Universal Journal of Accounting and Finance 6(2): 54-81, 2018
DOI: 10.13189/ujaf.2018.060204 http://www.hrpub.org

Text Mining and Reporting Quality in German Banks - A Cooccurrence and Sentiment Analysis

David Fritz[1,*], **Eugen Tows**[2]

[1]Department of Bank Management, University of Cologne, Germany

[2]Deutsche Bundesbank, Germany

Abstract A bank's annual risk report intends to reduce the information asymmetry between the bank and its stakeholders. Using automated text mining measures, we assess the quality of the reports in terms of their fulfillment of regulatory requirement and identify its main drivers in a panel regression. On a set of 343 risk reports from 30 German banks between 2002 and 2013, we further perform a cooccurrence and sentiment analysis and determine several additional characteristics of the reports' text. Our methods detect discrepancies for the reports of distressed and non-distressed banks and also for different types of banks. Some of these discrepancies might indicate an intended concealment of certain risks of a bank. We find that our text mining measures explain the variance of the reporting quality to a large extent. The number of words is an important factor for the determination of risk reporting quality. The share of positive words in a report reduces its reporting quality on average.

Keywords Text Mining, Sentiment Analysis, Cooccurrence Analysis, Bank, Risk Reports

1 Introduction

In recent years, banks' risks and particularly the management thereof have developed into a major component of the financial business. The continual issuance of new regulatory requirements, such as the Basel I, II, and III accords, oblige German banks to consider a series of risks. These include credit, market, and liquidity risk, in particular, but also operational, legal, and many other risks. It is in the financial institutions's interest to adequately account for its risks, to remain solvent and to be profitable at the same time. Each bank communicates its risks, the extent of these risks, and the risks' management in the bank's risk report. This report is an important instrument to reduce the information asymmetry between the bank and the readers, such as stakeholders, supervisors, or investors. It is also a tool for summarizing the past year's developments and to inform on the aftermath of crises.

Beside the public risk disclosure report, supervisors require German banks to report their risks in a regulatory disclosure report, the "Offenlegungsbericht". Our research concentrates on written text, i. e. on qualitative data, rather than on quantitative numerical data. While regulatory disclosure reports are standardized and do not deliver much qualitative information, risk disclosure reports are suited for qualitative text analysis. The report's informal structure provides banks with a certain degree of leeway to report their risks and risk management. Beside these regulatory requirements, banks have additional possibilities when reporting on the organization of their risk management. For example, some banks publish "Risk Maps" or "Management Cockpits". The description of risk measurements, the interaction of different customers' industries, and economic forecasts, in particular, are inevitably communicated in written form. Consequently, a bank's public risk disclosure report comprises key information that is not available as numerical data. There are two essential ways to process this information: either analysts closely read each single report or we use machine processing to quantify the text using text mining algorithms to extract the relevant information.

Close reading of risk reports is expensive in terms of time and personnel and is also prone to error and subjectivity. In contrast, we find that automated textual analysis performs quickly and is objective to a large extent. Moreover, employing a panel regression, we introduce a model to consistently assess the degree to which a report fulfills the requirements, i. e. its quality. To the best of our knowledge, there has been no automated quality evaluation of risk reports to date.

The analysis of financial content is mostly based on numerical data, although there is a vast amount of suitable data available in text form, such as analyst reports and newspaper articles. In the past, one of the largest problems in examining text was to transform large unstructured data sets into analyzable form. New approaches to handling big data sets and complex text mining algorithms offer the opportunity to analyze and quantify text and thereby extract knowledge and measure the information quality in a scientific way.

The finance-related textual analysis splits into three categories. The first category includes web mining techniques to analyze postings in blogs, wikis, and forums. Users share their thoughts about future market movements, new product information, or investment strategies and by doing this they generate a lot of textual information, which is worthwhile analyzing in many different ways. There are specified web-crawlers to search the web for certain keywords. The second category deals with public value-free text concerning financial topics, such as newspaper articles, analyst reports, or articles about companies in professional journals. These texts potentially have an influence on the recipients' opinion and actions. Therefore, firms should pay special attention to them to preserve their reputation. The third category considers finance-related content generated by the firms themselves, e. g. annual reports, risk disclosure reports, or other management communications. This category is often filled with highly sophisticated expressions, making the analysis even harder. The degree of sophistication depends on the targeted recipients and language characteristics. In this category, it is very important to differentiate between the targeted recipients of the studied text. In the case of financial documents, these could be employees, supervisors, customers or investors.

In general, text is much more versatile than numbers because it provides explanations, opinions, and even behavioral expressions. In the case of analyst reports, text mining algorithms can detect uncertainties and inconsistencies, which simple charts and numerical data analysis would ignore. With automated text analysis, it is also possible to determine sentiments of text. This so-called sentiment detection classifies the mood of texts in a computer-aided objective way. Sentiment analysis, in particular, is sensitive to various kinds of text paying special attention to every written word. On the one hand, supervisory authorities demand risk disclosure reports to be written in an objective manner similar to most of the finance related texts. On the other hand, Wiebe et al. [32] show that 44% of sentences in a finance news collection were found to be subjective although editorial and review articles were excluded from the analysis. Based on this finding, we expect our examined risk reports to be subjective to some extent. Therefore, we should be able to extract sentiments from these texts. The proper analysis of sentiments is one main goal of this study. We do this by applying a polarity analysis based on a finance-specific dictionary and we also apply a modified cooccurrence measure.

A cooccurrence analysis of a preset keyword provides a list of words that occur in the same window, e. g. a sentence or the keyword's neighboring words. This is an improvement over simple frequency analysis, since cooccurrences are able to identify the meaning of a sentence related to a certain keyword. We use the sentence as a window to ensure a wide range of potentially associated words. The derived list of words includes many words that give a certain tone to the sentence and thereby indicate the sentence's meaning. Possible keywords for this study could be "*risk*", "*crisis*", "*profit*", or "*loss*".

The nature of textual data in finance, however, presents further unique challenges. Concerning readability, uniqueness, and complexity, finance-based vocabulary entails some difficulties that we address in our study. For this purpose, we discuss certain techniques and the usage of common dictionaries to handle the above-mentioned challenges.

Another main goal is to look for alternative measures of risk reporting quality. Schlueter et al. [27] are the first to focus on the risk report's quality. They introduce a risk reporting index (RIX), which accounts for different risk types, their bank-specific definitions, and risk measurements. Instead of determining RIX by close reading single reports, we first use algorithmic factors generated by text mining techniques to explain their influence on the RIX. Then, we reproduce and predict the RIX with these factors.

Regarding the estimation and prediction of the RIX our sample consists of 27 German banks. A RIX value for these 27 banks is provided by Schlueter et al. [27]. The annual risk reports cover a period from 2002 to 2013. Hence, we can analyze up to 12 reports per bank. Due to some missing RIX values we base this panel regression on a set of 309 reports. We distinguish between healthy and distressed banks and also different bank types in our sample. Distressed banks are such banks, that defaulted, were downsized, taken over by another bank, or were in need of financial support provided by the government during the observation period. We map distressed banks to a negative pool and healthy banks to a positive pool. We also add three major banks to our descriptive analysis which do not have a RIX value produced by close reading. Applying these definitions, we construct a balanced sample of positive and negative banks and also of different types of banks.

The German banking industry splits into three sectors: private banks and banks with a special business model; cooperative banks; and public savings banks together with federal state banks. To cover the German banking industry, we choose a representative sample consisting of 11 private banks and banks with special business models, 5 cooperative banks, and 14 public savings, government-owned, and federal state banks. We stay in line with Schlueter et al. [27] and follow their bank type linkage. We identify 11 distressed banks according to the above definition. The descriptive analysis is based on 343 risk reports between 2002 and 2013. However, 3 banks lack RIX values and therefore, we exclude their reports from the calibration of the RIX estimation model. While companies in the United States are required to submit their filings to the Edgar system, which is publicly accessible online, there is no German equivalent. Thus, our sample is a hand-collected unique data set.

In the following, we provide a descriptive analysis of the data. Our used text mining measures extract information about the sentiment of texts, the different use of chosen keywords, such as "*risk*" and "*crisis*", and also about the text's readability. We analyze our data set from different perspectives. First, we look for differences between positive and negative banks in their reports' sentiment and the usage of the keyword "*risk*". We expect banks of the negative pool to use more positive words in their risk reports to try to conceal their real business development. Furthermore, "*risk*" is used more often in connection with positive words to color certain risk events. We find differences in risk reporting quality and risk measures considering different bank types. Due to the international connectivity and higher demands in risk reporting for banks, which act as a capital market participant, we expect these banks to provide a more comprehensive report and thus achieve a higher risk reporting quality. Local banks don't have to consider many particular risk types and spend less effort on external risk reporting, mainly giving information to customers and supervisors, but not to potential investors.

Second, in our analysis we also find that text mining measures are quite capable of explaining the variance in reports' quality to a large extent. We determine five factors that have a significant influence on RIX after controlling for the report's year and the applicable reporting standard. In particular, the word count and the proportion of positive words drive these study's results. Although of less significance, the number of a report's paragraphs, the proportion of negative words, and the cooccurrence of risky words are also important. While the word count is positively associated with the RIX we find that the increasing proportion of positive words reduces RIX on average. We conclude that the extensive use of positive words does not improve the presentation of a bank's risks but rather worsens it, e. g. by trying to conceal risks.

In any case, employed in a linear regression, these factors can reproduce and predict RIX most accurately. Consequently, we find that the advanced text mining techniques used in this study can replace close reading of risk reports to determine the reports' disclosure quality, even though, we only consider the textual information of the reports.

2 Literature review

Being a part of general data science, text mining is a quickly emerging discipline of science. We present a brief overview of the related literature that applies text mining algorithms to accounting and finance texts. Furthermore, we review the contemporary literature on the measurement of risk reporting quality.

Pioneer work in financial textual analysis

Frazier et al. [10] are the first to analyze the substance of narrative data in accounting disclosures. They use a method based on word-frequency logic. In particular, this method searches for certain words which occur and cooccur in annual corporate reports. Using computer-aided methods, they develop a dictionary containing 215 words and classify these as positive or negative words. Their sample consists of 74 annual reports of firms from the year 1978. Because computer technology and text mining methods have developed significantly in recent years, current studies use much larger dictionaries and more complex algorithms and factors than simple word count. Nevertheless, a report's length, i. e. its word count, is still important when quantifying text data, such as argued by You and Zhang [33] and Loughran and McDonald [23].

In their analysis of 45 Dow Jones Industrial Average and Internet Index companies, Antweiler and Frank [1] develop a bullishness index to predict stock returns. They study 1.5 million messages posted on Yahoo! Finance and Raging Bull and use newspaper articles concerning the Wall Street Journal as the control group. Based on their calculated linguistic measures, they cannot accurately predict stock returns; but, in an extended analysis they find that these measures help to predict the volatility of the stocks on a daily and intraday basis. Furthermore, they argue that public opinions influence market movements and reflect available information rapidly. Particularly postings with disagreement induce a greater trading volume. Overall, the authors conclude that talk about stocks is not only noise, but rather an influential factor in the analysis of their returns.

Tetlock et al. [31] published a very important and oftentimes referenced work on the textual analysis of financial documents. Based on a data set of S&P500 firms from 1980 to 2004, they use the Dow Jones News Service and Wall Street Journal news articles to generate trading signals. They also examine correlations between text, stock returns, and profits. Thereby, they find that a high percentage of negative words in the news text reduces profits and also that stock prices slightly underreact to these negative words. Furthermore, articles that focus on firm fundamentals, such as income statement, balance sheet, and cash flow, have the highest predictive power of profits and returns because these articles contain the most negative words.

Risk studies connecting quantitative and qualitative data

Li [20] investigates the frequency of risk terms in the risk reports of American companies. He discovers a link between the increase in the risk report's length, measured as the number of words, and changes in the profit of the following year. Moreover, he argues that trading strategies based on this analysis can generate significant excess returns. The occurrence of the words "*risk*" and "*uncertain*" is used as a proxy for the company's risk situation. He builds an index upon the changes in the frequencies of these words over time to quantify the described risk situation. Prospective developments and risk issues are rarely described pessimistically. Thus, Li [20] argues that wherever possible, companies use soft words, such as "*could*", "*may*", and "*might*". These findings are of particular interest to the research within the scope of our work. We look for certain risk-related buzzwords and jointly appearing words. These words give the chosen buzzwords a positive or negative meaning. Furthermore, we base our sentiment analysis on a financial dictionary that also contains soft words.

