Predicting Earnings Management from Qualitative Disclosures

Johannes Jaspersen

Andreas Richter

Sandra Zoller

Munich Risk and Insurance Center

Working Paper 40

January 1, 2021

Predicting Earnings Management from Qualitative Disclosures
Johannes Jaspersen, Andreas Richter, and Sandra Zoller
MRIC Working Paper No. 40
November 2020

ABSTRACT

While analysts, customers, and lenders rely on financial disclosures to make decisions regarding a company, executives often manage the disclosed earnings. Detecting such practices is thus a concern for company stakeholders and regulators. Qualitative disclosures are an additional source of information about a company's financial situation, but executives likely attempt to hide their earnings management activity in these disclosures, as well. We use supervised machine learning models to predict earnings management by property and casualty insurers from the Management’s Discussion and Analysis filings. For this, we utilize a new algorithm that interprets textual data conditional on the reported financial situation of the company. We show that the qualitative disclosures can predict earnings management, revealing that executives are unable to remove all subliminal messages from them. The results demonstrate that qualitative disclosures can be useful for learning about the accounting choices of companies.

Johannes Jaspersen
Institut für Versicherungsbetriebslehre
Leibniz University of Hannover
Otto-Brenner-Straße 7
30159 Hannover
Germany
jgj@ivbl.uni-hannover.de

Andreas Richter
Munich Risk and Insurance Center
Munich School of Management
Ludwig-Maximilians-Universität in Munich
Schackstraße 4/III
80539 Munich
Germany
richter@bwl.lmu.de

Sandra Zoller
Munich Risk and Insurance Center
Munich School of Management
Ludwig-Maximilians-Universität in Munich
Schackstraße 4/III
80539 Munich
Germany
zoller@bwl.lmu.de

1 Introduction

Earnings are a major explanatory factor for returns to equity, and they are considered an important item of a financial statement by analysts, investors, and boards of directors alike (Bhojraj et al., 2009; Degeorge et al., 1999; Hazarika et al., 2012; Kothari and Sloan, 1992; Leuz et al., 2003). Executives thus have a strong incentive to manage the firm's earnings. On the one hand, they might act out of self-interest because employee bonuses, employment decisions, or concerns about their external reputation are often tied to earnings. On the other hand, they might act on the stakeholders' behalf, since they expect a higher stock price from smoothed earnings (Graham et al., 2005). Executives can manage earnings by making choices about the real economic activity of a firm and by using the flexibility allowed by standard accounting principles – a practice that is often referred to as accrual-based earnings management (Roychowdhury, 2006; Dechow et al., 2010; Fields et al., 2001). Both forms of earnings management compromise the information value of financial reports. This is problematic for stakeholders who base their decisions on these statements. Detecting earnings management is thus an important question in accounting research (Bhojraj et al., 2009; Dechow and Skinner, 2000; Efendi et al., 2007).

In this study, we use companies' qualitative disclosures to provide a new approach to predicting earnings management. In general, qualitative disclosures contain a wealth of information (Armstrong et al., 2010; Guay et al., 2016; Lang and Stice-Lawrence, 2015) and provide executives with more opportunities to exercise discretion than quantitative disclosures (Brown and Tucker, 2011; Bozanic et al., 2018). It is thus possible that such disclosures contain information pertaining to the earnings management of the firm. Even though managers may not intend to disclose any such information, previous research in other contexts has shown that the subtext of qualitative statements might nevertheless include it (Hoberg and Maksimovic, 2014; Humpherys et al., 2011). In a first step, we use supervised machine learning techniques to calibrate a text-based classification model on qualitative disclosures in annual statements. In a second step, we enrich these models by also using financial disclosures as predictors. We are thus the first to use a classification protocol that integrates quantitative information with qualitative disclosures and allows such disclosures to have different meaning depending on the economic context of the firm. In our analysis, we use detailed financial filings by 722 property and casualty insurance companies and match them with Management's Discussion and Analysis (MD&A) filings. The insurance industry offers a precise and unbiased measure of earnings management, the so-called reserve error. We transform the sign of this reserve error into a binary classification of over- and underreserving firms, which we then use as the target variable of our earnings management classification protocol. We search for the best classification model in a broad set of models proposed in the literature. Subsequently, we fine-tune the parameters of the most promising candidates. We also explore different forms of text representation, including an unsupervised topic model. To ensure validity, we use a cross-validation approach throughout the model selection and evaluate the final classification model on a hold-out test sample.

Our results confirm that the qualitative disclosures published by a company are indicative of that company's earnings management. Prediction quality increases further when financial indicators are used in addition to the qualitative information. In our empirical setting, we find that, depending on the final model employed, using the information from the MD&A sections and financial indicators helps to predict the sign of the insurers' reserve error with an accuracy of 68 to 70 percent. This prediction is achieved by a Stochastic Gradient Descent model that uses term frequency–inverse document frequency (TF-IDF) statistics of uni- and bigram word combinations together with profit, growth, business concentration and financial distress information on the company. Considering the specific prediction errors, over three out of four predicted cases of overreserving are actually overstated reserves (precision) and the share of predicted overstated reserves among all actual overstated reserves is between 76 and 83 percent (recall).

We contribute to the literature by demonstrating the utility of MD&A sections for learning about the accounting choices of a company. Managers who aim to manipulate disclosed earnings to influence share prices or their own reputation should also aim for an MD&A section that is fully consistent with the firm's reported earnings. The fact that our model does render predictions based on qualitative information shows that managers are unsuccessful in this attempt. Our classification model thus presents a tool for practitioners to obtain early information on a company's earnings management. The model indicates the potential direction of the earnings management immediately after the financial disclosure. This is opposed to conventional calculations of earnings management, which often require a considerable time delay. The reserve error calculations for insurance companies, for example, have a time lag of five or more years.

We also provide a methodological contribution to the use of machine learning in accounting research. We test different ways of combining text-based information with quantitative disclosures. This includes a method that embeds the qualitative information in the economic context of the firm. Here, we allow qualitative disclosures to have different meanings based on the financial situation of the analyzed company. To achieve this, we condense the financial information into a set of binary indicators and interact those with the sparse TF-IDF matrix that represents the MD&A section. Previous literature which combines qualitative and quantitative information uses only qualitative information in the machine learning protocol and then uses the protocol's prediction in conjunction with the quantitative information. Since the protocol's prediction aggregates the information provided by the words used in the qualitative disclosures, such a procedure can only bestow a fixed meaning on a single word, which is less flexible than the approach developed here. Our results demonstrate that the integrated analysis of both qualitative and quantitative information in a machine learning model leads to better predictions – both in-sample and out-of-sample – than the previously used approaches.

Prior literature has used different approaches to identify earnings management. While some studies have exploited specific real activity choices, like asset sales and decreases in research and development expenditures (Dechow and Sloan, 1991; Bartov, 1993) or overproduction and sales discounts (Roychowdhury, 2006), the earnings management literature is mainly focused on accrual-based earnings management (see, e.g., Bergstresser et al., 2006; Cornett et al., 2008).[1] These studies can be categorized into three types of approaches (McNichols, 2000): aggregate accruals, industry-specific accruals, and the distribution of earnings. Aggregate accruals models attempt to decompose total accruals into their discretionary and non-discretionary components.

[1] Dechow and Skinner (2000), Dechow et al. (2010), Healy and Wahlen (1999) and Schipper (1989) provide comprehensive reviews of the earnings management literature.

Even though such measures are commonly used to study earnings management, results may be biased by the potential for a misspecification error in the discretionary accrual proxy (see McNichols, 2000, for a detailed analysis). This potential bias is not present in certain industry-specific accruals. For example, insurers' reserve errors allow for a more precise and largely unbiased measurement of earnings management (see, e.g., Beaver et al., 2003; Ding et al., 2020; Eckles et al., 2011; Grace and Leverty, 2012). However, this measure implies a substantial time lag for any analysis because such discrepancies can only be observed several years after the initial loss reserve is published. Distributional measures of earnings management do not suffer from this shortcoming as they investigate anomalies in the distribution of earnings in a specific year. Yet, these measures can, by definition, only provide an aggregated metric for a group of companies (Burgstahler and Dichev, 1997; Degeorge et al., 1999). Our prediction approach combines the measurement specification benefits of industry-specific accruals with the timely indication of earnings management as it is achieved by distributional earnings and aggregated accrual measures.

Our paper contributes to a growing literature examining the informative power of qualitative disclosures for accounting research. Gentzkow et al. (2019) and Loughran and McDonald (2016) provide surveys of historical advances and recent innovations in textual analysis in social sciences with an emphasis on finance. Early contributions find that executives are willing to provide more information in qualitative disclosures when their firms are performing well (see, e.g., Lang and Lundholm, 1993; Schrand and Walther, 2000). Subsequent research assesses the linguistic characteristics of qualitative information, such as tone (Tetlock, 2007), readability (Li, 2008), and forward-looking information (Bozanic et al., 2018; Muslu et al., 2014). Another stream of research explores a different approach by examining the similarity between disclosures. Applications range from creating new industry classifications (Hoberg and Phillips, 2016) to measuring financial constraints (Hoberg and Maksimovic, 2014). A recent line of research investigates how unsupervised machine learning topic models can retrieve information from the qualitative disclosure in 10-K filings (see, e.g., Bao and Datta, 2014; Dyer et al., 2017; Huang et al., 2018; Lopez-Lira, 2019). However, only a few papers use machine learning classification protocols to make predictions based on information from qualitative disclosures. An early application is presented by Antweiler and Frank (2004), who classify news on internet stock message boards as bullish, bearish, or neither with an accuracy of 84 percent. They show that these classifications help in predicting stock market volatility. Li (2010) uses a machine learning model to classify the forward-looking statements in MD&A sections. His model indicates the tone and content category of a statement with an accuracy of 63-67 percent. He finds that the tone of the statements is positively associated with future earnings.

