	(0.280)	(6.433)	(0.040)	(0.040)	(0.009)	(0.076)
	Adjusted R^2	0.586	0.576	0.635	0.768	0.768	0.724	0.617
	F-test (p-value)	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	Observations	754	754	760	752	752	752	602
Financial Crisis Excluded	TF.IDF.SCORE	-0.049	0.001	-0.064	-0.004**	0.004**	-0.000	0.012
		(0.153)	(0.015)	(0.376)	(0.002)	(0.002)	(0.001)	(0.008)
	Adjusted R^2	0.586	0.471	0.673	0.776	0.776	0.739	0.657
	F-test (p-value)	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	KWIC.SENTI	8.489**	0.606**	-18.448***	0.047	-0.047	-0.009	0.002
		(4.087)	(0.301)	(7.037)	(0.049)	(0.049)	(0.010)	(0.105)
	Adjusted R^2	0.591	0.475	0.679	0.773	0.773	0.740	0.653
	F-test (p-value)	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	Observations	646	646	649	641	641	641	494
Cyber Security Excluded	TF.IDF.SCORE	-0.073	-0.001	-0.033	-0.003*	0.003*	-0.000	0.012*
		(0.159)	(0.014)	(0.325)	(0.002)	(0.002)	(0.001)	(0.006)
	Adjusted R^2	0.580	0.570	0.618	0.763	0.763	0.712	0.607
	F-test (p-value)	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	KWIC.SENTI	9.559**	0.665**	-26.480***	0.038	-0.038	-0.021**	0.071
		(4.048)	(0.284)	(8.220)	(0.048)	(0.048)	(0.009)	(0.072)
	Adjusted R^2	0.582	0.571	0.628	0.762	0.762	0.719	0.602
	F-test (p-value)	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	Observations	754	754	760	752	752	752	602

Table A.2: Correlation matrix

This matrix reports Pearson correlation coefficients for the dependent variables, the digitalization measures, and the control variables.

	TF	REL.TF	TF.IDF.SCORE	KWIC.SENTI	ROAE	ROAA	CIR	PERS.EXP.NON.INT.EXP	OTHER.EXP.NON.INT.EXP	NON.INT.EXP.TA	NEW.PAT.EMPL	logTA	TETA	LIQU.ASSETS.TA	NPL.Ratio
TF	1	.64	.56	-.30	-.16	-.07	-.04	-.08	.08	-.18	.13	.02	.17	-.05	.05
REL.TF	.64	1	.39	-.05	.01	.03	-.13	.05	-.05	-.13	-.04	-.20	.03	-.09	-.10
TF.IDF.SCORE	.56	.39	1	-.25	-.14	-.08	.02	.04	-.04	-.12	.05	-.07	.16	-.08	.05
KWIC.SENTI	-.30	-.05	-.25	1	.29	.19	-.19	-.06	.06	.08	.08	.15	-.28	.13	-.19
ROAE	-.16	.01	-.14	.29	1	.92	-.48	.28	-.28	.03	-.01	.07	-.14	.05	-.53
ROAA	-.07	.03	-.08	.19	.92	1	-.55	.26	-.26	.04	.03	.05	.13	-.08	-.48
CIR	-.04	-.13	.02	-.19	-.48	-.55	1	-.30	.30	.43	-.06	.01	-.18	.18	.26
PERS.EXP.NON.INT.EXP	-.08	.05	.04	-.06	.28	.26	-.30	1	-1	-.27	-.13	-.13	-.06	.10	-.24
OTHER.EXP.NON.INT.EXP	.08	-.05	-.04	.06	-.28	-.26	.30	-1	1	.27	.13	.13	.06	-.10	.24
NON.INT.EXP.TA	-.18	-.13	-.12	.08	.03	.04	.43	-.27	.27	1	.17	.01	-.02	.05	.24
NEW.PAT.EMPL	.13	-.04	.05	.08	-.01	.03	-.06	-.13	.13	.17	1	.33	.09	.20	.06
logTA	.02	-.20	-.07	.15	.07	.05	.01	-.13	.13	.01	.33	1	-.18	.49	-.01
TETA	.17	.03	.16	-.28	-.14	.13	-.18	-.06	.06	-.02	.09	-.18	1	-.29	.12
LIQU.ASSETS.TA	-.05	-.09	-.08	.13	.05	-.08	.18	.10	-.10	.05	.20	.49	-.29	1	.01
NPL.Ratio	.05	-.10	.05	-.19	-.53	-.48	.26	-.24	.24	.24	.06	-.01	.12	.01	1

References

Aral, S. and P. Weill (2007). IT assets, organizational capabilities, and firm performance: How resource allocations and organizational differences explain performance variation. *Organization Science 18*(5), 763–780.

Beccalli, E. (2007). Does IT investment improve bank performance? Evidence from Europe. *Journal of Banking & Finance 31*(7), 2205–2230.

Bellstam, G., S. Bhagat, and J. A. Cookson (2019). A text-based analysis of corporate innovation. *Working Paper*.

Berger, A. N. and C. H. Bouwman (2013). How does capital affect bank performance during financial crises? *Journal of Financial Economics 109*(1), 146–176.