In another study, Li [21] establishes a link between risk reports' readability and the respective company's profits. He analyzes the Management Discussion and Analysis (MD&A) section of the annual report that describes and analyzes the firm's fundamentals. The author finds that firms with good current performance and a good readability of their MD&A-section, measured by the Gunning-Fog index, tend to have more positive forward-looking statements than firms with bad performance and bad readability. The average tone of these statements is positively associated with the future earnings. However, he finds no evidence that the measured tone in the MD&A section has an influence on the firm's future performance. The measured tone is based on commonly used dictionaries, such as *General Inquirer*, but also *linguistic inquiry*, and *word count*. Li [21] argues that these dictionaries are not suitable for the analysis of finance reports because the tone measurements that are based on these dictionaries cannot significantly contribute to the prediction of the companies' future performance. In order to analyze a text's sentiment and to measure its influence on the risk reporting quality, we use SentiWS, a comprehensive German dictionary.

In their study of financial risk reports of Japanese companies, Shirata and Sakagami [28] argue that a change in a company's financial situation is recognizable in text first and in numerical data later. Hence, the authors assume that text mining techniques can predict insolvency from a company's financial statements. They identify the 20 most important words in a group of companies' financial reports by using a simple frequency analysis. Then, the occurrence of these words is counted for insolvent companies as well as for solvent companies. This procedure reveals considerable differences in terms of frequency and occurrence of certain keywords, such as "*quality*", "*growth*", and "*interim dividend*" in the reports of solvent and insolvent companies. Furthermore, Shirata and Sakagami [28] find that certain non-financial keywords within financial reports can be used to evaluate a corporate's financial situation. The dividend section of the financial statements and words linked to this section play the most important role in their evaluation.

In an earlier study, Kloptchenko et al. [17] find empirical evidence that supports a similar assumption. Analyzing the financial reports of three major telecommunication companies, they argue that the numerical data of a company's report only reflects its past performance. The textual data, however, contains some indication of its future financial performance. In a subsequent study, Shirata et al. [29] confirm this finding.

Shirata et al. [29] perform a cooccurrence analysis of business specific words, such as "*dividends*" and "*income*". Thereby, they account for these words' context. They group descriptive words located around these context words according to bankrupt and non-bankrupt companies and find significant differences between the two groups, particularly concerning the word "*dividends*". As expected, the combination "*no dividends*" is a prominent representative of the insolvent group. The authors conclude that it is possible to derive predictions about the prospective insolvency of a company based on the combined occurrence of context words even if the quantitative data does not justify any such presumption.

This finding contradicts the results of Kohut and Segars [18]. In their linguistic analysis of president's letters of the top and bottom 25 firms of the Fortune 500 based on return on equity, they argue that risks are often masked by usage of the passive voice. The analysis uses word count, record length, number of syllables per word, and defined themes, such as environmental factors, growth and operating philosophy. However, only word count exhibits a significant influence on the companies' profits. They argue that profitable companies use more verbose descriptions than less profitable companies. This finding suggests that good news and public announcements lead to additional elaboration of financial statements.

Risk reporting quality

Campbell and Rattanataipop [3] provide a good overview of empirical studies on risk disclosure. Most of the mentioned studies only concern a few reports or a short observation period. Therefore, compared to previous studies our list of analyzed risk reports is quite extensive. In their study, Campbell and Rattanataipop [3] focus on the six largest banks in the UK between 1995 and 2010. They report that almost all kinds of risk reports contain a discussion of the actual risk situation, that are worth being analyzed. Furthermore, they find that the proportion of important news increases over time. The authors argue that market participants increasingly demand more realistic risk reports, instead of modified copies of previous reports. Kravet and Muslu [19] support this hypothesis. They investigate the association between changes in companies' textual risk disclosures in 10-K filings and changes in stock market and analyst activity around the filings. They find, that annual increases in risk disclosures have an effect on increased stock return volatility and trading volume around and after the filings. This leads to a more risk-sensitive perception of investors.

Studies that are based on the word lists developed by Loughran and McDonald [22], such as Huang et al. [14] and Feldman et al. [8], report stable results concerning the identification of polarity of financial statements. One of the main findings in Loughran and McDonald [22] is the unambiguous usage of negative words. Managers do not negate a negative word to make a positive statement. However, positive words are often used to describe negative facts. Therefore, in a carefully conducted polarity

Category	Maximum score	Percentage
Risk management (in general)	9.5	8.12
Risk capital management	7.5	6.41
Credit risk	35.5	30.34
Liquidity risk	20.5	17.52
Market risk	29.0	24.79
Operational risk	15.0	12.82
Sum	117.0	100.00

Table 1. Composition of the RIX.

analysis of financial text, we would expect a negative tone to be more ambiguous than a positive one. Positive words contain little incremental information according to Tetlock et al. [31], Loughran and McDonald [22], and Jegadeesh and Wu [15]. In order to incorporate this finding, Tetlock et al. [31] concentrate their research on negative words exclusively. Following this approach they produce accurate predictions of financial performance.

Loughran and McDonald [24] examine 10-K-filings published by companies from the US. They introduce a procedure that is based on sentiment analysis to identify positive and negative polarities of text. They find general word lists, such as *Diction*, to be inadequate for the sentiment analysis of financial text. In a financial context, these word lists have only little or no validity because certain words adopt a different meaning in finance, such as "*respect*", "*security*", "*power*" and "*authority*", which have an optimistic tone in *Diction*.

Schlueter et al. [27] are the first to measure the quality of risk reports in a representative sample of 30 German banks. They analyze 289 risk reports from 2002 to 2011 to introduce the risk reporting quality index RIX. This index measures the fulfillment of the different requirements for the risk report of German banks by grouping the reports into six categories. These categories are: general risk management; risk capital management; credit risk; liquidity risk; market risk; and operational risk. Each category has an individual maximum score dependent on its set requirements. The score depends on the quality and extent of strategic information provided, evaluation of the risk situation, and fulfillment of regulatory requirements. Credit and market risks claim the heaviest weight within the RIX. They amount to 55% of the RIX score in total. This weight reflects their share of a bank's overall risk and thus, their importance to the bank's risk management. Therefore, the reporting quality of these two risk categories in particular has an impact on the reliability of the bank's risk representation. Only Hope et al. [13] conduct a similar study. They use a computer algorithm to conduct the specificity of firms' qualitative risk-factor disclosures and uses this measure to examine the benefits of such disclosures being specific. They state, that their findings suggest that improved corporate risk reporting, in particular providing more specific risk-factor disclosures, enhances risk understanding and benefits users of financial statements. We support this hypothesis by dividing the RIX into a qualitative and a quantitative part and analyzing each part itself.

Table 1 introduces the composition of the RIX. The index sums up to a maximum score of 117, which breaks down to 146 qualitative and quantitative items. In our study, we transform the RIX to the unit interval to prevent unnecessary scaling during the analysis. The qualitative part outweighs the quantitative part 66 % to 34 % . Although most banks report quantitative items in figures and tables, the content is also contained in the written text. The description of risk measurements and models, assumptions about loss distributions, and risk horizons are partially assigned to the quantitative part. When we exclusively use written information, only a small amount of RIX-relevant information is lost. Strictly speaking, the risk reports meet the qualitative requirements to a larger extent than the quantitative requirements. Nonetheless, to paint a comprehensive picture of a bank's risk reporting quality, we analyze the RIX and also both its qualitative and quantitative part. We expect each part to have different drivers and consequently to have a different power of explaining variations of the RIX.

Schlueter et al. [27] produce several descriptive results. First, they observe a continuous increase in the quality of risk reporting in recent years, as shown in Figure 2. While the mean RIX value in 2002 is 42.0, it rises to 80.3 in 2011. Also, the overall minimum value of the respective years rises from 8.5 to 43.0 and the maximum value rises from 74.0 to 105.0. These findings show that particularly banks with small RIX scores considerably improved their risk reports' quality. However, the full RIX score has not been reached yet.

Second, there are considerable differences in the descriptive quality of the examined risk categories. Market risk yields the highest degree of fulfillment, although capital risk management has substainable increased since 2002. At the beginning of the analysis period, operational and liquidity risk were not reported and mostly not even considered by many banks.

The third result addresses differences between the three types of banks. Private banks and banks with a special business model exhibit the highest average RIX until 2006. Their reports' quality considerably outperforms the reporting quality of public savings banks and federal state banks during the period 2002 to 2006. In 2005, public guarantees for savings banks and federal state banks were abolished. In the aftermath, these banks reorganized their business models and since then have also been addressing a wider audience with their annual reports than before.

Fischer et al. [9] report on this change in business models. Since 2006, public savings banks and federal state banks achieve a higher average RIX score than private banks and banks with a special business model. Cooperative banks consistently retain the lowest average RIX scores over the whole period of the analysis. Schlueter et al. [27] find that risk reporting quality depends on the banks proximity to the capital market. Private banks, in particular, should set high standards for their disclosure reports because these reports primarily address potential investors. The study reveals a leap in risk reporting quality in 2007. In that year, the regulatory framework in Germany changed with the implementation of IFRS 7 (International Financial Reporting Standards) and Basel II. We find a similar impact on our results in 2007.

From a methodological perspective, the main shortcoming of the study is the extensive effort of close reading each single report to calculate the RIX. Furthermore, the RIX might not be independent of subjective evaluation, moreover, analysts make mistakes. By calculating computer aided indices, we eliminate these major disadvantages in our study. Our determined measures are objective to a high degree and can be calculated in very short time. We also predict current and future RIX scores accurately using linear regression with text mining factors.

3 Regulatory Framework

The German banking industry consists of three sectors. The first sectors comprises private banks, direct banks, and banks with a special business model. According to Deutsche Bundesbank [5], more than 70% of all German banks' assets are covered by these banks in 2014. The second sector contains cooperative banks and associated specialized financial institutions. The third sector contains public savings banks and associated specialized financial institutions. The second and third sectors together account for more than 80% of all German banks in 2014 (see Deutsche Bundesbank [5]).

The pillars are roughly categorized by their banks' business models. Private banks mainly pursue profit maximization. Cooperative and public savings banks focus on the financial support of small and medium sized enterprises (SMEs) as well as private customers. These different approaches have a direct influence on the importance of annual reports. Reports of private banks primarily address stakeholders, shareholders, and potential investors, whereas cooperative and public savings banks have fewer and oftentimes local recipients. Our sample comprises a representative cross section of all three types of German banks and it also includes banks with special business models.

The regulatory requirements of financial reports are based on four different areas of the relevant law, which lead to different disclosure obligations for reporting banks. These are: corporate law; trade law; supervisory law; and capital market law. All disclosure obligations have one common objective, that is the reduction of information asymmetries between banks' management and the recipients of their reports'.

There are different means of disclosure for banks, all of them are mandatory. The main disclosure texts concerning the risk situation divided into two different reports. On the one hand, there is an external regulatory disclosure report, the "Offenlegungsbericht" and on the other hand, there is a risk disclosure report that is included in the annual report. The "Offenlegungsbericht" is not part of the annual report and has to be published separately. It is highly standardized and comprises a rather small amount of qualitative information. In contrast, there are almost no reporting requirements for banks' risk management organization. To cover these, we rely on the analysis of risk disclosure reports.

By publishing risk disclosure reports, banks try to satisfy banking supervision by fulfilling regulatory requirements and investors, other banks, and customers by delivering demanded information. In this section, we provide a brief introduction to the regulatory requirements of risk disclosure reports for German banks. Although regulation evolves quickly, we describe the recent historic development of the risk reports' requirements. Furthermore, we try to explain how supervisory requirements influence the disclosure behavior of banks.

All German banks must file their financial statements in accordance with the German commercial code (”Handelsgesetzbuch“ (HGB)). § 340a HGB requires banks to publish a situation report. The risk disclosure report is part of this situation report. The latter reports economic forecasts and opportunities as well as risks that the bank is facing.

HGB provides the statutory framework, which is specified by the German Accounting Standards (GAS). The GAS consist of a variety of standards which address consolidated financial statements. The relevant standards concerning risk reports are GAS-5, GAS-5-10, GAS-15, and GAS-20. GAS-5 deals with risk reporting in particular. GAS-5-10 is concerned with risk reporting of financial institutions and financial service providers and GAS-15 deals with management reporting. GAS-5 and GAS-15 were introduced in 2001 and 2004 respectively. They were applicable until December 2012. GAS-20 regulates group management reports. After its introduction in 2012 it became obligatory in 2013. It replaced among others GAS-5, GAS-15, and other accounting standards.

GAS-15 requires financial institutions to publish forecasts on the development of their business based on available key information. All risks concerning these forecasts have to be named if they might have an influence on the actions of the reports' recipients. These forecasts are supposed to be written in a positive or negative manner. This supports our hypothesis that it might be worthwhile to examine the influence of reports' sentiment on the RIX.

The GAS require banks to use the management approach for their reporting. According to this approach, a bank shall not only publish balance sheet data (balance sheet approach) or regulatory required data (regulatory approach), but also data that is used for its internal management. The GAS enhance the risk disclosure quality of financial institutions. This standard has a particular impact on a bank's evaluation of its own business situation and on its reporting of non-financial key data. Our sample covers the period 2002–2013 and the applicable accounting standards of these years. Controlling for this factor should increase the robustness of our findings.