Our study is most closely related to three previous analyses. Frankel et al. (2010) show that historical MD&A sections can add explanatory power to aggregated accruals models by creating a new MD&A-based, independent variable with a machine learning model. Contrary to them, we study the predictive capabilities of MD&A sections across firms rather than across time. We also combine qualitative and quantitative disclosures within the machine learning model rather than considering quantitative information only after training the model. Frankel et al. (2010) thus enrich the aggregated accruals framework, while we develop a new and alternative framework for studying earnings management. Ding et al. (2020) use supervised machine learning to estimate the loss reserve of property and casualty insurers from financial disclosures. They thus also work outside the aggregated accruals framework, but consider only the loss reserve part of the earnings management process and ignore the managerial estimate of the reserve. While their results demonstrate the power of machine learning to improve accounting estimates like business line specific loss reserves, their study neither considers qualitative disclosures nor studies earnings management. Thus, while similar in methodology, our study differs from that by Ding et al. (2020) in purpose. We simply use machine learning tools to test hypotheses derived from our conceptual model. Lastly, in a study unrelated to earnings management but with a similar approach as ours, Humpherys et al. (2011) apply supervised machine learning to detect fraudulent financial statements from qualitative disclosures. Their final classification model, which has an accuracy of 67 percent, is only based on qualitative information. Similar to our result, their model shows the managers' lack of ability to remove all involuntary information disclosure from the MD&A section. Note, however, that Humpherys et al. (2011) use 10-fold cross-validation to test their model, while we use a hold-out set that was completely separated from the calibration process.[2]

The rest of the paper is organized as follows. In Section 2, we present a conceptual model, that links earnings, earnings management, and disclosure, and the institutional setting. In Section 3, we describe the sample selection process as well as the basic characteristics of the data. The methodological approach is outlined in Section 4. We present the results of the earnings management predictions in Section 5 before we conclude in Section 6.

2 Model and institutional setting

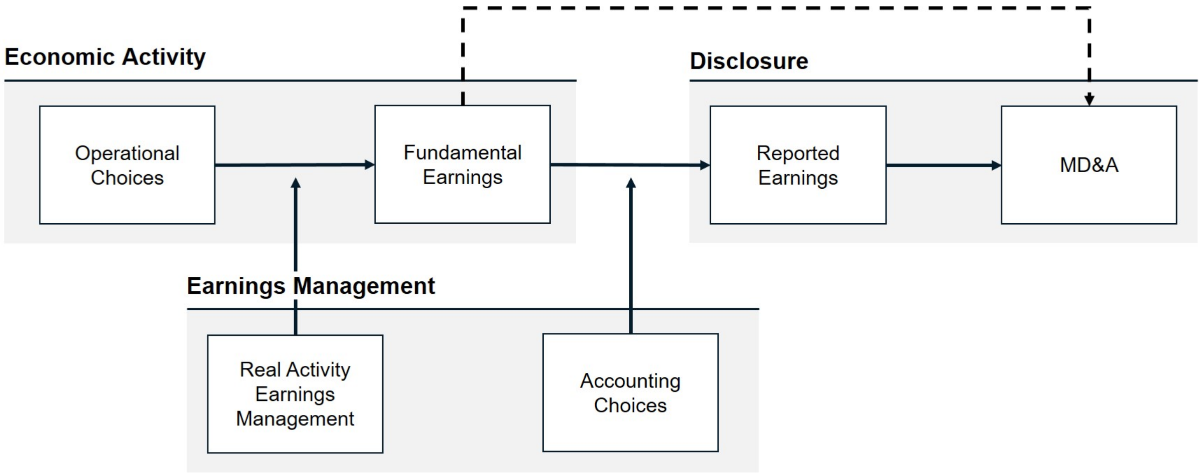
2.1 Conceptual model

We use the conceptual model in Figure 1 to outline how economic activity, earnings management, and disclosure are connected. Operational choices let fundamental earnings evolve during a business year. Managers can influence this process through real activity earnings management as describe above. Once fundamental earnings are realized for a business year, they are reported in the financial disclosure. The reported earnings differ from the fundamental earnings if the management decides to manage earnings through accounting choices. The earnings management literature makes the abstraction that fundamental earnings are the accurate depiction of economic activity through the firm's accounting system (Dechow et al., 2010). Earnings management in form of accounting choices adjusts the accurate picture to an inaccurate picture. The difference between real activity and accounting choice earnings management is that real activity earnings management changes the activity of the firm, such that the accurate depiction in the fundamental earnings would be changed. Hence, reported earnings are a function of fundamental earnings and accounting choices (Dechow et al., 2010). In addition to the financial

[2] Given certain assumptions on the data generating process, n-fold cross-validation can be used to construct a consistent estimator of out-of-sample prediction validity (Rabinowicz and Rosset, 2020). In practice, however, this property does not always seem to hold, even when only cross-sectional data is analyzed. This can also be observed in our results reported here. 10-fold cross-validation on the training set leads to significantly different results for our main fit criterion (the AUC) than the test on the hold-out set (compare results reported in Tables 6 and 7).

disclosures, managers create the accompanying MD&A section. Together, both parts form the annual statement of the company (Hoberg, 2016).

Figure 1: Conceptual model of earnings, earnings management, and disclosure



Managers engage in earnings management for various reasons. Fields et al. (2001) use the conditions of Modigliani and Miller (1958) to classify the goals for accounting choices in three categories: contracting, asset pricing, and influencing external parties. The first category is a result of agency costs in an incomplete market, where earnings management is used to influence a contractual arrangement, such as managerial compensation (see, e.g., Bergstresser and Philippon, 2006; Cornett et al., 2008; Eckles and Halek, 2010; Efendi et al., 2007, for empirical analyses). The second category stems from asymmetric information between managers and investors and attempts to influence asset prices (Beaver et al., 2003; Louis, 2004; Teoh et al., 1998a,b). The third category, driven by existing externalities, is to influence external non-contracting parties. By earnings management, managers hope to influence the decision of third parties, such as the regulator or the Internal Revenue Service (Folsom et al., 2017; Gaver and Paterson, 2004; Grace and Leverty, 2010; Scholes et al., 1990; Shrieves and Dahl, 2003).

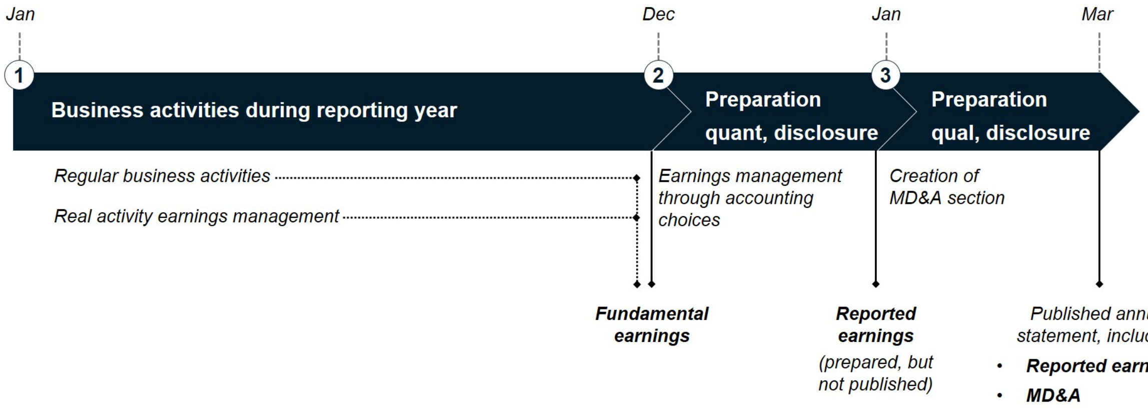
Given the common motivations for earnings management, managers usually should not be interested in disclosing any indications that they have engaged in the practice. Thus, managers will attempt to make any qualitative information they disclose consistent with the reported earnings (McCornack, 1992; Bloomfield, 2008). As such, the only path of influence that fundamental earnings should have on the MD&A section should be through the reported earnings. However, recent analyses of qualitative disclosures (Humpherys et al., 2011; Muslu et al., 2014) as well as applications of linguistic theory to disclosures (Li, 2008) let us hypothesize that managers fall short of this goal and that MD&A sections might in fact be influenced by fundamental earnings directly. This hypothesis is denoted by the dashed arrow in Figure 1 and constitutes the main research question of our analysis.

2.2 MD&A sections

The main qualitative disclosure required by the Securities and Exchange Commission (SEC) for annual and quarterly financial reporting is the MD&A section. While this section must cover certain topics, managers have flexibility in choosing the breadth and depth of what is discussed. The SEC specifies the purpose of the MD&A section as providing readers with information ”necessary to an understanding of [a company's] financial condition, changes in financial condition and results of operations” (SEC, 2003). To fulfill this purpose, the SEC requires managers to discuss and analyze the company through the eyes of the management. Therefore, the MD&A should not simply recite the financial statements in a qualitative form, but rather provide the management's perspective on the financial statements and the context within which they should be analyzed (SEC, 2003). For instance, the MD&A should address why earnings have changed or what liquidity needs the firm has. Despite these requirements, content of the MD&A disclosure is largely discretionary and not audited, which is in contrast to the notes accompanying the financial statements (Humpherys et al., 2011).

The timeline of how the MD&A section is created is portrayed in Figure 2. Operations and managerial choices throughout the business year result in fundamental earnings which are transformed to reported earnings through accounting choices. Based on these reported earnings, which are commonly determined in the month after the business year ends, managers then prepare the MD&A section. We can thus see a clear temporal structure in the creation of the different disclosures. Managers have the option to tailor the MD&A section to the reported earnings, which are determined and audited before its creation. At first glance, this option also seems achievable, given that managers both have sufficient time and substantial ressources (in the form of aides or consultants) to craft the MD&A section according to their preferences. However, linguistic theory informs us that the creation of any qualitative communication consists of attributes in several different dimensions, leading to virtually infinite combinations of choices available to the managers (Grice, 1989). Because there is no one obvious choice available for phrasing the MD&A section and because receivers of the communication might interpret it differently than senders (Shannon, 1948), it is likely that managers are unable to eliminate all subliminal messages in the communication. Thus, the MD&A section will likely include information which, against the intend of the managers, is not fully consistent with the reported earnings.