Bresnahan, T. F., E. Brynjolfsson, and L. M. Hitt (2002). Information technology, workplace organization, and the demand for skilled labor: Firm-level evidence. *The Quarterly Journal of Economics 117*(1), 339–376.

Brynjolfsson, E. (1993). The productivity paradox of information technology. *Communications of the ACM 36*(12), 66–77.

Brynjolfsson, E., L. M. Hitt, and S. Yang (2002). Intangible assets: Computers and organizational capital. *Brookings Papers on Economic Activity 2002*(1), 137–181.

Carr, N. (2003). IT doesn't matter. *Harvard Business Review 81*(5), 41.

Casolaro, L. and G. Gobbi (2007). Information technology and productivity changes in the banking industry. *Economic Notes 36*(1), 43–76.

Cecchini, M., H. Aytug, G. J. Koehler, and P. Pathak (2010). Making words work: Using financial text as a predictor of financial events. *Decision Support Systems 50*(1), 164–175.

Dedrick, J., V. Gurbaxani, and K. L. Kraemer (2003). Information technology and economic performance: A critical review of the empirical evidence. *ACM Computing Surveys (CSUR) 35*(1), 1–28.

Ganglmair, B. and M. Wardlaw (2017). Complexity, standardization, and the design of loan agreements. *Working Paper*.

Hagenau, M., M. Liebmann, and D. Neumann (2013). Automated news reading: Stock price prediction based on financial news using context-capturing features. *Decision Support Systems 55*(3), 685–697.

Hendry, S. and A. Madeley (2010). Text mining and the information content of Bank of Canada communications. *Working Paper*.

Hernando, I. and M. J. Nieto (2007). Is the internet delivery channel changing banks' performance? The case of Spanish banks. *Journal of Banking & Finance 31*(4), 1083–1099.

Hitt, L. M. and E. Brynjolfsson (1996). Productivity, business profitability, and consumer surplus: Three different measures of information technology value. *MIS Quarterly 20*(2), 121–142.

Ho, S. J. and S. K. Mallick (2010). The impact of information technology on the banking industry. *Journal of the Operational Research Society 61*(2), 211–221.

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

Jegadeesh, N. and D. Wu (2013). Word power: A new approach for content analysis. *Journal of Financial Economics 110*(3), 712–729.

Jegadeesh, N. and D. A. Wu (2017). Deciphering fedspeak: The information content of FOMC meetings. *Working Paper*.

Kleis, L., P. Chwelos, R. V. Ramirez, and I. Cockburn (2012). Information technology and intangible output: The impact of it investment on innovation productivity. *Information Systems Research 23*(1), 42–59.

Koetter, M. and F. Noth (2013). IT use, productivity, and market power in banking. *Journal of Financial Stability 9*(4), 695–704.

Kohli, R. and S. Devaraj (2003). Measuring information technology payoff: A meta-analysis of structural variables in firm-level empirical research. *Information Systems Research 14*(2), 127–145.

Lichtenberg, F. R. (1995). The output contributions of computer equipment and personnel: A firm-level analysis. *Economics of Innovation and New Technology 3*(3-4), 201–218.

Loughran, T. and B. McDonald (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *The Journal of Finance 66*(1), 35–65.

Martín-Oliver, A. and V. Salas-Fumás (2008). The output and profit contribution of information technology and advertising investments in banks. *Journal of Financial Intermediation 17*(2), 229–255.

Mithas, S., A. R. Tafti, I. Bardhan, and J. M. Goh (2012). Information technology and firm profitability: mechanisms and empirical evidence. *MIS Quarterly 36*(1), 205–224.

Parsons, D., C. C. Gotlieb, and M. Denny (1993). Productivity and computers in Canadian banking. *Journal of Productivity Analysis 4*(1), 95–113.

Saha, P., I. Bose, and A. Mahanti (2016). A knowledge based scheme for risk assessment in loan processing by banks. *Decision Support Systems 84*, 78–88.

Shah, R. and H. Shin (2007). Relationships among information technology, inventory, and profitability: An investigation of level invariance using sector level data. *Journal of Operations Management 25*(4), 768–784.

Shu, W. and P. A. Strassmann (2005). Does information technology provide banks with profit? *Information & Management 42*(5), 781–787.

Soh, C. and M. L. Markus (1995). How IT creates business value: A process theory synthesis. *ICIS 1995 Proceedings*, 4.

Solow, R. M. (1987). We'd better watch out. *New York Times Book Review 36*.

Stratopoulos, T. and B. Dehning (2000). Does successful investment in information technology solve the productivity paradox? *Information & Management 38*(2), 103–117.

Tetlock, P. C., M. Saar-Tsechansky, and S. Macskassy (2008). More than words: Quantifying language to measure firms' fundamentals. *The Journal of Finance 63*(3), 1437–1467.

Uysal, A. K. and S. Gunal (2014). The impact of preprocessing on text classification. *Information Processing & Management 50*(1), 104–112.

Wang (2010). Chasing the hottest IT: Effects of information technology fashion on organizations. *MIS Quarterly 34*(1), 63–85.