Another way to file financial statements is provided by the International Accounting Standards (IAS) and the International Financial Reporting Standards (IFRS). In particular, internationally operating private banks must file their balance sheet according to IAS 1 and IFRS 7. While IAS 1 is concerned with the representation of financial statements, IFRS 7 set standards for disclosing different types of risk, particularly credit, market, operational, liquidity, business, and strategic risks. IFRS 7 separates the nature and extent of risk exposure which arises from financial instruments into qualitative (IFRS–7.33) and quantitative (IFRS–7.34) disclosures. The qualitative disclosures describe the organization of the risk management and the risk exposures for each type of financial instrument that the bank is involved in. Furthermore, management's objectives, policies, and processes for managing those risks should also be described. Quantitative disclosures comprise quantitative data about the main risk types.

Natonal and international accounting standards differ in terms of the risk reports' design and volume, i. e. the number of words. In a recent study, Campbell and Rattanataipop [3] find that an increase in the length of risk reports does not lead to an increase in expressiveness. Following the authors' argumentation, a larger word count arises from new and additional regulatory requirements and therefore, fulfills the expectations of supervisors and investors. Hartmann [12] analyzes the risk reporting quality of European and US banks. He finds that word count is an indicator for the risk reporting quality. However, it is not the only explanatory factor. The author also argues that uniform standards for filing risk reports along with minimum reporting requirements would be of considerable help in order to evaluate banks in a consistent and comparable way. The risk reports which we examine in this study are filed in accordance with the HGB. Again we follow Schlueter et al. [27] in doing so. We propose text mining-based measures to consistently analyze the textual information in these risk reports.

The Basel Committee on Banking Supervision [2] formulates eleven principles to optimize the risk reporting process. It emphasizes the importance of an accurate risk data aggregation, which can be achieved by improving IT processes and risk data infrastructure. Some principles address risk reporting practices: risk reports should accurately convey aggregated risk data; and they should be written in a comprehensive, clear, and useful manner. Particularly the latter confronts financial institutions with new challenges in publishing adequate risk reports.

Based on the experiences from the financial crisis, a bank's transparency has become more and more important. Banks' stakeholders increasingly demand extensive and address-oriented transparency to reduce information asymmetry between the banks' management and themselves. Consequently, high quality risk reporting – embedded in a consistent annual report – should be an important factor to gain and secure trust on capital markets. Moreover, it should result in the attainment of refinancing advantages for the reporting bank.

For banks, the regulatory basics result in the complex but crucial situation of risk reporting in a not clearly specified mix of regulation and statutory frameworks, which complicates the generation of comparable information. It is challenging for banks to annually report their risks maintaining a balance between connecting the nationally and internationally established requirements and the adequate presentation of their risk situation. To increase transparency, banks should reduce discrepancies between internal risk management and external risk reporting. This especially applies to banks using capital market-based funding. These banks face massive reputational pressure and rely on low capital costs.

4 Methods

In the following we explain the pre-processing steps of the documents that are necessary to eliminate bias for the determined factors on the risk reporting quality, such as formatting and word elimination. We should say, that our analysis is based on the German language, which provides some difficulties, as described in this section. To make our study understable to a wider audience, we translate all the words into the English language.[1]Furthermore, we describe the sentiment and cooccurrence analysis as well as the estimation models' performance measurements.

4.1 Pre-processing the corpus

To create a consistent word corpus, we extract the risk report from the annual report and exclude figures and tables. The corpus is the basic data format for the analysis of text. It structures the set of texts and prepares these for further analysis.

[1] Our algorithms are mostly programmed in R. We use some text mining packages such as tm and SnowballC. Nevertheless, most of the work is self-programmed.

We define each paragraph of the report as one item of the corpus and add information, such as bank name and type, year of the report, title, subtitle, and sub-subtitle. We remove all numbers, hyperspaces, and special characters and we map certain terms to a single word. For instance, "*Value at Risk*", "*ValueAtRisk*", "*VaR*", etc. are mapped to *valueatrisk*.

In the next step, we fragment the text into sentences. Moreover, every single word is tagged, that means it is categorized into categories of nouns, verbs, adjectives, numbers, and many more. We store this information separately in the corpus' metadata. Then, we eliminate punctuation and remove all stop words. Stop words are common words which have no additional information in terms of a sentence's meaning, such as "*the*", "*are*", or "*a*".

We transform all words to lower case letters. Usually, the next step is to map different word inflections to one word by employing a stemming or lemmatization process. Stemming removes the words' suffixes, so that words such as "*high*", "*higher*", "*highest*", and "*highly*" are all reduced to *high*. In comparison to the English language, stemming is less practicable and more prone to error in German because the inflection process is more complex. In contrast to that, the procedure of lemmatization reduces words to their dictionary form and, therefore, needs most comprehensive dictionaries. It works more precisely than stemming but is not suitable for text which contains special vocabulary. Due to these issues, we do not stem or lemmatize but rather map a list of certain words manually. We do this for all possible inflections of words that are contained in the *SentiWS* list and also for all the different risk types.

In the next step, we eliminate frequently used words, such as "*respectively*", "*referred to*", and "*following*". These words have no meaning in a text mining sense but are not part of the stop word list. We also eliminate rarely used words which are used less often than five times per bank. Finally, we create the document-term-matrix (DTM) to calculate some of our indices. Each call of the DTM matrix counts the frequency of one term in one document. The amount of terms is equal to the amount of columns and the amount of documents is equal to the amount of rows.

4.2 Quantification methods and basic assumptions

Loughran and McDonald [22] argue that it is unlikely that negative words are negated in financial reports, e. g. "*non-negative profits*". However, positive words contain little information according to the literature. Therefore, many studies only focus on negative words and negations. Although it is easy to account for simple negation, most forms of negation are rather difficult to detect.

Word and paragraph count

Word count is a variable counting the amount of words. It also includes stopwords. Other volumetric measures, such as the amount of sentences are highly correlated with word count, thus it is sufficient to analyze word count alone.

Furthermore, we count the number of paragraphs used. We separate the corpus into three different hierarchical levels of sections. These are section, subsection, and sub-subsection e. g."*market risk*" (section), "*strategic orientation of market risk*" (subsection), and "*limits in market risk*" (sub-subsection). We observe this variable to measure the content diversification in a report and its detailedness to some degree. In Table 1 six possible main sections are listed. Many banks in our sample divide their risk report into similar parts.

Sentiment and cooccurrence analysis

Sentiment analysis detects opinions, emotions, and sentiments in text. Risk disclosure reports are assumed to be written in an objective and neutral manner. However, the description of key risk data and its measuring methodologies give banks the opportunity to paraphrase quantitative data. It might be in the banks' interest to provide extensive information in addition to its quantitative data or to withhold certain information even if it is already included in the quantitative data. Using a certain rhetoric, banks might conceal negative modifications and adjustments to their balance sheet or profit and loss account.

Sentiment analysis consists of two parts. These are subjectivity and polarity analysis. The former examines the difference between an objective and a subjective character of terms, sentences, or complete text. The latter analysis identifies positive, neutral, or negative sentiments of a text. Pang et al. [25] classify short text forms into positive and negative terms by using statistical methods, such as support vector machines. These are trained on a small text database to predict values of larger data sets. In their study on movie reviews, Kennedy and Inkpen [16] introduce a rule-based approach. They consider polarity bearing words combined with modifiers, negations, weakenings, or amplifications. Naturally, the approach is based on existing dictionaries that already contain a comprehensive list of positive and negative words. However, due to the specific character of financial text, these dictionaries are not applicable to our data.

The influence of dictionaries in sentiment analysis is crucial to its results. Loughran and McDonald [22] find that financial text mining studies essentially use four different dictionaries. These are: *Henry*; *Harvard's General Inquirer*; *Diction 7*; and *Loughran and McDonald's dictionary*. For languages other than English, Strapparava and Valitutti [30] recommend *WordNet-Affect*, Esuli and Sebastiani [6] propose *SentiWordNet*, and Remus et al. [26] propose *SentiWS*.

In this study, we perform a polarity analysis and try to detect differences in the tone of risk disclosure reports. We also compare the determined tone of distressed banks to that of non-distressed banks. In order to account for finance-specific terms, we use a modified dictionary that is based on *SentiWS*, a finance-related German dictionary provided by the ASV Leipzig (see Remus et al. [26]).

We employ sentiment analysis in several ways. First, we count the frequency of positive and negative terms or phrases. *SentiWS* is a list of words that hold positive and negative polarity. These words have weights within the interval of $[-1, 1]$ according to their polarity. The absolute value of their polarity index shows the intensity of the semantic orientation of the word. Moreover, the list provides a part of speech tag for each word and, if applicable, its inflections. Words such as "*fear*", "*risk*", and "*outlawing*" have a large negative weight. Words such as "*success*", "*importance*", "*happiness*" have a large positive weight. *SentiWS* contains a lot of finance-specific words but since the meaning of several words that are often used in risk disclosure reports differs from their regular meaning, we modify the list slightly. For example, the word "*flat*" has a negative weight within *SentiWS*. Risk reports, however, often use "*flat*" in connection with e. g. "*flat interest rate*" or "*flat hierarchies*". Both expressions are rather neutral than negative. Another example is "*risk*", which naturally appears very often in risk disclosure reports. It has a negative weight within the original *SentiWS*-dictionary. Beside the removal of these words, we also change some semantic orientations. Terms such as "*reduction*", "*cuts*", and "*minimization*" have a different meaning in a financial compared to a non-financial context. Particularly in combination with risk types or risk measurements, their original weight would bias the results of the analysis. Therefore, we invert the signs of these words' polarity weight.

The modified word list contains 1,646 unique positive words and 1,801 unique negative words, net of inflections. Once again, we point out that *SentiWS* is a word list that provides a lot of finance-related words and our modification makes it even more suitable for the analysis of German financial reports.

Furthermore, we calculate cooccurrences. Computing cooccurrences means looking for words that occur in the same window as a preset keyword. Commonly used windows are sentences and paragraphs, however, they might also be neighboring words. We use the sentence as a window to ensure a wide range of potentially associated words that give a certain meaning to our keywords. We choose risk disclosure-specific words, such as "*risk*" and "*crisis*" for the analysis and examine corresponding words, which provide the chosen keywords with a certain tone. For example, consider the following two fictional sentences concerning the key term "*risk*":

> "*We use a range of quantitative tools and metrics to monitor our credit risk reducing activities.*"
>
> "*Operational risk events have become acute.*"

The first sentence obviously has a different tone compared to the second sentence. A simple frequency analysis cannot detect the difference in the tone by counting the term "*risk*" in the two sentences. In both cases, the result is 1. A cooccurrence analysis, however, delivers a list of cooccurrences. In the first case, the list comprises the terms "*monitor*", "*reducing*", and "*range of*", which give the sentence a positive tone. The cooccurrence list of the second sentence comprises only the term "*acute*", which results in a negative tone concerning the term "*risk*". Therefore, a cooccurrence analysis processes and delivers more information than a frequency analysis. However, the interpretation of the former is more complicated.

We only concentrate on risk-related words. Many English terms are composed of two or more words, such as "*default risk*" or "*annual report*". In German, however, these terms are mostly combined into one word, e. g. "*Ausfallrisiko*" or "*Jahresbericht*", respectively. This is particularly true for the different types of risks that banks report. Some banks report up to 94 different risk types. Thus, our list of key risk terms is quite exhaustive. Based on the assumption of no double-negation, we determine 192 negative words and 95 positive words in joint occurrence with the word "*risk*", inflections not counted.

The word "*crisis*" is used less often than "*risk*". Consequently, its cooccurrence list is shorter. We find 7 negative and 4 positive words.

In our cooccurrence analysis, we determine cooccurrences using two different measures. Both measures are widespread in the text mining literature and use the following variables:

$$\begin{aligned} n_a &\text{ number of windows that contain word a} \\ n_b &\text{ number of windows that contain word b} \\ n_{ab} &\text{ number of windows that contain words a and b} \\ n &\text{ number of all windows} \\ wc &\text{ word count.} \end{aligned} \tag{1}$$

The first measure that we use is a modification of the log-likelihood measure. The original log-likelihood test uses the generalized likelihood ratio λ for two parametrized distributions. The counting of the occurrence of two words results in two binomial distributions. For each sentence the value is either 1 in case one of these words is present in the sentence, or 0 in case it is not present. The likelihood ratio λ compares both hypotheses. We determine the first distribution's parameters by the independence hypothesis of both words. The second distribution's parameters are associated with observed joint occurrences of these words within the given window. Hence, λ is driven by the total word count (independence hypothesis) and the words' joint occurrences (cooccurrence hypothesis).