Figure 2: Exemplary firm's creation of an annual statement


Jan
Dec
Jan
Mar
1
2
3
Business activities during reporting year
Preparation quant, disclosure
Preparation qual, disclosure
Regular business activities
Real activity earnings management
Earnings management through accounting choices
Creation of MD&A section
Fundamental earnings
Reported earnings
(prepared, but not published)
Published annual statement, including:
• Reported earnings
• MD&A

Empirical findings also suggest that managers unintentionally disclose information that they (at least by best assumptions) do not want to disclose. Hoberg and Lewis (2017) and Humpherys et al. (2011) are able to differentiate fraudulent from non-fraudulent firms based on their qualitative disclosure. One can conjecture that managers have no incentive to be detected in their fraudulent activities. However, even if managers are able to remove all indications of fraudulent behavior from their disclosure, the very act of deception itself can lead to cues in their qualitative statements which would normally be absent (for a summary, see Humpherys et al., 2011). In addition to the results on fraudulent behavior, these studies also reveal that managers will subconsciously use ”nonimmediate” language to disassociate themselves from bad events and give greater weight to internal factors as explanation for good events. This self-serving attribution is also found to increase overconfidence, which manifests another bias that causes managers to overestimate their ability to predict future firm performance and influences their disclosure behavior (Billett and Qian, 2008; Larcker and Zakolyukina, 2012).[3]

2.3 Property and casualty insurers and earnings management through loss reserves

In this study, we analyze the property and casualty insurance industry which is frequently used to study earnings management (Beaver et al., 2003; Gaver and Paterson, 2004; Grace and Leverty, 2012). The industry has the key advantage that it provides an unbiased industry-specific measure of earnings management, the loss reserve error, which is based on publicly available financial disclosures. In comparison to methods based on aggregate accruals, the measurement error is relatively low, which makes the loss reserve error an attractive earnings management measure for our study aimed at predicting earnings management.

Property and casualty insurance is an umbrella for different types of insurance covering personal and commercial property and legal responsibility for losses stemming from damage to another’s property or personal well-being. Exemplary property and casualty insurance types are homeowners insurance, auto liability insurance and medical malpractice insurance. The industry does not provide certain other types of insurance coverage, such as health or life insurance. The property and casualty insurance industry contributes to the global economy with \$1.6 trillion in gross written premiums. In North America, the industry generates \$723 billions in gross written premiums.[4] The regulation of the property and casualty industry in the United States is performed exclusively by the states. However, the National Association of Insurance Commissioners (NAIC) provides expertise, data, and analysis for insurance commissioners to effectively regulate the industry and collects all financial filings. This ensures consistent disclosure practices throughout all states.

Earnings management in property and casualty insurers is primarily measured through loss reserve errors (Beaver et al., 2003; Grace and Leverty, 2012; Petroni, 1992). The loss reserve represents the estimated future cost of settling claims which occurred in the current business year but are not yet fully settled. It is thus a material accrual, that is estimated subjectively and the development of which is observable over time. The loss reserve is usually the largest liability on a property and casualty insurer’s balance sheet and due to its subjective nature, offers the largest amount of discretion for earnings management. After receiving a recommendation for

[3] See Baker and Wurgler (2013) for a survey of the related behavioral finance literature.

[4] Figures for 2018, based on McKinsey & Company (2020).

an acceptable loss reserve by the firm's actuaries, management will choose the final amount of the reserve.

Over time, as claim settlement progresses, the insurer adjusts its estimated loss reserve. These revisions indicate whether the insurer under- or overreserved in the business year of the loss. The loss reserve error for a company i is calculated as the estimated total incurred losses in a calendar year t minus the revised estimate of the incurred losses in t reported in the calendar year $t+n$ (Beaver et al., 2003; Gaver and Paterson, 2004; Grace and Leverty, 2012):

$$Error_{i,t,n} = Incurred\ losses_{i,t,t} - Incurred\ losses_{i,t,t+n} \tag{1}$$

If the estimated loss reserve is overstated, the insurer overreserved and the loss reserve error is positive. With an increasing n the estimate of the reserve error becomes more acurate as more claims are settled and more information about still outstanding claims is available. The common assumption for n is 5 (Beaver et al., 2003; Gaver and Paterson, 2004; Eckles and Halek, 2010; Grace and Leverty, 2012). That is, the reserve error compares the estimated loss reserve with the revised estimate after 5 years.

3 Data

In our study, we use detailed financial filings by 722 group and unaffiliated property and casualty insurance companies and match them with hand collected MD&A sections of filings. Because prediction algorithms in general and machine learning techniques in particular cannot with standard techniques be applied to panel data, we focus our analysis on a single year of data.[5] We use the year 2012 because it is sufficiently long ago such that we can calculate the reserve error reliably. For financial information, the primary data source are the annual statement filings with the NAIC. The sample is collected using the following criteria:

(i) the company's total assets are positive,

(ii) the company reports a positive loss reserve in 2012,

(iii) the company's developed loss reserve after five years (i.e., in 2017) is available,

(iv) the company is not primarily a reinsurer, whose direct premiums written exceed its assumed premiums,

(v) the company is based in the U.S.,

(vi) the company's reserve error is not extreme (i.e., the absolute reserve error scaled to total assets is smaller than one),

(vii) the company is a stock or mutual firm, and

(viii) the company's MD&A section is available.

[5] Prediction in panel data needs to take the autocorrelated nature of the data into account. This is typically done by assuming an autocorrelation mechanism and estimating its parameters from the available data. The estimation error introduced by this process could influence our result and would distract from the research question analyzed here. For machine learning, standard methods of cross-validation are dependent on non-correlated data, such that a panel structure would bias the results (Rabinowicz and Rosset, 2020).

For qualitative information, we manually collect the MD&A section for the reporting year 2012 for each insurance company, that meets the selection criteria (i) to (vii). The selection process yields a sample of 722 insurance companies as summarized in Table 1. The MD&A section has been filed as part of a 10-K filing with the SEC for 60 publicly traded companies and as part of the annual statements for 370 privately held stock insurance companies and 292 mutual insurance firms. Property and casualty insurance companies are required to file annual statements with the NAIC. We transformed all filings to raw text files. The details of the MD&A collection and document transformation are described in Appendix A.3.

Table 1: Sample selection

Selection criteria	**Number of firms**
Total assets > 0	1222
Reported reserve > 0	980
Developed reserve > 0	945
No primary reinsurer (DPW $>$ Assumed Premiums)	909
Domestic US firm	907
$\lvert$Scaled reserve error$\rvert < 1$	904
Stock or mutual ownership form	733
Available MD&A section	722
Total	**722**

Data comes from the companies' respective NAIC annual statements for 2012.

Table 2 shows the descriptive statistics for our sample. Consistent with prior studies (see, e.g., Beaver et al., 2003; Eckles and Halek, 2010; Grace and Leverty, 2012), we examine the five-year reserve errors, so $n = 5$ in Equation (2), and scale the reserve error to total assets to control for the company size.[6] On average, we observe the loss reserve to be overstated by 1% of total assets. Consequently, the mean original reserve exceeds the developed reserve.

Table 2 also presents descriptive statistics for other firm characteristics. Besides total admitted assets, net income, direct premiums written and revenue growth in direct premiums written, we also consider business concentration and financial distress. Business concentration is measured using the Herfindahl index, which denotes the sum of the squared percentage shares of premiums earned in each of the 45 property and casualty lines of business. The higher the Herfindahl index, the more concentrated is the business of a company. To measure financial distress, we add information about the Risk-Based Capital (RBC) ratio. The RBC method was developed as an additional monitoring tool to assist regulators. It establishes a minimum capital requirement based on a company's risk profile. A company's RBC ratio can trigger regulatory actions. Actions range from filing a business plan to the regulator being required to take control over the company unless the RBC level is corrected within 90 days.[7]

[6] Prior studies find that results are robust regarding different choices of the scaling variable and the development window (see, e.g., Beaver et al., 2003; Eckles and Halek, 2010; Grace and Leverty, 2012).

[7] The four RBC levels are defined: company action (RBC ratio $\leq$ 200%), regulatory action (RBC ratio $\leq$ 150%), authorized control (RBC ratio $\leq$ 100%), and mandatory control (RBC ratio $\leq$ 70%) (National Association of Insurance Commissioners (NAIC), 2019).

Table 2: Summary statistics

	Count	**Mean**	**SD**	**25%**	**50%**	**75%**
Scaled reserve error	722	0.0149	0.1006	-0.0061	0.0151	0.0514
Reported reserve	722	734.2433	3,868.8758	2.4380	16.0270	116.4905
Developed reserve	722	702.3914	3,790.7891	2.1318	13.4585	109.8980
Total assets	722	2,048.1794	11,574.8734	17.8359	80.7031	445.2465
Net income	721	49.0255	403.9909	0.0229	1.1751	8.7625
Direct premiums written	722	631.7138	3,193.1869	7.2752	36.7996	186.5501
Growth of direct premiums written	691	12.2079	46.2521	-0.0071	5.7713	13.0655
Business concentration index	695	0.5480	0.3309	0.2476	0.4753	0.9813
RBC ratio	683	1,297.1244	2,914.9587	496.2897	826.6821	1,281.0761
MD&A characters	722	32.1458	82.7371	8.3875	13.1140	20.6820

All numbers are taken from the companies' respective NAIC annual statements for 2012. *Reported reserve, developed reserve, total (admitted) assets, net income, and direct premiums written* are presented in million US-Dollars. *Growth of direct premiums written* and *RBC ratio* are expressed in percentage rates. *Reserve error* is scaled to total admitted assets. *Developed reserve* is calculated for a five-year development window, i.e., it is measured in 2017. *Premiums growth* measures the annual growth rate of direct premiums written from 2011 to 2012. *Business concentration index* is a Herfindahl Index across the property and casualty lines of business and is based on direct premiums earned. *RBC ratio* indicates the ratio of adjusted capital to the authorized control level. *MD&A characters* denote the count of characters (in 1,000) in the MD&A section of a company.