We then transform the likelihood ratio λ to $-2\log(\lambda)$, which is χ^2-distributed. Now, we can use the usual thresholds to determine significance. Evert [7] suggests a different approach to avoid issues with overly extreme probability values. We follow his approach and normalize the measure with the text's word count to avoid biases. We then define the log-likelihood measure as

$$\tilde{\lambda} = \frac{1}{wc}\left(\begin{array}{l} n\cdot\log(n) - n_a\cdot\log(n_a) - n_b\cdot\log(n_b) + n_{ab}\cdot\log(n_{ab}) \\ +(n-n_a-n_b+n_{ab})\cdot\log\left(n-n_a-n_b+n_{ab}\right) \\ +(n_a-n_{ab})\cdot\log(n_a-n_{ab}) + (n_b-n_{ab})\cdot\log(n_b-n_{ab}) \\ -(n-n_a)\cdot\log(n-n_a) - (n-n_b)\cdot\log(n-n_b) \end{array}\right). \tag{2}$$

In order to distinguish between significant cooccurrence and significant non-cooccurrence, we calculate the log-likelihood measure as

$$c_{\text{loglik}}(a,b) = \begin{cases} -2\log(\tilde{\lambda}), & \text{if } n_{ab} < \frac{n_a\cdot n_b}{n_{ab}} \\ 2\log(\tilde{\lambda}), & \text{otherwise.} \end{cases} \tag{3}$$

The dice measure, however, is much simpler and thus, more intuitive than the log-likelihood measure. It is calculated as

$$c_{\text{dice}}(a,b) = \frac{2\cdot n_{ab}}{n_a+n_b}. \tag{4}$$

We determine the cooccurrence list A with respect to the whole data set. The cooccurrence value of each report for a given keyword b and the corresponding cooccurrence list A is then calculated as

$$c_{\text{cooc}}(b) = \sum_{a\in A} c_{\text{dice}}(a,b)\cdot\text{sign}(a), \tag{5}$$

where sign(a) is either $+1$ in cases where a is a positive word concerning the keyword b, or -1 in cases where it is a negative word concerning b. Words that are neither positive nor negative are not considered by $c_{\text{cooc}}(b)$.

Based on our data set, both measures $c_{\text{loglik}}(a,b)$ and $c_{\text{dice}}(a,b)$ deliver nearly the same list of cooccurrence words. Moreover, their cooccurrence values correlate with a factor of 0.969. Therefore, we only employ the dice measure in our analysis. To best of our knowledge, we are the first to measure these cooccurrences. Cheng and Ho [4] study financial reports and look for metaphors for the description of certain events. Nevertheless, their research is based on a metaphor database. Our method is independent of external metaphor data and can easily be transferred to other areas.

4.3 Performance measurements

In order to use the coefficient of determination R^2 to measure the explanatory power of the variance in the RIX. For the evaluation of the estimations, we use the mean absolute error (MAE) and the root mean squared error (RMSE).

These are common methods to measure the performance of estimation models. With RIX and RIX^* denoting the realized and estimated RIX, respectively, and n being the number of observations, we calculate MAE and RMSE according to the following definition

$$\text{MAE} = \frac{1}{n}\sum_{j=1}^{n}|\text{RIX}_j - \text{RIX}^*_j|, \tag{6}$$

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{j=1}^{n}\left(\text{RIX}_j - \text{RIX}^*_j\right)^2}. \tag{7}$$

While RMSE punishes greater deviations between predicted and realized value harder, a low parameter outcome is preferable in both cases.

5 Data set

Our data set comprises the annual risk reports of 30 German banks between 2002 and 2013. We ensure a representative cross section of the German banking sector including the largest banks from each sector in terms of asset size in 2013. There is no German equivalent to the EDGAR-database provided by the SEC in the US. Thus, our data set is hand collected. Most of the reports can be found on the banks's website or requested by the communication departments within the banks. 11 out of 30

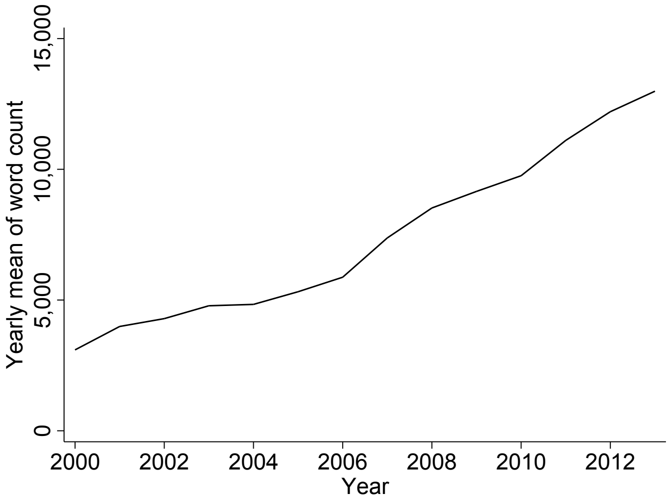

Figure 1. Evolution of the annual reports' average length over time. We average the factor over the available banks.

banks were in distress during the observation period. Employing the risk reports of healthy and distressed banks should increase the robustness of our results. Three banks lack RIX values and, therefore, we exclude their reports from the empirical analysis. Nevertheless, they are included in the following descriptive analysis. Some banks are hard to assign to one of the three sectors, because they have a special business model. Nevertheless, to be in line with the previous work of Schlueter et al. [27], we follow their linkage[2]. The descriptive results using the methods in the precedent chapter are described in table2.

We assign 11 banks to the negative pool due to the following reasons. The first German banking sector contains private banks and banks with a special business model. 6 out of 11 banks from this sector were in distress during our observation period. Dresdner Bank AG was acquired by Commerzbank in 2009. However, Commerzbank itself received massive state aid in the wake of the financial crisis. The same applies to IKB and Hypo Real Estate. Sal Oppenheim had problems with its subsidiaries starting with the financial crisis in 2008. The bank was acquired by Deutsche Bank AG in 2010. Eurohypo or later named Hypothekenbank Frankfurt was founded in 2001 as a merger of three banks focusing on real estate banking. The bank had to change its business model due to the financial crisis in 2008, but was not successful in doing so. Therefore, it was downsized starting in 2012.

The second sector contains cooperative banks. We assign 4 of its 5 banks to the positive and one to the negative pool. There are two cooperative central institutions, DZ Bank and WGZ Bank, and the specialized cooperative bank DG HYP, which focuses on covered bonds. Apobank is the largest cooperative bank in Germany. It faced immense losses during the financial crisis. Therefore, the National Association of German Cooperative Banks established a bailout fund to stabilize the institution. Consequently, we assign Apobank to the negative pool. The fifth cooperative bank is the DVB Bank.

The third banking sector contains public savings, government-owned and federal state banks. We identify 4 distressed banks out of 14. These are BayernLB, HSH Nordbank, Sparkasse KölnBonn and WestLB. All banks suffered intensively from the financial crisis in 2008. Finally, WestLB AG was downsized in 2012. However, high losses of BayernLB, HSH Nordbank, and Sparkasse KölnBonn were absorbed due to the support of their partners. Nevertheless, the banks remained solvent, but are mapped to the negative pool.

Word count

Between 2002 and 2013, the length of the risk reports ranges from two to more than 100 pages. Over time, we observe an increase in the average length of the reports in Figure 1. This increase results from more complex regulatory requirements and a growing public interest in banks' risks and business models. We observe a similar evolution for the number of sentences in a risk report (see Table 13). The number of paragraphs is highly correlated to these two measures and also increases over time.

Cooccurrence analysis

In related literature such as Gomaa and Fahmy [11], the log-likelihood and the dice measure are established instruments for measuring cooccurrences of words in textual data. Naturally, the modified log-likelihood and the dice measure exhibit a high

[2] In fact, we assign KfW as a governement-owned development bank and Haspa to public savings and federal state banks, DVB to the cooperative banks, and Eurohypo to private banks.

Factor	Obs	Mean	Std	Min	Max
wc	343	7,312.000	5,653.000	350.000	53,303.000
pc	343	36.845	24.051	1.000	164.000
dicer	340	0.555	0.272	−0.103	1.488
diceg	340	0.404	0.230	−0.139	1.225
lnsentipa	343	4.525	0.491	2.303	5.617
lnsentina	343	3.300	0.615	1.386	4.682
lnsentips	343	5.096	0.706	2.485	7.245
lnsentins	343	3.900	0.826	1.386	6.423
sentigp	343	−13.262	14.147	−119.138	4.118
rix	309	0.565	0.175	0.073	0.919

Table 2. General descriptive statistics of the relevant text mining factors for RIX estimation. All of the 30 banks are considered. The factors are: word count (wc); paragraph count (pc); cooccurrence value of the keyword "*risk*" measured with the dice measure (dicer) (only positive and negative cooccurrence words are counted); cooccurrence value of the keyword "risk" measured with the dice measure (diceg) (all values are counted); logarithm of number of unique positive words (lnsentipa); logarithm of number of unique negative words (lnsentina); logarithm of sum of positive words (lnsentips); logarithm of sum of negative words (sentips); netted sentiment index (sentigp) which is the total product of the polarized words and their weights; and risk reporting quality index (rix). Obs is the number of observations; Std is the standard deviation and Min and Max are the minimum and maximum values respectively. We report further text mining factors concerning readability, lexicographics, and countings in Appendix 13.

correlation, which amounts to 0.969 in our sample. Therefore, we only use the dice measure for our analysis of the RIX. This measure's value is driven by two factors. The first driver is the frequency of different words that appear in joint occurrence with a given keyword. The second driver influences the dice measure directly. It is these words' weight, i. e. the frequency of their individual occurrence. The frequency of the keyword itself influences the dice measure only implicitly because a higher frequency of the keyword implies more possibilities for the cooccurrence with additional words. We focus on the keywords "*risk*" and "*crisis*". Although we consider all kinds of crises, the term's cooccurrence list is rather short. For the whole data set, it contains 7 negative and four positive words. Due to the small variance in the cooccurrence to "*crisis*", we find no link of this factor to the RIX. Therefore, we only consider the keyword "*risk*".

In our analysis, we solely consider those terms of risk that describe a type of risk, such as credit risk and liquidity risk. Thereby, we consistently consider the relevant risks. Because most types of risk are single words in German, e. g. "*credit risk*" – "*Kreditrisiko*", "*market risk*" – "*Marktpreisrisiko*", and "*liquidity risk*" – "*Liquiditätsrisiko*", we break these terms down into their "*risk*"-components and different risk categories. However, we exclude certain terms and other confounding strings from the analysis, such as "*risk-free interest rate*" or "*risk appetite*". Nevertheless, the list of risk types is quite exhaustive. The banks report 278 different types of risk in total. These can be summarized in 94 risk categories, which is a quite exhaustive list.

We find a slightly higher average cooccurrence value of the word "*risk*" (dicer) for distressed banks compared to non-distressed banks (see Table 9). However, according to an unequal variances t-test testing for significant different mean values, we do not find significant differences. Nevertheless this might be explained by the different use of the term "*risk*". Distressed banks use "*risk*" more frequently in combination with negative words, for instance, "*risk increases*", "*higher risk*", or "*risk*" accompanied by "*enter value adjustments*". On the other hand, the cooccurrence list of distressed banks might be shorter than that of non-distressed banks because the latter use the term "*risk*" less often. This would support our hypothesis that banks try to conceal certain risks or do not even consider them.

When we differentiate the banks according to their type (see Table 10), again, we expect to observe a difference in the RIX between the three banking sectors. This difference might result from the international connectivity of the respective sector. Private and federal state banks are relevant market participants and, consequently, must consider more risks than a cooperative bank, which mainly focuses on credit risks. Nevertheless this significant difference cannot be shown within our data.

We expect our risk measures to have a significant influence on the RIX because an increase in the dice measure originates from a larger number of different types of risk. Banks that report more risk types report their risks to a larger extent and in more detail. This extensive reporting should consequently result in a higher RIX value.

Sentiment analysis

The most frequent words of the *SentiWS* list in our data are "*loss*", "*current*", "*to lead*", and "*appropriate*" (see Table 11). Out of these words, "*loss*" has the largest polarity weight by far. Combined with its extremely high frequency, this word has a particularly strong impact on the sentiment index. Banks often use "*loss*" in terms of the expected and unexpected loss but also to describe their profit and loss situation. Further crucial words are "*fraud*","*danger*", and "*minor*", which often occur in the description of operational risks. These words are used frequently and have a particularly large negative weight. Nevertheless, to prevent bias, we did all of our calculations without the terms "*loss*" , "*fraud*", "*minor*", and "*danger*". Results stayed the same, so these terms are still included in the analysis.

Concerning the sentiment index, we find out that it remains negative over time. This outcome is rather unexpected because there are more positive than negative words in the reports on average. The reason is that the detected negative words have a significantly larger polarity weight than the positive words. However, these weights cannot be compensated with the frequency of the positive words in their linear combination. Consequently, the netted sentiment index exhibits negative values. As mentioned above, we replicate the same calculations, removing the keyword "*loss*", "*fraud*", "*minor*", and "*danger*" from the analyzed text and all results remain the same.

Based on our sample, we find a significant connection between the sentiment and the quality of a bank's risk report. Our results indicate that in particular additional positive words reduce a reports quality on average.

This finding indicates that banks not only repeate but also complement their quantitative data and associated risk situation with qualitative arguments. Unlike with the quantitative data, text might gloss over potential financial difficulties.

Readability

Readability measures aim to differentiate between degrees of difficulties in a text. They are one way to quantify a text's complexity. The Gunning-Fog index (fog) is a combination of the average sentence length and the number of complex words. The algorithm considers words with three or more syllables to be complex. The index is computed as

$$\text{fog} = 0.4 \cdot \left((\text{average number of words per sentence}) + 100 \cdot \left(\frac{\text{\# complex words}}{\text{\# words}} \right) \right). \tag{8}$$

Loughran and McDonald [23] find that the readability of financial text is measured with fog in most cases. We consider two more popular readability measures. The Flesch-Kincaid index is named after its creators and is calculated as

$$fk_{\text{grade}} = 0.39 \left(\frac{\text{\# words}}{\text{\# sentences}} \right) + 11.8 \left(\frac{\text{\# syllables}}{\text{\# words}} \right) - 15.59. \tag{9}$$

It determines a text's difficulty measured in U.S. grade levels. We also consider the Forcast index. It mainly focuses on functional literacy, such as annual reports. It is based on the examination of samples of 150 words and averages these to the final index. The index is calculated as

$$\text{forc} = \frac{1}{n} \sum_{i=1}^{n} 20 - \frac{N_i}{10}, \tag{10}$$

with N denoting the number of words with one syllable in sample i and n the number of samples.

All three readability measures account for the number of words and syllables of different text windows, e. g. sentences or the whole text. Therefore, theoretically, the measure should be independent of the text's language to some extent. Each index returns a score that translates to a certain school grade according to the text's complexity.

The risk reports' readability outcomes are as high as expected (see Table 12). As expected, the values are large. Loughran and McDonald [23] determine an average Gunning-Fog index of 18.69 in their analysis of 10-K forms. They argue that text with values larger than 18 are mostly unreadable. We find values as high as 22.54 on average. The Flesh-Kincaid and Forcast indices also return values far above average. According to the results of our readability analysis, readers can hardly understand risk reports. This is certainly not the case. Many financial terms, in particular, in German comprise several syllables but are rather easy to understand. For instance

- "*risk management*" – "*Risikomanagement*" – 6 syllables,
- "*credit default risk*" – "*Kreditausfallrisiko*" – 7 syllables, and
- "*overall risk-bearing capacity*" – "*Gesamtrisikotragfähigkeit*" – 9 syllables.

Naturally, supposedly complex words such as these appear frequently in risk reports. However, they increase the readability measures greatly.

Over time, the measures remain almost constant. Similar to Loughran and McDonald [23], we do not find a link between these measures and the report's quality. We attribute this outcome to the small variance in their values over time and, therefore, we do not analyze them further. To put it in a nutshell, we should say that our results of readability does not provide true new insights. Nevertheless our aim is to connect classical text mining studies and the finance area. Therefore, the analysis and the descriptions of these readability measures is not excludable.

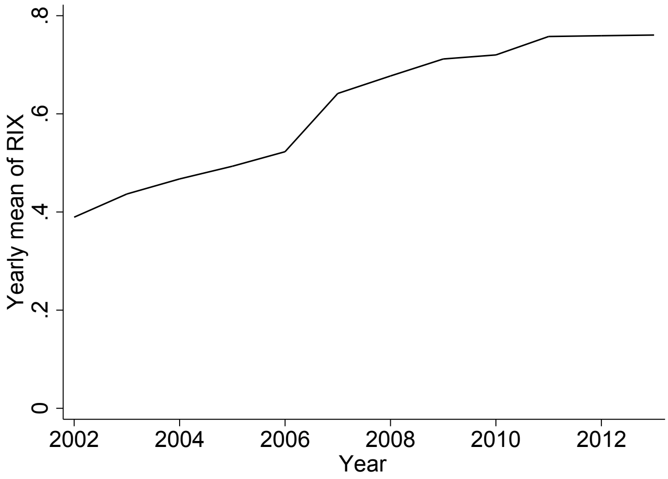

Figure 2. Evolution of the average risk reporting quality of the annual reports over time. We average the factor over the available banks.

Risk reporting quality

The measurement of risk reporting quality using the RIX starts in 2002. We see in Figure 2 that the average risk reporting quality increases monotonously. The average RIX value doubles within 10 years. However, its deviation is rather small. In 2007, we observe an increase in the RIX value that is significantly above its average growth rate. We attribute this jump to the introduction of IFRS 7. With this new regulatory approach, banks using capital market-based funding must fulfill several additional reporting requirements. These are for instance, a description of the bank's risk management structure and the extended explanation of its risk measurements, particularly, of credit, market, and liquidity risks, and the report of risks concentration. At the same time, the Basel II Accord was implemented into German national laws, i. e. the *Solvabilitätsverordnung* (SolvV) and the *Mindestanforderungen an das Risikomanagement* (MaRisk).

Naturally, the data set is incomplete because e. g. WestLB was downsized in June 2012 and Portigon became its legal successor. However, Portigon's business model and its risk management are completely different from those of WestLB. Hence, Portigon achieves RIX values that are far from those of WestLB. Therefore, we cut off the data of WestLB in 2012.

In summary, the banks' reports differ concerning their determined text mining factors when we pool the banks according to their type or performance. However, some of these differences are rather small, particularly, when it comes to cooccurrence measures. It is in the bank's best interest to produce a satisfying external risk report and to enrich the quantitative data with explanatory text. A data-driven classification of the banks, e. g. according to their reports' sentiment or the cooccurrence index, would be inappropriate. Therefore, we distinguish the employed banks by their level of distress which we define manually. Thus, the banks are either distressed or non-distressed.

6 Results

Our list of determined factors is quite exhaustive. These factors quantify many aspects of the underlying text. However, if we group these factors, for instance, according to counting measurements, e. g. counting words, paragraphs, or even letters, naturally the factors' correlations are relatively high. With more text, it is very likely that there are more objects of any kind, such as letters, paragraphs, punctuations, and nouns. In order to extract unbiased influences of these factors on the RIX, we reduce the number of influencing factors to an eligible sample and standardize counting measures by the respective total word count. The reduction is also necessary in order to account for the moderate number of 309 observations from 27 banks covering the period from 2002 to 2013.

To analyze rudimentary relationships between the calculated measures, we provide a comprehensive covariance matrix in Table 3. We see in Figure 2 that the average RIX value and the word count are growing monotonously. They exhibit the highest correlation. Consequently, RIX and number of paragraphs are also positively correlated. Naturally, the number of paragraphs is positively as well as highly associated with the word count. Concerning the RIX, only the proportion of positive words (sentipswc) and the cooccurrence of risky words (dicegwc) have negative correlations. In case of sentipswc, this indicates that a larger proportion of positive words reduces the quality of the report. Thus, additional positive words might be an instrument to trivialize risks in general. The negative correlation of dicegwc with rix supports this hypothesis. An increase in the dice measure might have several reasons. However, it is not only influenced by a higher frequency of the keyword, i. e. "*risk*". Increasing

	rix	lnwc	lnpc	sentipswc	sentinswc
lnwc	0.819				
lnpc	0.709	0.872			
sentiposwc	−0.426	−0.281	−0.194		
sentinegswc	0.159	0.109	0.003	−0.124	
dicegwc	−0.356	0.425	0.399	0.005	−0.078

Table 3. Covariance matrix of selected factors. rix is the RIX; lnwc is the logarithm of word count; lnpc is the logarithm of the number of paragraphs; sentiposwc and sentinegswc are the ratios of positive and negative words to word count; dicegwc is the ratio of the overall dice measure to word count.

usage of the keyword tends to extend the list of cooccurring words. In our data, we find more positive than negative words cooccurring with the keyword "*risk*". The correlation of sentipswc with rix is negative because these positive words tend to decrease information quality and, hence, lower the report's RIX value. In order to improve the results in terms of estimation accuracy and interpretability, we control for several general and idiosyncratic factors. In Figure 1 we see an increasing average number of words per report. To control for this endogenous effect in the regressions, we add dummy variables for each year of the observation period.

6.1 Regression and estimation of RIX

To begin with, we determine factors which have a significant influence on the RIX. We do this by multivariate linear regression with ordinary least squares. As indicated before, we establish a consistent set of factors that comprises lnwc, lnpc, sentiposwc, sentinegwc, and dicegwc. To control for differences arising from the particular reporting year, we add year dummies. Additionally, we control for the introduction of GAS-15 and GAS-20 reporting standards and also for the banks' total assets.

In order to calibrate an adequate model, we regress RIX_t on the found factors in a panel regression

$$\mathrm{RIX}_{it} = \alpha_i + \sum_{j=1}^{n} \beta_j \mathrm{VAR}_{j,it} + u_{it}, \tag{11}$$

where α is the intercept of group i, β_j is the slope coefficient of the variable VAR_j in period t with j the number of variables, $u_{i,t}$ is the error term, and n is the number of independent variables.

Table 4 reports the regression results. First, we use the simple counting measures word count and number of paragraphs and control for our set of controls. The word count has a significant and positive coefficient. With a value of 76.2%, R^2 is already quite large.

We find R^2 worsening when we extend the regression to the qualitative measures, i. e. the cooccurrence of risky words and the ratio of positive and negative words to the report's word count. In contrast to other factors, the proportion of positive words (sentiposwc) is negatively associated with the RIX. This finding corresponds to our hypothesis that an extensive use of positive words reduces the reporting quality. Negative words (sentinegwc) on the other hand tend to increase RIX. Both sentiment measures are significant when not accounting for the set of controls. Surprisingly, the cooccurrence with the word "*risk*" does not significantly influence RIX.

Regression 4.3 accounts for all these factors and additionally controls for year, reporting standard, and bank specifics. These controls put the influence of most factors into perspective. Thus, negative words lose their significance. Now the number of paragraphs becomes significant but has only a small coefficient. Again, positive words have a particularly negative effect on the RIX. The word count maintains its highly significant influence. However, compared to Regression 4.1, we obtain a slight decrease in the overall R^2. In Regression 4.3, it is as high as 75.2%. Nonetheless, the model explains variances in the RIX to a very large extent.

The final regression extends the analysis to the lagged dependent variable rix_{t-1}. Now, we can examine the influences of our text mining factors more clearly. Naturally, rix_{t-1} is highly significant and has a strong positive influence on the RIX. However, the word count remains significant and adds to the regression's explanatory power. The increase in R^2 to 93.0% mostly results from the notable increase in the coefficient of determination between the panels.

One of our major goals is to assess the quality and extent of banks' risk reports. Because close reading is very expensive in terms of invested time and personnel and is also subject to subjectivity and prone to error, we implement an automated procedure to determine the reports' quality. Using Regressions 4.3 and 4.4, which explain much of the variance in the RIX, we estimate the RIX for each single risk report. We report the estimation's mean absolute error (MAE) and root mean squared error (RMSE) at the bottom of Table 4. These errors are particularly low. On average, we miss the realized RIX value by only 7.1 percentage points. This sturdy estimation becomes obvious in Figure 3. Here only a few estimates are far from their realized value. However, these

Reg	4.1	4.2	4.3	4.4
R^2				
Overall	0.762	0.724	0.752	0.930
Within	0.855	0.731	0.861	0.915
Between	0.724	0.730	0.697	0.943
lnwc	0.0666**	0.1789***	0.0761***	0.0407**
lnpc	0.0248	0.0280	0.0099*	-0.0031
sentiposwc		-8.3748***	-4.6263**	-1.8538
sentinegwc		9.6373***	-1.3458	0.4820
dicegwc		0.1054	0.0269	0.0322
rix_{t-1}				0.6235***
Controls				
Year	yes	no	yes	yes
Accounting	yes	no	yes	yes
Balance sheet	yes	no	yes	yes
Bank fixed effects	yes	yes	yes	yes
MAE			0.0713	0.0338
RMSE			0.0872	0.0436

Table 4. Panel regression of RIX on text mining factors with fixed effects and clustered standard errors. The regressions are based on 309 observations of 27 banks. Regressions 4.1–4.4 report the regression results with different factors. R^2 is the coefficient of determination. It is determined for the full set of observations, within each panel, i. e. each bank, and between the panels. The factors are: logarithm of word count (lnwc); logarithm of the number of paragraphs (lnpc); ratio of the sum of positive words to word count (sentiposwc); ratio of the sum of negative words to word count (sentinegwc); ratio of the overall dice measure to word count (dicegwc); and lagged RIX (rix_{t-1}). The t-statistics are based on clustered standard errors. The coefficients' two-tail test displays their significance at levels of 10% (*), 5% (**), and 1% (***). As controls, we add dummy variables (Year) for each year of the observation period. We also control for the introduction of new accounting standards (Accounting). In addition we control for the banks' total balance sheet (Balance sheet) and bank fixed effects (Bank fixed effects). MAE is the mean absolute error and RMSE is the root mean squared error of the RIX prediction, according to Equations (6) and (7), respectively.