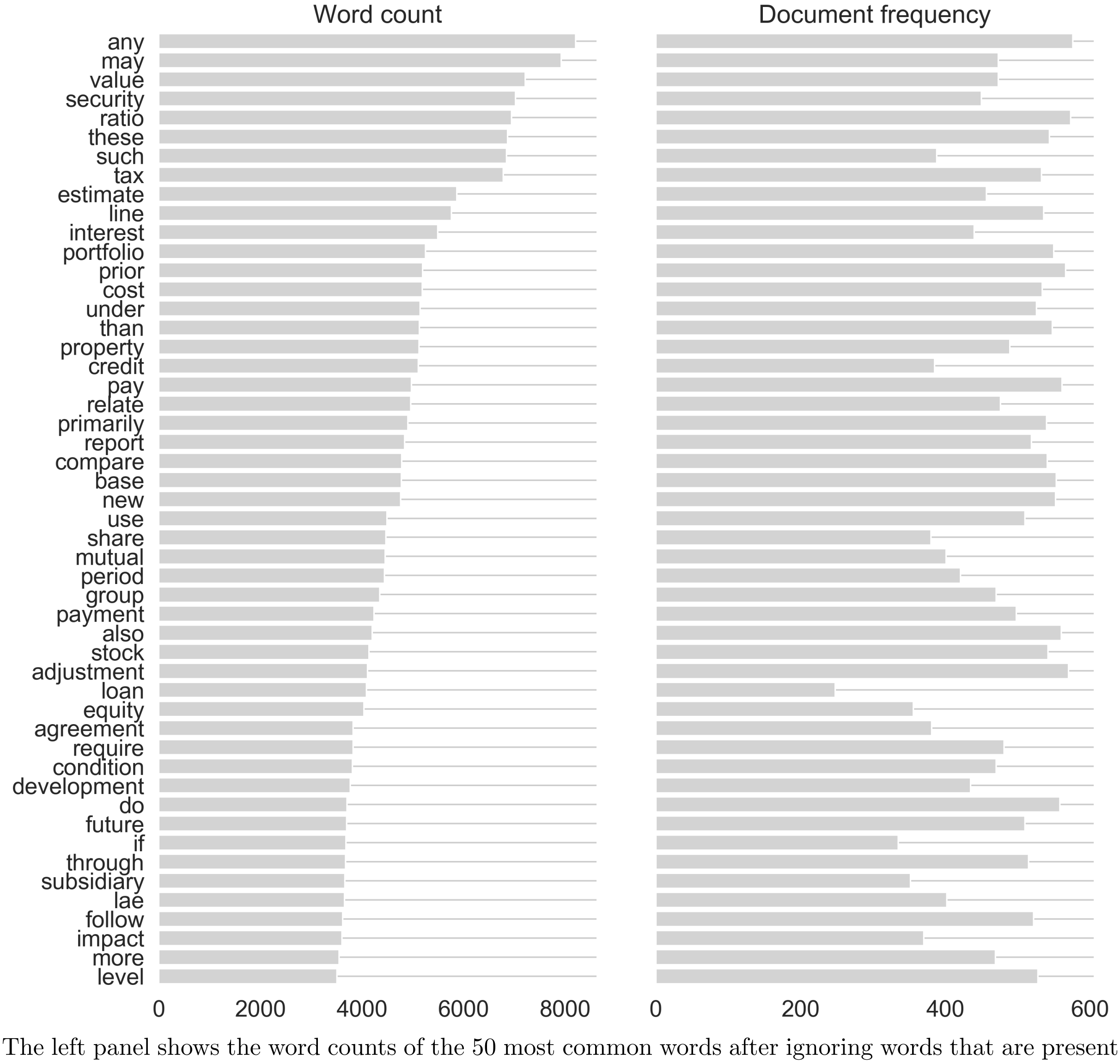
Mean (median) total assets are approximately \$2,048 million (\$81 million), which indicates that the sample is skewed by some particularly large insurance companies. This pattern is also reflected in the mean (median) net income of \$49 million (\$1 million), in the mean (median) direct premiums written of \$632 million (\$37 million), and in the mean (median) RBC ratio of 1,297 percent (827 percent). Growth of direct premiums written and business concentration are comparatively less skewed with a mean (median) of 12 percent (6 percent) and a Herfindahl Index of 0.55 (0.48). Note that the reserve error scaled to total assets is not skewed anymore, since those companies with high assets also display high original and developed loss reserves. Some differences between the two analyzed organizational forms exist. Mutual insurers have a higher average reserve error with otherwise similar distributional characteristics (see Figure A.2 in Appendix A.2). Stock insurers exhibit a larger size, higher growth, higher concentration of lines of business, higher RBC ratio, and longer MD&A sections compared to mutual companies (see Appendix A.2). Despite these differences, we will see that our model performs well in classifying stock and mutual companies in one sample.

Regarding the companies' qualitative disclosure, the MD&A sections comprise on average 32,150 characters. Figure 3 describes the most common words after ignoring words that are present in more than 80 percent of the documents and, thereby, largely uninformative.[8] The

[8] This filter is also a result of the text preprocessing fine-tuning which is part of the parameter tuning and presented in Section 4. Abbreviations and verbally written numbers are also ignored in Figure 3.

second column of the figure indicates the number of documents a specific word appears in. Even after filtering out words that appear in more than 80 percent of the documents, we can still observe some words without informative value among the remaining list of the most common words. For instance, the two most common words are ”any” and ”may” with approximately 8,000 mentions. Yet, most of these uniformative words are ignored as result of the filter. The remaining set of commonly used words consists mainly of industry-specific terms, such as ”security”, ”ratio”, ”line”, or ”property”, and evaluative terms, such as ”value”, ”estimate”, ”under”, ”than”, ”primarily”, or ”compare”.

Figure 3: Most common words in MD&A sections



The left panel shows the word counts of the 50 most common words after ignoring words that are present in more than 80 percent of the documents. The right panel indicates in how many documents a word appears in.

4 Methodology

This study tests whether the qualitative information in the MD&A section is indicative of earnings management. To that end, we transform the sign of the five-years reserve errors into a binary classification of over- or underreserving firms, which we then use as the target variable of our earnings management classification protocol. The machine learning approach follows six steps:

(i) the definition of feature matrix $\mathbf{X}$ and target variable $\mathbf{y}$,

(ii) the split into a training and a (hold-out) test set for evaluation and text vectorization,

(iii) the model selection and parameter tuning,

(iv) the optimization of text preprocessing,

(v) the integration of financial information, and

(vi) the model evaluation.

Steps (iii) to (v) are carried out solely on the training set, while only the last step uses both the training set (for calibrating the model) and the test set (for evaluating the model fit).

4.1 Definition of feature matrix and target variable

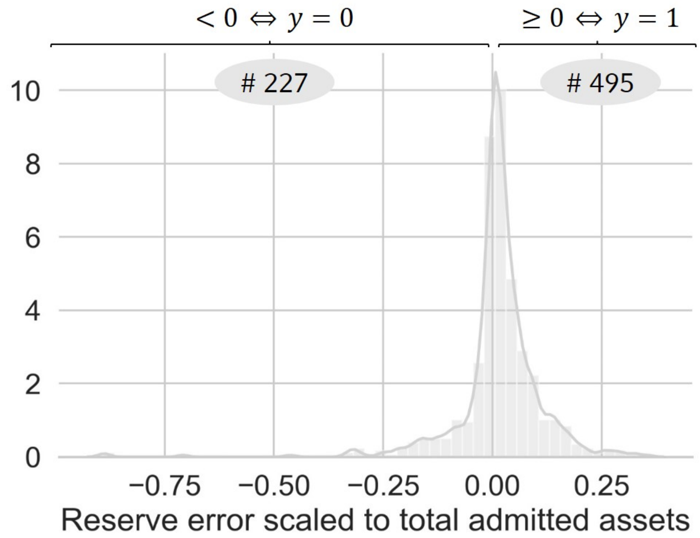
In a first step, we define the feature matrix $\mathbf{X}$ and the target variable $\mathbf{y}$. Our basic feature set comprises the MD&A sections of the insurers. From that, we create lemmatized tokens from each MD&A section's raw text. A token is a segment of the text in a document that is discovered based on text boundaries, such as white space, punctuation, or special case rules (e.g., N.Y. is not split into two tokens). A lemma is the base form of a word (i.e., the word's dictionary form), and allows to identify words with the same meaning (e.g., ”is”, ”are”, and ”being” become ”be”). As target variable $\mathbf{y}$ for our classification model, we create a binary indicator of over- and underreserving firms based on the sign of the five-years reserve errors.[9] While 227 insurance companies (31.44%) underreserve in our sample, a majority of 495 companies (68.56%) overestimate the loss reserve as presented in Figure 4.

4.2 Hold-out test set approach and text vectorization

In a second step, we split the sample of 722 property and casualty insurance companies into a training and a test set. This is necessary for any prediction model, including supervised learning models, in order to address the problem of overfitting: The prediction should be accurate out-of-sample, but we can only fit the model in-sample (Hastie et al., 2009). In our study, we use 80 percent of the observations for the training set (577 companies) and hold out the remaining 20 percent of the observations for the test set (145 companies). The assignment into training and test set is stratified on the target variable such that the proportion between under- and

[9] To evaluate different definitions, we also test the classification performance for an alternative target variable that splits the five-years reserve errors at the median. The results of this specification are described in the Section 5.

Figure 4: Distribution of loss reserve error scaled to total assets



The figure shows the distribution of the loss reserve error scaled to total assets. The annotations indicate the number of over- and underreserving firms.

overreserving companies in each subset is the same as in the complete sample. We fit the model on the training set and eventually evaluate it on the hold-out test set to assess how the model performs on new instances.

For all calibration steps (steps (iii) to (v) above), we use a cross-validation approach with folding to identify the best model based on the data in the training set. Under the condition of uncorrelated error terms, this procedure gives an unbiased estimate of the prediction error in out-of-sample predictions (Rabinowicz and Rosset, 2020). Cross validation implies that we fit the model on only a part of the training set and ask which parameters perform best on the other part of the data. Folding means that this process is repeated multiple times with different, equally sized subsamples of the training set – so-called folds – used for the evaluation in each step. In our study, we split the training set into five folds.[10] We hold out one of the folds for evaluation while fitting the model for a range of parameters on the remaining four folds. Then, we pick the parameter with the best average performance across all folds. As with the training-test-split, the folding procedure is stratified on the target variable.

Before we make any classification, we rescale the text features according to their informativeness. For this, we ignore all numbers and any punctuation in the MD&A sections. For the verbal content, we transform the lemmatized tokens of an MD&A section into a vector of TF-IDF weights (Salton and McGill, 1986). The term frequency counts the word appearances in a document and divides it by the document's total word count. The term frequency for a token is multiplied with the inverse document frequency, which is defined as the logarithm of the total documents divided by the number of documents that contain the particular word. Intuitively, a high weight is given to any term that appears often in a particular document but not in many

[10] For step five, the integration of financial information, we apply a ten-fold cross-validation approach due to a reduced sample size. Full financial information is only available for 649 firms.

documents. Thus, the token is descriptive for the content of the document. The result is a term-by-document matrix whose columns contain the TF-IDF values for each document in the sample. Thus, the TF-IDF transformation reduces documents of arbitrary length to fixed-length lists of term weights.[11]

4.3 Model selection and parameter tuning

In a third step, we consider twelve different machine learning classification models in a horse race for the most promising model. The models belong to different model families, including neighbor models, tree-based models, ensembling models like gradient boosting, vector machines, and neural networks.[12] For a brief summary of these models, see Appendix A.5 or refer to Hastie et al. (2009) for a more detailed discussion. Table 3 lists the average test scores of the cross-validation predictions on the training set. We report four different evaluation criteria. Accuracy is defined as the share of correct predictions. Precision measures the share of true positives (that is, the case of overreserving in our model) among all predicted positives. The F1 score is calculated as the harmonic mean of precision and recall (which measures the share of predicted positives among all actual positives). The area under the receiver operating characteristic curve (AUC) estimates the probability that a random positive is ranked before a random negative (in our model, the case of underreserving), without specifying a particular decision threshold.

We choose the most promising candidates based on the AUC, which is a standard measure of performance ranking (Hanley and McNeil, 1982; Ferri et al., 2011; Müller and Guido, 2016). The AUC is particularly appropriate when the sample is imbalanced, that is when the target variable is not uniformly distributed. In the case of an imbalanced sample with a binary target variable, accuracy should not be used alone, since the accuracy score would also indicate a high performance when the model only predicts the most prominent class of the target variable.[13]

Next, we fine-tune the model parameters of the most promising models, a Gradient Boosting Machine and a Stochastic Gradient Descent model. The Gradient Boosting Machine is an ensemble method that combines several decision trees. Gradient Boosting builds trees sequentially, where each tree tries to correct the errors of the previous one. The Stochastic Gradient Descent is a method associated with discriminant learning of linear classifiers under convex loss functions. A Gradient Descent measures the local gradient of the cost function with regards to the parameter vector and it moves in the direction of the descending gradient. Once the gradient is zero, it reaches a minimum. The gradient vector contains all the partial derivatives

[11] The term frequency assigns a high weight to tokens often appearing in a document, while the inverse document frequency scales down commonly used tokens in a sample. Since the inverse document frequency depends on the documents under consideration, we use a pipeline that first converts each feature subset with the TF-IDF method and then predicts with a particular classification model. If we were not to use a pipeline, the TF-IDF weights would be calculated based on the complete data set rather than only on the training subset.

[12] Namely, the twelve machine learning classification models in the horse race are K Neighbors, Logistic Regression, Naive Bayes Classifier, Decision Tree, Random Forest, Gradient Boosting Machine, Support Vector Machine with linear kernel, Support Vector Machine with RBF kernel, Linear Support Vector Machine, Stochastic Gradient Descent, Linear Discriminant Analysis, and Neural Network (lbfgs solver).

[13] For example, in our full sample, 68 percent of instances are of positive value. Thus, a model that always predicts a positive value would have an accuracy of 68 percent.

Table 3: Model selection

	Accuracy	Precision	F1	AUC	AUC ranking
K Neighbors	0.6049	0.7080	0.7125	0.5399	11
	(0.044)	(0.0267)	(0.0531)	(0.0385)	
Logistic Regression	0.6881	0.6875	0.8148	0.6100	6
	(0.004)	(0.003)	(0.0021)	(0.0328)	
Naive Bayes Classifier	0.6863	0.6863	0.8140	0.5540	9
	(0.0013)	(0.0013)	(0.0009)	(0.0578)	
Decision Tree	0.6133	0.7128	0.7183	0.5495	10
	(0.0562)	(0.0246)	(0.0554)	(0.0602)	
Random Forest	0.6898	0.6935	0.8128	0.5994	8
	(0.0197)	(0.01)	(0.0136)	(0.0304)	
Gradient Boosting Machine	0.6915	0.7164	0.8022	0.6366	1
	(0.0286)	(0.0173)	(0.018)	(0.0397)	
SVC (linear kernel)	0.6898	0.6887	0.8157	0.6115	5
	(0.0049)	(0.0037)	(0.0026)	(0.0173)	
SVC (RBF kernel)	0.6863	0.6863	0.8140	0.6090	7
	(0.0013)	(0.0013)	(0.0009)	(0.0141)	
Linear SVC	0.6915	0.6997	0.8110	0.6262	4
	(0.0075)	(0.005)	(0.0059)	(0.0304)	
Stochastic Gradient Descent	0.6759	0.7136	0.7847	0.6332	2
	(0.0437)	(0.0224)	(0.0598)	(0.0302)	
Linear Discriminant Analysis	0.6621	0.7250	0.7694	0.6317	3
	(0.0279)	(0.0255)	(0.0163)	(0.0298)	
Neural Network (lbfgs solver)	0.6672	0.7051	0.7848	0.5243	12
	(0.0174)	(0.0095)	(0.0163)	(0.0307)	

The table summarizes the performance scores of the cross-validation prediction of the alternative classification models with TF-IDF unigram word tokens of the MD&A sections. The table reports mean and standard deviation over the five cross-validation folds in brackets. Accuracy is defined as the share of correct predictions. Precision measures the share of true positives (that is, the case of overreserving in our model) among all predicted positives. The F1 score is calculated as the harmonic mean of precision and recall, which measures the share of predicted positives among all actual positives. The AUC estimates the probability that a random positive is ranked before a random negative (in our model, the case of underreserving), without specifying a particular decision threshold.

of the cost function for the complete training set. A Stochastic Gradient Descent is an efficient learning approach for linear classifiers with a convex cost function. The gradients are computed based on that single instance rather than on full training set.

The optimal parameters are found in a cross-validation grid-search with five folds. For the Gradient Boosting, we first search for optimal number of trees and then fine-tune tree-specific parameters. The tree-specific parameters are evaluated sequentially. First, we evaluate the maximal depth, which limits the number of nodes in each tree, and minimal samples to split an internal node. Second, we choose the optimal number of minimal samples at a leaf node. Third, we consider the maximal number of features considered at each split. The results are summarized in Section A.5.1 in Appendix A.5. The best cross-validation score is achieved by a Gradient Boosting model with 80 trees, where each tree has a maximal depth of 10, at least

20 samples at an internal node to be split, and at least 10 samples at a leaf node. In the best fitting model, all features are considered at each split.

For the Stochastic Gradient Descent, we search for the highest AUC score by varying the learning rate and the regularization term. The learning rate determines the strength of the model update after each loss gradient estimation. In general, the model is updated with a decreasing strength. If the learning rate is too small, then the algorithm will have to go through many iterations to converge. A too large learning rate may causes drastic updates that can hinder conversion. The regularization term is a penalty, which is added to the loss function and shrinks model parameters towards zero using either the squared euclidean norm (L2) or the absolute norm (L1) or a combination of both (Elastic Net). We evaluate different specifications of the regularization term together with a range of values for the learning rate in a cross-validation gridsearch. The results are shown in Table A.7 in Appendix A.5. The best Stochastic Gradient Descent model in the parameter tuning has a learning rate of 0.001 and implements regularization with an L2 penalty.

4.4 Optimization of text preprocessing

In a fourth step, we optimize the text preprocessing parameters to achieve the best prediction results. The initial model selection used the complete lemmatized text of the MD&A sections as a feature set. However, the individual lemmas vary in their informative value. If a lemma appears only in a small number of documents or is used in a large share of documents, its presence in a new document, for which a classification should be made, has little information. As such, we evaluate the prediction performance for possible combinations of different values for the maximal share of documents that a word appears in and for the minimum number of documents that a word appears in. The preprocessing parameter tuning results are shown in Appendix A.5.[14] We find that we can improve the performance of the Stochastic Gradient Descent model by fine-tuning the text preprocessing to ignore words that appear in less than 10 documents and words that appear in more than 80 percent of the documents. The Gradient Boosting model performs best with the complete text corpus.

We also investigate whether we can increase out-of-sample performance by removing certain categories of words from the lemmatized text. Specifically, we consider dropping individual and company names, locations, verbally expressed numbers, and abbreviations. These word categories are typically highly informative regarding an individual company, but may lead the model to overfit on the training set, which reduces out-of-sample performance.[15] We combine this analysis in a cross-validation grid-search together with an analysis of how may words are considered for each lemma. The meaning of a word often not only depends on the word itself but also on its neighbors. We compare the performance of unigram, uni-/bigram, and uni-/bi-/trigram word models. Unigram contain single tokens, bigrams contain two tokens that follow each other, and trigrams are a series of three subsequent tokens in a document. For example, ”loss” is a unigram, ”loss reserve” is a bigram, and ”loss reserve decrease” is a trigram.

[14] See Table A.6 for the Gradient Boosting model and Table A.8 for the Stochastic Gradient Descent model.

[15] Appendix A.6.3 shows the full list of words for each described word category.

We search for the most promising text representation in cross-validation gridsearch that explores all possible combinations of ngram ranges and stop words. We find that a Stochastic Gradient Descent model using uni- and bigrams and ignoring company names, individual names, and verbally expressed numbers is most promising as shown in Table A.9 in Appendix A.6.1. The results of the evaluation of the text representation for the Gradient Boosting model are shown in Appendix A.6.2. The best cross-validation result for the Gradient Boosting model is achieved by using unigrams and no stop word exclusions. However, the best Gradient Boosting model achieves an AUC score more than 1 percentage point lower than that of the best Stochastic Gradient Descent model (scores are 0.6602 and 0.6732, respectively). We thus continue with analyzing the Stochastic Gradient Descent model in the main text and only report the remaining results for the Gradient Boosting model in the appendix.[16] Since we continue by analyzing the best-performing model, the stop word selection decreases the length of the dictionary from 5,205 words to 5,020.

4.5 Integration of financial information

In a last step, we examine whether the combination of the text-based information with financial information can increase prediction performance. Adding financial information provides an economic context for the qualitative disclosures in the MD&A section by allowing to take the financial situation of the company into consideration. Specifically, we test information on profit, growth, business concentration, and financial distress. However, in prediction models, using continuous scales on financial indicators can lead to overfitting and poor out-of sample prediction. Additionally, incorporating an interaction of financial and qualitative information is conceptually difficult with continuous information and likely uninformative. For each of the four types of financial information we not only consider the reported continuous values, but also construct a binary indicator.

The profit indicator is positive in case the company has a net income greater than zero. As such, the indicator separates the profitable from loss-burdened companies. The growth indicator identifies companies with growing direct premiums written. Because the majority of firms reported growing premiums in 2012, we also evaluate a second version of the growth indicator that splits the sample at the growth median. Business concentration is measured using the Herfindahl index and we partition the sample at the median to form the indicator. We create an indicator of financial distress by setting it to one if a company breaches the company action RBC level (RBC ratio $\leq 200\%$). A company that breaches the company action level needs to produce a plan to restore its RBC levels. While this is an intuitive threshold for the indicator, only 16 companies in our sample are below company action level. We thus additionally consider a second version of the RBC ratio indicator that splits the sample at the median of the RBC ratio. The resulting sample splits are summarized in Table 4.