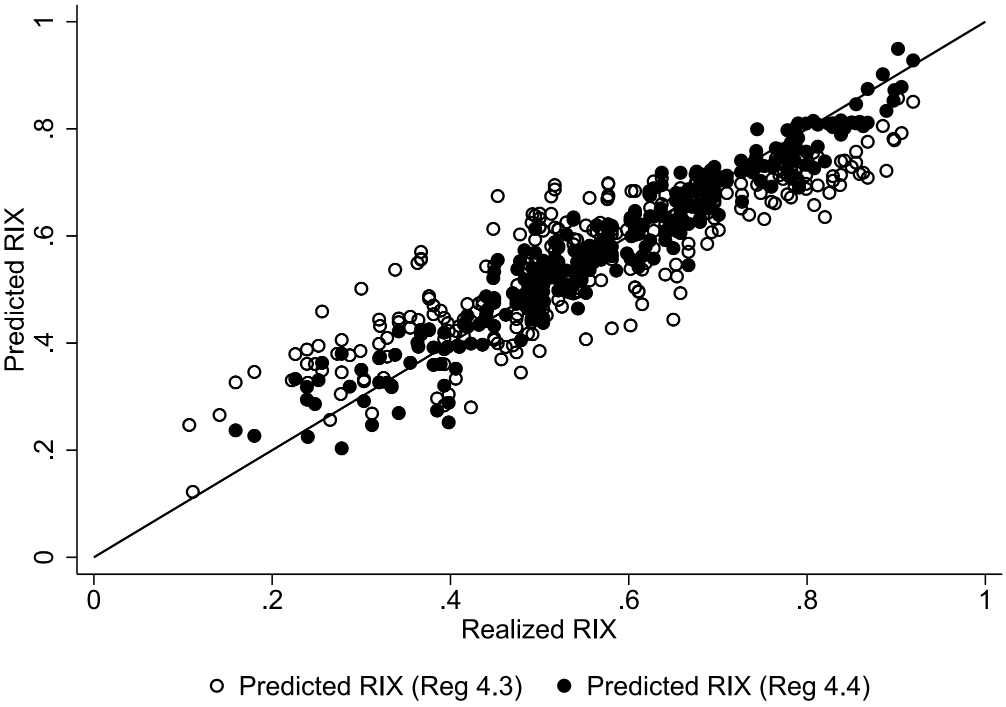

Figure 3. Scatter plot of realized and predicted RIX. Hollow circles mark the predictions according to Regression 4.3, solid circles are calculated according to Regression 4.4 of Table 4. A simple diagonal line marks the perfect prediction.

few estimates are particularly punished by the RMSE, which averages at 0.087. When we estimate the RIX with Regression 4.4, which incorporates the lagged RIX, naturally, the error decreases. In particular, the RMSE drops to 0.044. Again, this reduction becomes quite clear in Figure 3. Now, the estimates are even closer to the diagonal, which indicates the exceptional accuracy of the estimation.

Although the estimation can obviously be improved by adding the one year lag of the RIX, we refrain from using this factor for the prediction of unseen reports. The reason is that the RIX is determined by close reading of the employed risk reports. However, our objective is to calibrate a model that is based on automatically determined algorithmic factors. Therefore, incorporating RIX_{t-1} would make the model dependent on this hand collected factor because an iterative method would have to start with a RIX determined by close reading. On the other hand, the estimation error of the lagged RIX would propagate in subsequent periods and, hence, result in biased estimates. Moreover, an estimation model without the lagged RIX allows for the estimation of the RIX of the reports of banks that were not yet analyzed.

6.2 Regression of RIX by type of bank

The German banking industry builds upon three sectors as described in Section 5. Covering these three with a representative sample, we can analyze each sector separately. Furthermore, we expect different drivers of the RIX for the different types of banks due to their business models and the related different risk taking profiles. Potential investors and shareholders pay particular attention to private banks' risk reports. Cooperative and public savings banks act in colligated solutions and do not use capital market-based funding. Therefore, less attention is paid to their risk reports by potential investors. The recipients of their annual reports are customers or regulators. According to Fischer et al. [9], some of the federal state banks changed their business models during the observation period. This change might result in a change of the drivers of the RIX during the observation period. We expect the variables word count, sentiposwc, and sentinegwc to have a larger impact on the reporting quality of private and federal state banks than on the reporting quality of cooperative banks.

Table 5 reports the regression outcomes for the three bank types. Regressions 5.1 and 5.3 exhibit the highest R^2, when disregarding the lagged RIX. Therefore, we focus on their factors in the following only. Regression 5.1 of private banks in Table 5 is similar to its counterpart of all banks in Table 4. lnwc is significantly and positively associated with the RIX. However, the overall R^2 decreases to 72.7% and indicates that private banks are less homogeneous in terms of the used factors than the set

	Private		Cooperative		Public	
Reg	5.1	5.2	5.3	5.4	5.5	5.6
Obs	101	101	48	48	160	160
Banks	9	9	4	4	14	14
R^2	0.727	0.667	0.854	0.890	0.782	0.813
lnwc	0.0883**	0.0895**	0.0192	0.0149	0.0659*	0.0673**
lnpc	0.0131	0.0062	0.0576	0.0765**	0.0000	−0.0181
sentiposwc		−2.1531		−7.6622*		−9.0999***
sentinegwc		−10.8275		−0.2445		0.0226
dicegwc		−0.1001		0.4602*		−0.0345
Controls						
Year	yes	yes	yes	yes	yes	yes
Accounting	yes	yes	yes	yes	yes	yes
Balance sheet	yes	yes	yes	yes	yes	yes
Bank fixed effects	yes	yes	yes	yes	yes	yes

Table 5. Panel regression of RIX on text mining factors per type of bank with fixed effects and clustered standard errors. The types are: private banks and banks with a special business model (Private); cooperative banks (Cooperative); public savings and federal state banks (Public). Regressions 5.1–5.6 report the results of the regression with different factors. R^2 is the coefficient of overall determination. The factors are: logarithm of word count (lnwc); logarithm of the number of paragraphs (lnpc); ratio of the sum of positive words to word count (sentiposwc); ratio of the sum of negative words to word count (sentinegwc); and ratio of the overall dice measure to word count (dicegwc). The t-statistics are based on clustered standard errors. The coefficients' two-tail test displays their significance at levels of 10% (*), 5% (**), and 1% (***). As controls, we add dummy variables (Year) for each year of the observation period. We also control for the introduction of new accounting standards (Accounting). In addition we control for the banks' total balance sheet (Balance sheet) and bank fixed effects (Bank fixed effects).

	Positive		Negative	
Reg	6.1	6.2	6.3	6.4
Obs	199	199	110	110
Banks	17	17	10	10
R^2	0.681	0.617	0.616	0.585
lnwc	0.0424	0.0497*	0.1042**	0.1123**
lnpc	0.0174	−0.0081	0.0244	0.0204
sentiposwc		−8.6028***		−3.1832
sentinegwc		−0.8537		−1.7976
dicegwc		0.1479		−0.2918
Controls				
Year	yes	yes	yes	yes
Accounting	yes	yes	yes	yes
Balance sheet	yes	yes	yes	yes
Bank fixed effects	yes	yes	yes	yes

Table 6. Panel regression of RIX on text mining factors pooled by the banks' distress level with fixed effects and clustered standard errors. Regressions 6.1–6.4 report the results of the regression with different factors. R^2 is the coefficient of determination. It is determined for the full set of observations, within each panel, i. e. each bank, and between the panels. The factors are: logarithm of word count (lnwc); logarithm of the number of paragraphs (lnpc); and ratio of the sum of positive words to word count (sentiposwc); ratio of the sum of negative words to word count (sentinegwc); ratio of the overall dice measure to word count (dicegwc). The t-statistics are based on clustered standard errors. The coefficients' two-tail test displays their significance at levels of 10% (*), 5% (**), and 1% (***). As controls, we add dummy variables (Year) for each year of the observation period. We also control for the introduction of new accounting standards (Accounting). In addition we control for the banks' total balance sheet (Balance sheet) and bank fixed effects (Bank fixed effects).

of all banks. Adding the qualitative text mining factors in Regression 5.2, we obtain a similar result with the text mining factors being insignificant.

The sample of cooperative banks consists of 4 banks and 48 reports in total. Its results should be treated with caution. Presumably according to the high homogeneity of this group's reports, we observe the highest R^2 of all bank types.

Finally, the reports of federal state and public savings banks are significantly influenced by the word count in Regression 5.5. Its coefficient is in line with that of Table 4. Similar to Regression 5.2, including the qualitative factors in Regression 5.6 results in sentiposwc also being significantly different from zero. Its sign is as expected, however, the coefficient differs notably from the regression with all banks. Thus, public banks seem to be particularly negatively affected by an increasing proportion of positive words. Regression 5.4 indicates that the number of paragraphs and the dice measure have a positive and significant effect on RIX. Also, sentiposwc becomes significant with a negative coefficient which is in line with the regressions in Table 4. R^2 is remarkably high with 81.3%.

6.3 Regression of RIX by distressed and non-distressed banks

For our final analysis of the influencing factors of the RIX, we pool the data into a positive and negative pool with distressed and non-distressed banks respectively in accordance with Section 5. Again, our hypothesis is that the RIX of both pools is driven by different factors. To begin with, we find medium high R^2 for both regressions with non-distressed banks. Neither the word count nor the number of paragraphs have a significant influence on RIX in Regression 6.1 of Table 6. In Regression 6.2, the prominent significance of the proportion of positive words stands out. Moreover, the coefficient is particularly large and comparable to the regressions of cooperative and public banks in Table 5. In contrast to Table 5 the large coefficient of the proportion of negative words mostly compensates its effect.

The factors of the negative pool behave similar to those of the positive pool. However, only word count seems to be driving RIX. R^2 decreases in both regressions to 61.6% and 58.5% for the regressions without and with qualitative factors, respectively. Both regressions indicate that additional words increase the RIX significantly. Compared to Regression 6.1, the coefficient of word count almost doubles. Hence, distressed banks can enhance their RIX value by reporting risks more extensively.

6.4 Regression and prediction of future RIX

We see a large potential for text mining measures to predict future RIX values. This prediction takes the current year's report into account and estimates the RIX of the following year. This procedure is reasonable because the reports contain a special section for the banks forecast of the upcoming fiscal year. Moreover, Figure 2 shows that RIX tends to increase continuously

Reg	7.1	7.2	7.3	7.4
R^2				
Overall	0.627	0.594	0.633	0.817
Within	0.698	0.524	0.700	0.753
Between	0.623	0.682	0.594	0.897
lnwc_t	0.0738***	0.1581***	0.0846***	0.0435*
lnpc_t	-0.0094	0.0105	-0.0131	-0.0240*
$\mathrm{sentiposwc}_t$		-5.5909**	-1.9634	0.4951
$\mathrm{sentinegwc}_t$		9.9631**	2.2388	3.0328
$\mathrm{dicegwc}_t$		0.0536	-0.1490	-0.2155
rix_t				0.5936***
Controls				
Year	yes	no	yes	yes
Accounting	yes	no	yes	yes
Balance sheet	yes	no	yes	yes
Bank fixed effects	yes	yes	yes	yes
MAE			0.0895	0.0583
RMSE			0.1114	0.0781

Table 7. Panel regression of RIX_{t+1} on text mining factors with fixed effects and clustered standard errors. The regressions are based on 281 observations of 27 banks. Regressions 7.1–7.4 report the results of the regression with different factors from the current period t. R^2 is the coefficient of determination. It is determined for the full set of observations, within each panel, i. e. each bank, and between the panels. The factors are: logarithm of word count (lnwc_t); logarithm of the number of paragraphs (lnpc_t); ratio of the sum of positive words to word count ($\mathrm{sentiposwc}_t$); ratio of the sum of negative words to word count ($\mathrm{sentinegwc}_t$); ratio of the overall dice measure to word count ($\mathrm{dicegwc}_t$); and the realized RIX value (rix_t). The t-statistics are based on clustered standard errors. The coefficients' two-tail test displays their significance at levels of 10% (*), 5% (**), and 1% (***). As controls, we add dummy variables (Year) for each year of the observation period. We also control for the introduction of new accounting standards (Accounting). In addition we control for the banks' total balance sheet (Balance sheet) and bank fixed effects (Bank fixed effects). MAE is the mean absolute error and RMSE is the root mean squared error of RIX_{t+1} prediction, according to Equations (6) and (7), respectively.