The question how to combine qualitative and financial information in a prediction model is not well explored in financial and accounting research. The most common alternative is a

[16] The results for the integration of financial information can be found in Appendix A.8. The hold-out test set results are shown in Appendix A.10.

Table 4: Financial information variables and indicators

Variable	**0**	**1**	**Total**	**Missing**
Profit				
Profit Indicator	172	549	721	1
Growth				
Growth Indicator (Positive)	175	516	691	31
Growth Indicator (Median)	345	346	691	31
Business concentration				
Business concentration Indicator	347	348	695	27
RBC ratio				
RBC ratio Indicator (Action)	667	16	683	39
RBC ratio Indicator (Median)	341	342	683	39

All numbers are taken from the companies' respective NAIC annual statements for 2012. *Profit* is 1, if the net income is greater than 0. *Growth (Positive)* is 1, if the annual growth rate of direct premiums written from 2011 to 2012 is greater than 0. *Growth (Median)* is 1, if the company's growth rate is above the median. *Business concentration* is 1, if the Herfindahl index, the sum of the squared percentage shares of premiums earned in each of the 45 property and casualty lines of business, is above the median. *RBC ratio (Action)* is 1, if the RBC ratio is below the company action level. *RBC ratio (Median)* is 1, if the RBC ratio is below the sample median.

two-stage approach. In the first stage, only the qualitative information is used in a prediction model. The classification of this prediction model and the financial information are then used as regressors in a standard regression model in the second stage. The advantage of this model is that it provides utilizable standard errors and can thus be used for testing hypotheses (see, e.g., Antweiler and Frank, 2004). However, the primary purpose of such a model is hypothesis testing rather than prediction, which makes it less applicable to our case. We nevertheless test this model as a first alternative for the integration of qualitative and financial information. To make the results of the model comparable to those of typical machine learning classification models, we use a logistic regression for forming binary predictions and test both a model with continuous financial information and one with binary indicators.

The second alternative is a pure prediction model in which both the qualitative information and the quantitative information are used as potential predictors. We test this procedure with the reported continuous variables as well as with the financial indicators. The biggest problem with it is, however, that while some machine learning procedures, such as random forests, can form interaction terms themselves, linear models, such as the stochastic gradient descent model used in this analysis, cannot do so. As such, the prediction model we use here will include both qualitative and financial information if it is useful for prediction, but is unable to assign qualitative information a different meaning based on the financial information. Since taking into account the economic context of the firm was our primary motivation for integrating qualitative and financial information, this second approach is also not fully suitable for our application.

We thus use a novel approach to integrating both types of information as our third alternative. As in the second approach, we include both the qualitative TF-IDF statistics of uni- and bigram word combinations and the financial indicators in our set of potential predictors. In addition, we interact each financial indicator with the TF-IDF statistics and use the resulting matrix of interaction terms as additional predictors in the model.

To determine which of the financial information should be included in the prediction models, we again use a cross-validation grid search. For every possible combination of financial indicators, we report the average AUC score of the prediction models using binary financial indicators with and without interaction terms.[17] Table 5 shows that the highest performance according to the AUC score is achieved when adding information on profit, growth (Median), and the RBC ratio (Median), where the RBC ratio indicator is built based on the median cut-off and the growth indicator is created by partitioning the sample at the median.

Table 5: Selection of financial information model

		Profit	**Growth (Positive)**	**Growth (Median)**	**Concentration**	**RBC ratio (Action)**	**RBC ratio (Median)**	**Average AUC score**	**Rank AUC score**
1.	**1.**	x						0.6334	45
	2.		x					0.6255	62
	3.			x				0.6427	29
	4.				x			0.6265	60
	5.					x		0.6299	53
	6.						x	0.6696	3
2.	**1.**	x	x					0.6341	43
	2.	x		x				0.6468	25
	3.	x			x			0.6278	58
	4.	x				x		0.6377	39
	5.	x					x	0.6647	7
	6.		x		x			0.6205	68
	7.		x			x		0.6284	56
	8.		x				x	0.6614	13
	9.			x	x			0.6425	31
	10.			x		x		0.6399	37
	11.			x			x	0.6689	5
	12.				x	x		0.6247	64
	13.				x		x	0.6633	9
3.	**1.**	x	x		x			0.6234	66
	2.	x	x			x		0.6321	51
	3.	x	x				x	0.6594	17
	4.	x		x	x			0.6417	33

[17] Note that we did not consider models which included multiple binary indicators for a single piece of financial information (such as including both the above median and positive growth indicators in a single model).

Table 5: Selection of financial information model (continued)

		Profit	**Growth (Positive)**	**Growth (Median)**	**Concentration**	**RBC ratio (Action)**	**RBC ratio (Median)**	**Average test score**	**Rank test score**
	5.	x		x		x		0.6462	27
	6.	x		x			x	0.6705	1
	7.	x			x	x		0.6328	47
	8.	x			x		x	0.6567	21
	9.		x		x	x		0.6157	70
	10.		x		x		x	0.6579	19
	11.			x	x	x		0.6345	41
	12.			x	x		x	0.6602	15
4.	**1.**	x	x		x	x		0.6323	49
	2.	x	x		x		x	0.6533	23
	3.	x		x	x	x		0.6407	35
	4.	x		x	x		x	0.6619	11
Baseline								0.6286	55

The table summarizes the AUC scores of the ten-fold cross-validation prediction of the alternative classification models with TF-IDF uni- and bigram word tokens of the MD&A sections. The AUC estimates the probability that a random positive is ranked before a random negative (in our model, the case of underreserving), without specifying a particular decision threshold. The columns *Profit* to *RBC ratio (Median)* indicate which financial information is included. The models are specified in two variants: only with indicators and with indicators and interaction terms. The average of the test scores of both variants is reported in the column *Average AUC score*. The column *Rank AUC score* indicates the rank of a specific average AUC score. The *Baseline* category shows the results without financial information.

In Table 6, the alternative models are evaluated with a comparison of cross-validation results on the training set. In this comparison, we can see that the models that using qualitative and quantitative information in a single stage returns the most accurate predictions according to the AUC. This renders credibility to the idea that financial and qualitative information should be used in conjecture when making predictive statements. The two-stage models, while more useful for hypothesis testing, can not perform as well. The table also shows, as conjectured, that continuous financial information likely leads to worse predictions than using binary indicators. This is true for both the two-stage model (model 1.1 in Table 6) and when integrating financial and qualitative information in a single stage (model 2 in Table 6). We nevertheless evaluate the performance of all five models in Table 6 on the test set in Section 5, such that their value for out-of-sample predictions is also assessed.

Table 6: Classification results for models integrating financial information

		Accuracy	**Precision**	**F1**	**AUC**
(1.1)	Two-stage model with reported continuous information	0.6975	0.6997	0.8194	0.5132
(1.2)	Two-stage model with financial indicators	0.6860	0.7477	0.7854	0.5933
(2)	Prediction model with MD&A and reported continuous information	0.5026	0.6487	0.5587	0.4738
(3)	Combination of MD&A and financial indicators	0.6802	0.7520	0.7786	0.6723
(4)	Combination of MD&A, financial indicators and interaction terms	0.6764	0.7402	0.7780	0.6630

The table summarizes the average performance scores of a ten-fold cross-validation on the training set of the alternative models. Accuracy is defined as the share of correct predictions. Precision measures the share of true positives (that is, the case of overreserving) among all predicted positives. The F1 score is calculated as the harmonic mean of precision and recall, which measures the share of predicted positives among all actual positives. The AUC estimates the probability that a random positive is ranked before a random negative, without specifying a particular decision threshold. Model (1) is a classification based on a logistic regression model with financial information and the predicted sign of the reserve error of Stochastic Gradient Descent model using MD&A information. Model variant (1.1) uses reported continuous information, whereas variant (1.2), includes financial indicators. Model (2) is a Stochastic Gradient Descent classifier with MD&A information and reported continuous information. Model (3) is a Stochastic Gradient Descent classifier with MD&A information and financial indicators. Model (4) is a Stochastic Gradient Descent classifier with MD&A information and financial indicators and interaction terms. MD&A information represents TF-IDF uni- and bigram word combinations of MD&A sections. Reported continuous information summarizes continuous data on net income, growth of direct premiums written, and RBC ratio. Used financial indicators are profit, growth (Median) and RBC ratio (Median) indicators.

4.6 Robustness: Alternative text representation with unsupervised learning model

The entire process of model selection has, so far, only considered supervised machine learning models. Alternatively, text can also be decomposed into topics instead of word combinations by using unsupervised learning models. To test whether this approach renders better predictions than our model, we evaluate an unsupervised machine learning model as alternative text representation model. Specifically, we test whether a Latent Dirichlet Allocation (LDA) topic model can outperform the Stochastic Gradient Decent model.

The idea of an LDA topic model is that ”documents are represented as random mixtures over latent topics, where each topic is characterized by a distribution over words” (Blei et al., 2003). Here, the weighting of TF-IDF is not necessary, since the LDA is a probabilistic model that tries to estimate probability distributions for topics in documents and words in topics. The LDA topic model requires only few parameter inputs. One key assumption is that there exists a finite number of topics and that every document consists of a mix of these topics. It is thus important to determine the appropriate number of topics and we use a cross-validation approach to determine it from a set of possible values (1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 25, 50, 75, 100, 125, 150, 175, 200, 225, 250).

We approach the optimization from two directions. First, as suggested by Blei et al. (2003), we compare the ”perplexity” of topic models with different numbers of topics. A lower perplexity score indicates a better generalization performance. Since perplexity is calculated as the inverse of the geometric mean per-word likelihood, it monotonically decreases with the likelihood. Thus, we select the number of topics that maximizes the log-likelihood. Then, we use an LDA model with the best identified number of topics (3) as input for the earnings management classification model. However, this model only achieves an average AUC of 49 percent, when evaluated in a cross-validation classification with the Stochastic Gradient Descent model on the training set.

Next, we explore a different perspective and search for the best number of topics as part of our overall classification model. That is, we look for the number of topics that optimizes the performance of the earnings management classification model. This model achieves an AUC score of 61 percent with 150 topics as optimal model parameter (see Appendix A.7 for detailed results). However, the model using TF-IDF weights and uni- and bigrams still yields a higher performance with an AUC score of 67 percent. Therefore, we do not use the unsupervised topic model for the prediction.