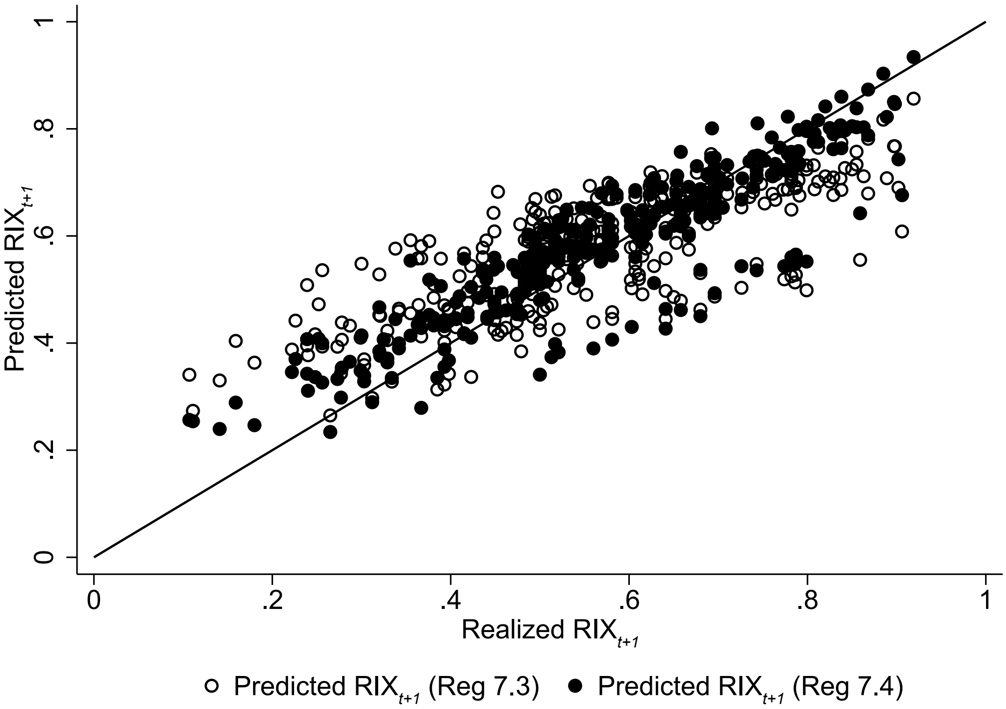

Figure 4. Scatter plot of realized and predicted RIX_{t+1}. Hollow circles mark the predictions according to Regression 7.3, solid circles are calculated according to Regression 7.4 of Table 7. A simple diagonal line marks the perfect prediction.

over time, therefore the RIX value of the previous period and its drivers should deliver accurate predictions of future RIX values. We employ the model from Equation (11) and replace RIX_{it} with RIX_{t+1}.

We report the regression results in Table 7. Naturally, as we are forecasting future RIX, the overall R^2 reduces to a lower level compared to previous regressions. Regressions 7.1–7.3 mostly achieve higher coefficients of determination when we incorporate additional factors. The number of words is highly significant in each case. Particularly in Regression 7.3, we find that after we control for the year and the reporting standard, only the word count remains significant. Here, R^2 reaches its peak at 63.3%. All signs are retained.

Adding the current period's RIX_t to the regression is very beneficial to the overall R^2. It rises to 81.7%. Still, along with RIX_t, also the counting measures for word and paragraph count retain a significant influence on RIX_{t+1}.

Using Regression 7.3, we predict RIX_{t+1}. The MAE amounts to 0.090, which refers to an average deviation of the same amount from the realized RIX, i. e. 9% points. As before, RMSE penalizes large deviations particularly strongly, it amounts to 0.111. When we predict RIX_{t+1} according to Regression 7.4, the prediction error even drops to about half of the amount of that in Regression 7.3. Estimates produced by Regression 7.4 exhibit MAE and RMSE values of 0.058 and 0.078, respectively. Concerning the high coefficient of determination, RIX_t explains most of the variation of RIX_{t+1}. Moreover, its incorporation increases the estimation accuracy significantly.

Figure 4 visualizes the goodness of fit outlined in a scatter plot. As the employed measures MAE and RMSE already indicate, the deviation of the predicted from their realized values is rather small on average. The prediction from Regression 7.3 in Figure 4 is already very accurate. Most estimates are close to their realized RIX values. As expected, Regression 7.4 produces even more accurate predictions.

7 Conclusion

A bank's annual risk report is a comprehensive public source for information on all issues related to the bank's risk. The report is of particular interest to the bank's stakeholders and its customers. It is regulated to some extent by prescribing the required topics on which a bank must file a statement such as its credit risks and the risk management in place. Other topics are less standardized and provide room for interpretation and forecasts.

Using the risk reporting index (RIX) introduced by Schlueter et al. [27], we analyze the risk reports of 30 major German banks over the period of 2002–2013. We find that there are two essential ways to assess the quality and quantity of the information provided in the risk report. The first of which is close reading of each single report. However, this is an increasingly time-consuming matter because the reports' length on average increases by a total of 300% from a bank's first to its most current risk report of the observation period. Also, the assessment result naturally depends on the individual analysts, their mistakes, and their degree of subjectivity concerning the scoring. The second way to assess risk reports is the automated analysis of the reports using text mining techniques. This procedure overcomes most of the shortcomings of the close reading process. On the one hand, it assesses the reports in a very short time. On the other hand, depending on the model's calibration, it is also error-free and objective to a large extent.

Our descriptive analysis indicates differences between the determined text mining factors of distressed and non-distressed banks. These differences arise from the banks' use of different expressions and vocabulary. Moreover, banks using capital market-based funding and local banks differ particularly in terms of their report's complexity, vocabulary, and sentiments.

We find that our panel regression model which employs the determined text mining measures can reproduce the close reading RIX values very accurately. The relevant factors have a robust and significant influence on the RIX and explain up to 90% of its variance. As expected, the number of words per report is a particularly important factor for determining the RIX. The RIX value rises with an increasing report length, indicating that a long report provides more information on the bank's risk than a shorter one. Moreover, we find that the positive sentiment of the text is negatively associated with the RIX. Hence, the reporting of risks with too many and at the same time, overly positive words reduces the RIX value. We conclude that this outcome indicates the bank's attempt to deflect certain disadvantageous developments in its risk situation and therefore gives reason to deeper analysis of such reports.

Using the calibrated model of this study, new and also historic risk reports of all German banks can be assessed automatically. However, further work could be done on precisely reproduce the score system underlying the RIX.

8 Acknowledgements

The author thanks the Institute for Bank Management and Banking Law at the University of Cologne for the financial support.

REFERENCES

[1] Antweiler, W., Frank, M.Z., 2004. Is All That Talk Just Noise? The Information Content of Internet Stock Message Boards. The Journal of Finance 59, 1259–1294.

[2] Basel Committee on Banking Supervision, 2013. Principles for effective risk data aggregation and risk reporting. Bank for International Settlements.

[3] Campbell, D., Rattanataipop, P., 2014. Risk reporting by UK banks, 1995-2010: an exercicse in futility? Unpublished.

[4] Cheng, W., Ho, J., 2015. A Corpus Study of Bank Financial Analyst Reports. International Journal of Business Communication , 1–25.

[5] Deutsche Bundesbank, 2014. Bankenstatistik Dezember 2014, Statistisches Beiheft 1 zum Monatsbericht.

[6] Esuli, A., Sebastiani, F., 2006. SentiWordNet: A Publicly Available Lexical Resource for Opinion Mining, in: In Proceedings of the 5th Conference on Language Resources and Evaluation (LRECï¿½06), 417–422.

[7] Evert, S., 2004. The Statistics of Word Cooccurrences: Word Pairs and Collocations. Ph.D. thesis. University of Stuttgart.

[8] Feldman, R., Govindaraj, S., Livnat, J., Segal, B., 2010. Managementï¿½s tone change, post earnings announcement drift and accruals. Review of Accounting Studies 15 4, 915–953.

[9] Fischer, M., Hainz, C., Rocholl, J., Steffen, S., 2014. Government Guarantees and Bank Risk Taking Incentives. ESMT Research Working Papers ESMT-14-02. ESMT European School of Management and Technology.

[10] Frazier, K.B., Ingram, R.W., Tennyson, B.M., 1984. A Methodology for the Analysis of Narrative Accounting Disclosures. Journal of Accounting Research 22, pp. 318–331.

[11] Gomaa, W.H., Fahmy, A.A., 2013. A Survey of Text Similarity Approaches. International Journal of Computer Applications 68, pp. 1–6.

[12] Hartmann, W., 2014. Risk transperency of large banks in the U.S. and Europe, in: Yearbook 2014. The Frankfurt Institute for Risk Management and Regulation, Frankfurt am Main, 14–19.

[13] Hope, O.K., Hu, D., Lu, H., 2016. The benefits of specific risk-factor disclosures. Review of Accounting Studies 21, 1005–1045.

[14] Huang, X., Hong Teoh, S., Zhang, Y., 2014. Tone Management. The Accounting Review 89, 1083–1113.

[15] Jegadeesh, N., Wu, D., 2013. Word power: A new approach for content analysis. Journal of Financial Economics 110, 712–729.

[16] Kennedy, A., Inkpen, D., 2006. Sentiment Classification of Movie Reviews Using Contextual Valence Shifters. Computational Intelligence 22, 2006.

[17] Kloptchenko, A., Magnusson, C., Back, B., Visa, A., Vaharanta, H., 2004. Mining Textual Contents of Financial Reports. The International Journal of Digital Accounting Research 4, 1–29.

[18] Kohut, G.F., Segars, A.H., 1992. The President's Letter to Stockholders: An Examination of Corporate Communication Strategy. International Journal of Business Communication 29, 7–21.

[19] Kravet, T., Muslu, V., 2013. Textual Risk Disclosures and Investors' Risk Perceptions. Review of Accounting Studies 18, 1088–1122.

[20] Li, F., 2006. Do Stock Market Investors Understand the Risk Sentiment of Corporate Annual Reports? Working paper.

[21] Li, F., 2010. The Information Content of Forward-Looking Statements in Corporate Filings–A Naïve Bayesian Machine Learning Approach. Journal of Accounting Research 48, 1049–1102.

[22] Loughran, T., McDonald, B., 2011. When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks. The Journal of Finance 66, 35–65.

[23] Loughran, T., McDonald, B., 2014. Measuring Readability in Financial Disclosures. The Journal of Finance 69, 1643–1671.

[24] Loughran, T., McDonald, B., 2015. The Use of Word Lists in Textual Analysis. Journal of Behavioral Finance 16, 1–11.

[25] Pang, B., Lee, L., Vaithyanathan, S. (Eds.), 2002. Thumbs up? Sentiment Classification using Machine Learning Techniques, Proceedings of EMNLP.

[26] Remus, R., Quasthoff, U., Heyer, G., 2010. SentiWS – A Publicly Available German-language Resource for Sentiment Analysis, in: Proceedings of the 7th International Language Resources and Evaluation (LREC'10), 1168–1171.

[27] Schlueter, T., Hartmann-Wendels, T., Weber, T., Zander, M., 2014. Die Risikoberichterstattung deutscher Banken: Erhebung des Branchenstandards. zfbf – Schmalenbachs Zeitschrift fuer betriebswirtschaftliche Forschung 5-6, 386–427.

[28] Shirata, C.Y., Sakagami, M., 2009. An Analysis of the "Going Concern Assumption": Text Mining from Japanese Financial Reports. The Journal of Emerging Technologies in Accounting 5, 1–16.

[29] Shirata, C.Y., Takeuchi, H., Ogino, S., Watanabe, H., 2011. Extracting Key Phrases as Predictors of Corporate Bankruptcy: Empirical Analysis of Annual Reports by Text Mining. Journal of Emerging Technologies in Accounting 8, 31–44.

[30] Strapparava, C., Valitutti, A., 2004. WordNet-Affect: An Affective Extension of WordNet, in: Proceedings of LREC, 1083–1086.

[31] Tetlock, P.C., Saar-Tsechansky, M., MacKassy, S., 2008. More Than Words: Quantifying Language to Measure Firms' Fundamentals. The Journal of Finance 63, 1437–1467.

[32] Wiebe, J., Wilson, T., Bell, M., 2001. Identifying collocations for recognizing opinions, in: Proceedings of the ACL-01 Workshop on Collocation: Computational Extraction, Analysis, and Exploitation, Toulouse, France. 24–31.