5 Results

5.1 Prediction results on the hold-out test set

We evaluate the final classification model on a hold-out test sample and report the results in Table 7. The final classification model is a Stochastic Gradient Descent model that uses TF-IDF statistics of uni- and bigram word combinations in the MD&A section together with profit, growth, and financial distress information on the company. The model variant with financial interaction terms achieves a 70%-accuracy of the predicted sign of the company's reserve error and achieves an AUC score of 62 percent. The corresponding values for the model without interaction terms are a lower accuracy of 68% but a higher AUC score of 64%.[18] [19] Our results confirm that the MD&A section published by a property and casualty insurer is indicative of the insurer's reserve error and, hence, the company's fundamental earnings. Managers thus seem unable to remove all indications of earnings management from their qualitative disclosures.

The test set evaluations of Table 7 also allow for a comparison of the quality of the out-of sample predictions by the different analyzed models. As with the in-sample predictions, we can see that binary indicators of financial information lead to significantly better predictions than continuous variables. Additionally, we can see that the two-stage model (1.2) still performs worse than the integrated models (3 and 4), but that the gap is smaller. In both accuracy and AUC, the two-stage model always performs similar to the worse of the two preferred models. Nevertheless, when considering both in-sample and out-of-sample predictions, the integrated approach seems more promising for predictive accuracy than the two-stage analysis. It needs to be emphasized though that this result only applies to predictions and not for other purposes such as hypothesis testing, where the two-stage approach has significant advantages.

[18] The Gradient Boosting model with financial information achieves a accuracy of 70 percent, but only a AUC score of 55 percent. The test set classification results for the Gradient Boosting model are reported in Appendix A.10.

[19] If we alternatively split the target variable at the median of the reserve error, the AUC score of 61 percent in a model with interaction terms and 62 percent in a model without interaction term.

Table 7: Classification results for models integrating financial information on test set

		Value Target Variable	**Precision**	**Recall**	**F1**	**Accuracy**	**AUC**
(1.1)	Two-stage model with reported continuous information	0	0.0000	0.0000	0.0000		
		1	0.6899	0.9889	0.8128	0.6846	0.4944
		Weighted average	0.3450	0.4944	0.4064		
(1.2)	Two-stage model with financial indicators	0	0.4865	0.4500	0.4675		
		1	0.7634	0.7889	0.7760	0.6846	0.6194
		Weighted average	0.6782	0.6846	0.6811		
(2)	Prediction model with MD&A and reported continuous information	0	0.2841	0.6250	0.3906		
		1	0.6429	0.3000	0.4091	0.4000	0.4625
		Weighted average	0.5325	0.4000	0.4034		
(3)	Combination of MD&A and financial indicators	0	0.4884	0.5250	0.5060		
		1	0.7816	0.7556	0.7684	0.6846	0.6403
		Weighted average	0.6914	0.6846	0.6876		
(4)	Combination of MD&A, financial indicators and interaction terms	0	0.5161	0.4000	0.4507		
		1	0.7576	0.8333	0.7937	0.7000	0.6167
		Weighted average	0.6369	0.6167	0.6222		

The table summarizes the performance scores on the hold-out test set of the alternative models. Accuracy is defined as the share of correct predictions. Precision measures the share of true positives (that is, the case of overreserving) among all predicted positives. The F1 score is calculated as the harmonic mean of precision and recall, which measures the share of predicted positives among all actual positives. The AUC estimates the probability that a random positive is ranked before a random negative, without specifying a particular decision threshold. The weighted average calculates the weighted mean per label. Model (1) is a classification based on a logistic regression model with financial information and the predicted sign of the reserve error of Stochastic Gradient Descent model using MD&A information. Model variant (1.1) uses reported continuous information, whereas variant (1.2), includes financial indicators. Model (2) is a Stochastic Gradient Descent classifier with MD&A information and reported continuous information. Model (3) is a Stochastic Gradient Descent classifier with MD&A information and financial indicators. Model (4) is a Stochastic Gradient Descent classifier with MD&A information and financial indicators and interaction terms. MD&A information represents TF-IDF uni- and bigram word combinations of MD&A sections. Reported continuous information summarizes continuous data on net income, growth of direct premiums written, and RBC ratio. Used financial indicators are profit, growth (Median) and RBC ratio (Median) indicators.

Table 7 also shows that all tested classification models are significantly better at predicting overreserving (when the value of the target variable is 1) than they are at predicting underreserving (when the target variable is 0). Between 76 and 78 percent of the firms predicted to overreserve by our preferred prediction models, actually overstate their reserves (precision). Further, the share of predicted overstated reserves among all actual overstated reserves by these models are between 76 and 83 percent (recall). The corresponding values for underreserving firms are lower, but still better (on average) for our preferred models than for all other ones tested.

Do our preferred models provide a good prediction result? As of yet, only few studies in the accounting literature have used text-based machine learning classification. In most textual analyses, the similarity of documents is used to create a new variable for statistic regressions. For studies with a machine learning classification, often only accuracy is reported, and it ranges between 50 to 84 percent (Antweiler and Frank, 2004; Li, 2010; Humpherys et al., 2011; Kang et al., 2013). However, in some studies, the model is either not evaluated on a hold-out test set, which may cause overfitting, or it is not clear whether a hold-out set has been used to derive the final prediction scores. For our preferred models, the accuracy score lies within the range of values observed in the literature. This is true both when evaluated on the training set and when evaluated on the test set. In addition, the models' AUC scores lie well above 50% (the value for an uninformative model) in both evaluations. We thus argue that our preferred models provide a good prediction. The qualitative disclosure in the MD&A section is informative of the reserve error and, thus, for fundamental earnings.

As described in section 4.5, we use a model that integrates financial information with the qualitative information of MD&A sections. The methodological contribution provided here is that in our approach the linear classification models are capable of embedding qualitative information in the economic context in which it is collected. As can be seen in Table 7, the procedure with financial indicators as well as the procedure with financial indicators and interaction terms provide the best predictive performance on the test set of all models which integrate qualitative and financial information. This is particularly true when considering the AUC criterion, which we have used throughout the entire model selection procedure.

5.2 Qualitative information in the MD&A

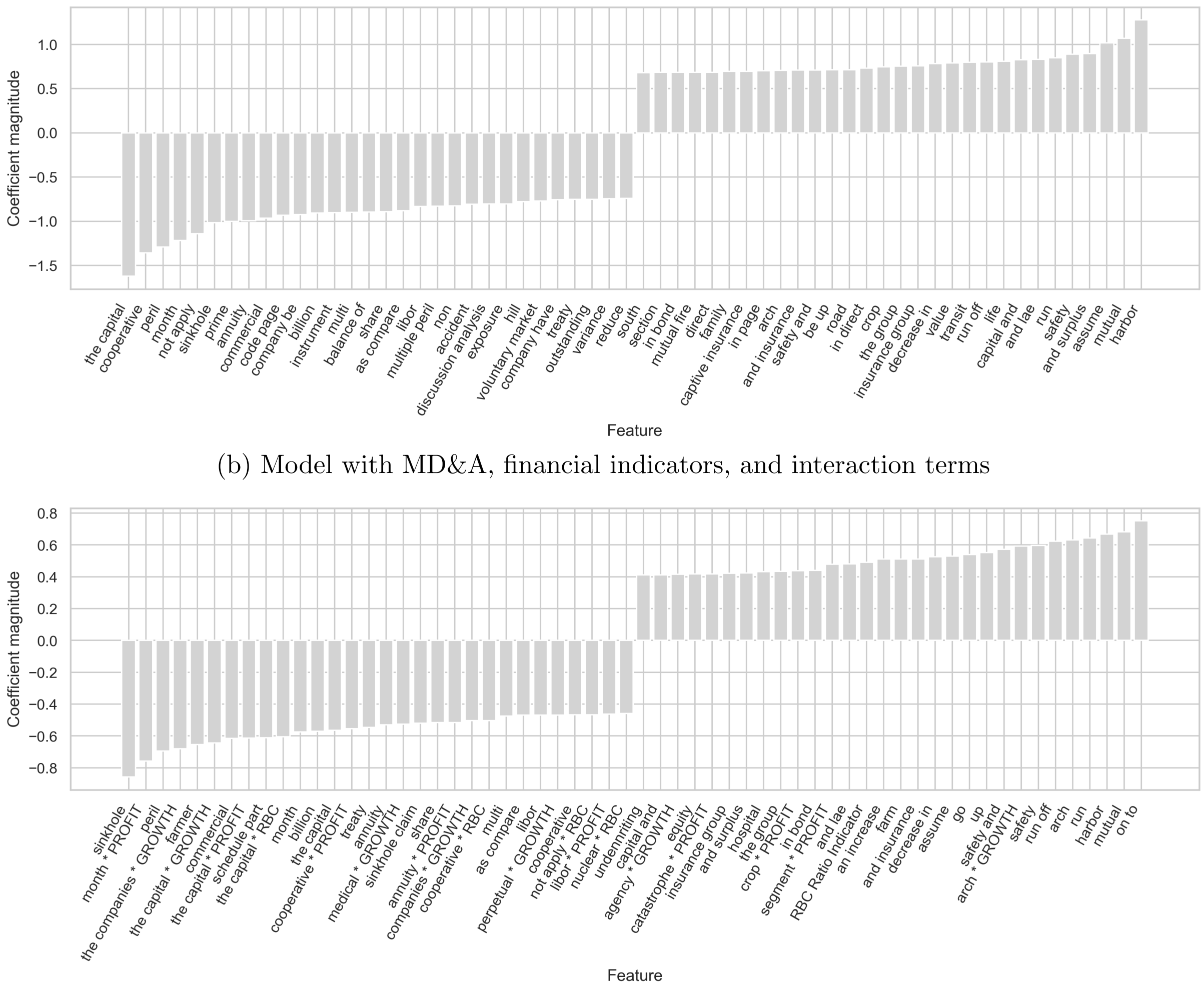
Our final model contains a total of 21,065 uni- and bigram word combinations, which leads to 84,263 features after including the financial indicators and interaction terms with the word combinations. To examine how the model creates the prediction, we can inspect the most important features in the model. Panel (a) in Figure 5 shows the features with the top and bottom 30 coefficients for the model with MD&A and financial indicators and panel (b) shows the same for the model with MD&A, financial indicators and interaction terms.[20] We can see that some words that indicate financial threats, such as "peril", "sinkhole", "multiple peril", and "exposure" have strong negative coefficients. On the contrary, expressions for safety, such as

[20] We excluded locations from this analysis of the word features, since they do not relate to the managers' wording choices, but are a result of the business context of the firm. An representation including locations is shown in Appendix A.9.