[33] You, H., Zhang, X.j., 2009. Financial reporting complexity and investor underreaction to 10-k information. Review of Accounting Studies 14, 559–586.

Appendix A

Data set

Bank	Type of bank	Total assets in 2013 in € bn	Pool	RIX
Deutsche Bank	private	1,395	pos	yes
Commerzbank	private	592	neg	yes
KfW	public	465	pos	yes
LBBW	public	286	pos	yes
DZ Bank	cooperative	223	pos	yes
BayernLB	public	206	neg	yes
Helaba	public	157	pos	yes
NRW.Bank	public	146	pos	yes
Postbank	private	143	pos	yes
NordLB	public	129	pos	yes
HSH Nordbank	public	120	neg	yes
Deka Bank	public	117	pos	yes
Hypothekenbank Frankfurt /Eurohypo	private	101	neg	yes
LBBH	public	71	pos	yes
L-Bank	public	71	pos	no
Hypo Real Estate	private	68	neg	yes
DKB	private	67	pos	yes
WGZ Bank	cooperative	51	pos	yes
DG HYP	cooperative	49	pos	no
Aareal Bank	private	42	pos	yes
Haspa	public	41	pos	yes
ApoBank	cooperative	34	neg	yes
SEB	private	30	pos	yes
Sparkasse KölnBonn	public	29	neg	yes
IKB	private	26	neg	yes
Kreissparkasse Köln	public	24	pos	yes
DVB	cooperative	23	pos	yes
Dresdner Bank	private	–	neg	no
Sal Oppenheim	private	–	neg	yes
WestLB	public	–	neg	yes

Table 8. List of banks and their key characteristics. The type of bank differentiates three kinds of banks: private banks and banks with a special business model (private); cooperative banks (cooperative); and public savings and federal state banks (public). "–" indicates that no balance sheet data is available as of 2013 because the bank was either downsized or taken over. We classify the banks into pools of distressed (neg) and non-distressed (pos) banks. For some banks, no RIX-values are available, which is indicated in the last column (RIX).

	Positive pool					Negative pool				
Factor	Obs	Mean	Std	Min	Max	Obs	Mean	Std	Min	Max
wc	225	7,486.422	6,298.596	350.000	53,303.000	118	6,978.678	4,158.499	1,225.000	19,400.000
pc	225	39.331	26.033	1.000	164.000	118	34.008	19.518	7.028	88.000
loglikr	222	5.182	1.934	0.913	11.513	118	5.254	1.744	1.029	9.086
dicer	222	0.543	0.263	−0.095	1.353	118	0.577	0.290	−0.103	1.488
diceg	222	0.410	0.233	−0.099	1.225	118	0.393	0.225	−0.139	1.019
lnsentipa	225	4.516	0.517	2.303	5.616	118	4.542	0.437	3.401	5.278
lnsentina	225	3.276	0.620	1.609	4.682	118	3.346	0.603	1.387	4.407
lnsentips	225	5.083	0.746	2.485	7.245	118	5.120	0.623	3.610	6.274
lnsentins	225	3.892	0.837	1.609	6.423	118	3.915	0.807	1.387	5.393
sentigp	225	−13.510	15.622	−119.137	4.118	118	−12.790	10.835	−50.841	0.824
rix	199	0.565	0.191	0.073	0.919	110	0.566	0.145	0.256	0.829

Table 9. General descriptive statistics of the relevant text mining factors for RIX estimation according to the banks' pool assignment. The factors are: word count (wc); paragraph count (pc); cooccurrence value of the keyword "*risk*" measured with the log-likelihood measure (loglikr); cooccurrence value of the keyword "*risk*" measured with the dice measure, only words which where identified as positive and negative in cooccurrence with "*risk*" are counted (dicer); multiplication of all values of identified cooccurrence words, not only positive and negative risky words (diceg); logarithm of number of unique positive words (lnsentipa); logarithm of number of unique negative words (lnsentina); logarithm of sum of positive words (lnsentips); logarithm of sum of negative words (lnsentips); netted sentiment index (sentigp); and risk reporting quality index (rix). Obs is the number of observations; Std is the standard deviation and Min and Max are the minimum and maximum values, respectively. We report further text mining factors concerning readability, lexicographically, and counting measures in Table 13.

	Private					Cooperative					Public				
Factor	Obs	Mean	Std	Min	Max	Obs	Mean	Std	Min	Max	Obs	Mean	Std	Min	Max
wc	120	9,078.800	7,150.430	1,645.000	5,330.000	60	6,384.767	6,236.487	714.000	32,164.000	163	6,352.067	3,491.110	350.000	13,822.000
pc	120	42.421	25.145	7.000	164.000	60	28.783	24.557	2.000	140.000	163	35.707	22.106	1.000	91.000
loglikr	120	5.643	2.072	1.029	11.513	59	5.060	1.850	2.069	9.501	161	4.935	1.651	0.913	9.042
dicer	120	0.544	0.233	−0.103	1.033	59	0.599	0.299	0.000	1.353	161	0.546	0.289	−0.026	1.488
diceg	120	0.409	0.241	−0.139	1.225	59	0.393	0.201	0.000	0.733	161	0.404	0.233	−0.099	1.097
lnsentipa	120	4.701	0.372	3.737	5.617	60	4.366	0.552	2.773	5.521	163	4.454	0.508	2.302	5.153
lnsentina	120	3.510	0.601	1.791	4.682	60	3.021	0.717	1.387	4.585	163	3.248	0.529	1.609	4.078
lnsentips	120	5.348	0.606	4.111	7.245	60	4.855	0.787	2.833	6.773	163	4.999	0.691	2.485	6.043
lnsentins	120	4.221	0.835	1.946	6.423	60	3.525	0.946	1.387	5.841	163	3.801	0.680	1.609	5.011
sentigp	120	−17.945	17.325	−119.138	0.316	60	−11.405	16.938	−98.554	0.887	163	−10.499	8.504	−36.538	4.118
rix	101	0.585	0.156	0.141	0.919	48	0.539	0.178	0.107	0.906	160	0.559	0.187	0.073	0.889

Table 10. General descriptive statistics of the relevant text mining factors for RIX estimation according to the bank type. We group private banks and banks with a special business model in the category Private; cooperative banks in Cooperative; and public savings and federal state banks in Public. The factors are: word count (wc); paragraph count (pc); cooccurrence value of the keyword "*risk*" measured with the log-likelihood measure (loglikr); cooccurrence value of the keyword "*risk*" measured with the dice measure, only words which where identified as positive and negative in cooccurrence with "*risk*" are counted (dicer); multiplication of all values of identified cooccurrence words, not only positive and negative risky words (diceg); logarithm of number of unique positive words (lnsentipa); logarithm of number of unique negative words (lnsentina); logarithm of sum of positive words (lnsentips); logarithm of sum of negative words (lnsentips); netted sentiment index (sentigp); and risk reporting quality index (rix). Obs is the number of observations; Std is the standard deviation and Min and Max are the minimum and maximum values, respectively. We report further text mining factors concerning readability, lexicographically, and countings in Table 13.

Sentiment detection

Term	Term frequency	Sentiment	Weight	Term	Term frequency	Sentiment	Weight
loss	2514	neg	−0.520	to reduce	579	pos	0.005
current	1324	pos	0.004	responsibility	572	pos	0.004
to lead	1314	pos	0.004	responsible	516	pos	0.004
appropriate	1017	pos	0.004	individual	508	pos	0.004
economical	977	pos	0.004	to fulfill	495	pos	0.004
fundamental	939	pos	0.004	default	482	neg	−0.216
historic	929	pos	0.004	enable	480	pos	0.004
independent	880	pos	0.004	maximum	475	pos	0.004
compliance	850	pos	0.004	large	462	pos	0.369
consistent	781	pos	0.004	detailed	455	pos	0.004
minor	776	neg	−0.662	important	426	pos	0.382
fraud	767	neg	−0.491	restriction	424	pos	0.005
to raise	760	neg	−0.004	appropriate	423	pos	0.097
active	737	pos	0.091	quality	421	pos	0.004
legal	720	pos	0.004	improvement	408	pos	0.004
percentage	715	pos	0.004	value	406	pos	0.004
danger	712	neg	−1.000	special	396	pos	0.004
qualitative	697	pos	0.004	extensive	390	pos	0.004
goal	691	pos	0.004	continuous	386	pos	0.004
strong	681	pos	0.004	dependence	384	neg	−0.365
increase	655	neg	−0.004	collaboration	379	pos	0.089
sufficient	639	pos	0.004	sustainable	367	pos	0.004
decline	604	neg	−0.210	low	358	neg	−0.362
meaning	588	pos	0.004	law	353	pos	0.004
unexpected	581	neg	−0.035	classic	349	pos	0.004

Table 11. List of the 50 most frequent terms from the SentiWS dictionary in the reports.

Readability

		Fog		fk		Forcast	
Year	Obs	Mean	Std	Mean	Std	Mean	Std
2002	24	22.881	0.257	21.086	0.298	13.783	0.056
2003	29	22.950	0.195	21.346	0.254	13.802	0.054
2004	30	22.754	0.172	21.192	0.198	13.772	0.057
2005	30	22.707	0.217	21.096	0.229	13.757	0.069
2006	30	22.705	0.174	21.116	0.199	13.763	0.061
2007	30	22.313	0.161	20.733	0.187	13.721	0.048
2008	30	22.221	0.187	20.620	0.210	13.673	0.046
2009	28	22.459	0.151	20.833	0.174	13.689	0.044
2010	28	22.400	0.170	20.791	0.203	13.666	0.047
2011	28	22.350	0.155	20.744	0.186	13.664	0.046
2012	28	22.397	0.160	20.812	0.180	13.677	0.042
2013	28	22.336	0.156	20.764	0.189	13.673	0.047

Table 12. Evolution of the determined readability measures in a year-by-year analysis. We consider all of the 30 banks with a total of 343 observations. The factors are: Gunning-Fog-Index (Fog); Grade level of Flesch-Kincaid Index (fk); Forcast Index (Forcast). Obs is the number of observations, i. e. the number of reports, and Std is the standard deviation.

Text mining measures

		Lexical diversity				Counting measure				Average measures			
		TTR		MTLD		lnnuadj		sentc		avgsentcl		avgwordl	
Year	Obs	Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean	Std
2002	24	0.395	0.014	134.577	3.117	6.086	0.131	253.166	35.910	16.827	0.359	7.439	0.045
2003	29	0.390	0.014	131.511	3.244	6.096	0.143	259.552	30.669	17.072	0.330	7.474	0.036
2004	30	0.384	0.011	131.040	2.999	6.125	0.120	265.066	27.962	16.553	0.279	7.498	0.031
2005	30	0.372	0.012	130.011	3.261	6.256	0.124	305.664	29.893	16.679	0.349	7.464	0.037
2006	30	0.351	0.009	127.087	3.123	6.458	0.091	345.995	29.150	16.661	0.299	7.473	0.041
2007	30	0.331	0.009	124.043	2.984	6.632	0.101	449.099	40.333	15.919	0.272	7.451	0.040
2008	30	0.314	0.009	125.698	2.972	6.849	0.100	554.633	53.111	16.122	0.288	7.416	0.043
2009	28	0.311	0.010	120.851	2.619	6.884	0.109	562.145	52.750	16.301	0.282	7.466	0.036
2010	28	0.303	0.010	119.381	2.430	6.948	0.114	611.397	61.310	16.284	0.285	7.447	0.039
2011	28	0.295	0.009	118.481	2.667	7.068	0.099	668.212	64.719	16.223	0.267	7.437	0.034
2012	28	0.295	0.011	118.799	2.466	7.076	0.120	717.853	98.191	16.096	0.273	7.468	0.033
2013	28	0.293	0.010	118.658	2.490	7.101	0.121	751.496	113.527	16.070	0.278	7.453	0.042

Table 13. Additional text mining measures concerning lexicographic, counting measures, and readability. We consider all of the 30 banks. The Type-Token-Ratio (TTR) is the ratio of the number of unique words (types) to all words (tokens). The measure of textual lexical diversity (MTLD) uses a special factorized text form and is independent of the text's length. Further factors are: logarithm of adjective count (lnnuadj); the number of sentences per report (sentc); the average sentence length (avgsentcl); and the average word length (avgwordl). Obs is the number of observations, i. e. is the number of reports, and Std is the standard deviation. We exclude the factors of the lexical diversity from our descriptive analysis and the prediction of the RIX because they vary only slightly and, therefore, have almost no influence and explanatory power concerning the RIX. The counting measures are highly correlated with the word count and, thus, we omit these factors from our analysis. The average sentence and word lengths primarily influence the readability measures but also have little explanatory power concerning the variance of the RIX.