"harbor", "safety", "safety and", and "arch", as well as forward-looking statements like "run", "be up", and "road" exhibit high positive coefficients. Also information on the "mutual" and "group" company form are present among the most positive coefficients.

Figure 5: Most important feature coefficients in the prediction model with MD&A and financial indicators

(a) Model with MD&A and financial indicators



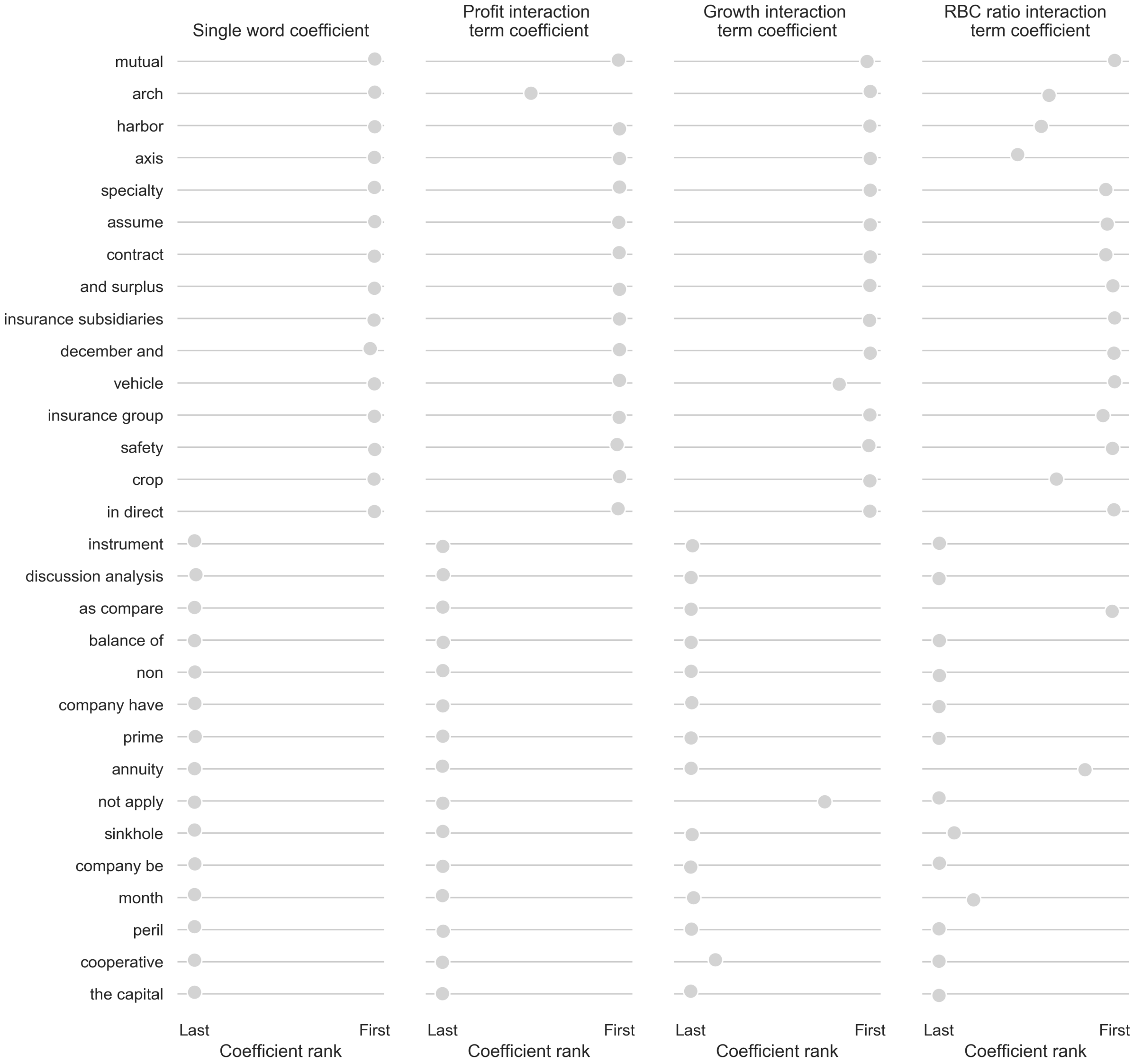
(b) Model with MD&A, financial indicators, and interaction terms

The figure shows features with the 30 largest and the 30 smallest coefficients for two classification models for over- and underreserving. Both models use a Stochastic Gradient Descent classifier with TF-IDF uni- and bigram word. Models include financial indicators for profit, growth (Median), and RBC ratio (Median). Word features that contain information on locations and abbreviations are not shown in the figure.

In the model with interaction terms, again words that indicate financial threats have strong negative coefficients. In contrast to the model without interaction terms, also the information on the business type, "farmer", "commercial", "treaty" or "medical" exhibit strong negative coefficients. Moreover, combinations with "of the capital" bear strong negative coefficients. Among the strongest positive coefficient, the emerging scheme is very similar to the model without interaction terms. Forward-looking statements, expressions for safety as well as information on "mutual" and "group" have strong positive coefficients.

In a next step, we evaluate whether individual words are similarly important among the subsamples created by the financial indicators. Figure 6 illustrates the top 15 and bottom 15 aggregated coefficients of word combinations. The aggregated coefficient of a word combination is the result from taking into account that each word combination is also part of the interaction terms with the financial indicators. We can see that the importance of the single words as well as the profit and growth interaction terms are similar, while the RBC ratio interaction terms seem to value some words in a different manner. For instance, the word ”harbor” has a strong positive coefficient as a stand-alone word as well as in the case of a profitable, growing company. However, for companies with an higher RBC ratio, the word ”harbor” is associated with a smaller coefficient.

Figure 6: Feature coefficient rankings of single word tokens and interaction terms



The figure shows word combinations with the 15 largest and the 15 smallest aggregated coefficients for the classification model for over- and underreserving, that uses a Stochastic Gradient Descent classifier with TF-IDF uni- and bigram word combinations of MD&A sections and profit, growth (Median), and RBC ratio (Median) indicators and interaction terms with the word tokens. Aggregated coefficient magnitude is calculated as sum of coefficients of single words and the interaction terms of the specific word with profit, growth (Median), and RBC ratio (Median). The figure shows in the first panel the coefficient magnitude for the single word and the other panels the coefficient maghnitude for the interaction terms.

Comparing the top 15 and bottom 15 aggregated coefficients in Figure 6 with the top 30 and bottom 30 individual coefficients in Figure 5, a common scheme emerges: Combinations with ”mutual” and expressions for safety exhibit strong positive coefficients, while ”capital” and words for financial threats bear strong negative coefficients. This link is not only provided by the statistical model, but also makes sense from an intuitive perspective. Expressions of safety are used because the financial situation of the firm is safe (due to untapped reserves). Words for financial threats imply a threatened situation because there are unpaid liabilities in the balance sheet.

6 Conclusion

In this study, we examine the association between a company's fundamental earnings and the qualitative information in the MD&A section. Our conceptual model is based on the earnings management process within a company and exhibits the role of qualitative information disclosed in the MD&A section. Managers seek to allow only a link between reported earnings and the MD&A section and to cover any direct indication on fundamental earnings in the MD&A section, that would reveal accrual-based earnings management. We use a new machine learning approach to predict earnings management based on the disclosed qualitative information and apply it to property and casualty insurers. Using machine learning techniques on the MD&A section, we document that the qualitative disclosure is indicative of an insurer's reserve error, the industry-specific earnings management measure. This suggests that managers are unable to cover earnings management in the qualitative disclosure. Our results show that it is possible to link the information in the MD&A section directly to fundamental earnings.

Our findings suggest that the qualitative information in the MD&A section has an informational value, as opposed to previous doubts whether the disclosure provides useful information (Brown and Tucker, 2011). Regarding the MD&A section of 10-K filings, the SEC has repeatedly urged firms to reduce needless information and provide helpful information for investors (SEC, 2003, 2013). Our findings underpin that the MD&A section indeed helps investors to evaluate a company's performance by gaining information about fundamental earnings. Benefiting from machine learning applications, company stakeholders can make better-informed decisions.

We contribute to the growing literature on machine learning in accounting research. For this purpose, we analyze different approaches for integrating financial information and qualitative disclosures in a prediction model. This includes a novel approach which allows the interpretation of qualitative information in the economic context of the firm even in linear models. We find that models using a single stage and integrating both text processing and financial information in the cross-validation procedure lead to better predictions than models doing so in a two-stage approach. In the specific field of earnings management, we are the first to utilize qualitative information for predicting the reserve error of property and casualty insurance companies. While our results are industry-specific, the approach can be applied to other industry-specific measures or be generalized to cross-industry indicators of earnings management.

Our study has certain limitations. First, our analysis develops a prediction model and does not focus on causal inference. We can thus not answer why managers choose to manager their earnings or what the specific employed disclosure strategies are. We simply provide a test of the link between qualitative disclosures and earnings management. Future research can utilize this link to investigate causal research questions of earnings management, strategic disclosure and managerial behavior. Second, we provide a stepwise introduction of our machine learning model, using cross-validation grid-search algorithms in several stages from model selection, over parameter tuning, text preprocessing and finally the integration of financial information. It is theoretically likely that the predictive power of our model would be increased by integrating all these steps into a single grid-search. However, as of now, this approach is not computationally feasible. Lastly, we only briefly cover the class of unsupervised learning models, such that we may not have investigated their full potential. Nevertheless, the model which we did test had a relatively poor performance with an AUC score 6 percentage points lower than the best supervised model after parameter tuning. There is thus a reasonable indication that supervised learning models perform better in the prediction of earnings management than their unsupervised counterparts.

References

Antweiler, W. and Frank, M. Z. (2004). Is all that talk just noise? The information content of internet stock message boards. Journal of Finance, 59(3):1259–1294.

Armstrong, C. S., Guay, W. R., and Weber, J. P. (2010). The role of information and financial reporting in corporate governance and debt contracting. Journal of Accounting and Economics, 50(2-3):179–234.

Baker, M. and Wurgler, J. (2013). Behavioral corporate finance: An updated survey. In Handbook of the Economics of Finance, volume 2, pages 357–424. Elsevier.

Bao, Y. and Datta, A. (2014). Simultaneously discovering and quantifying risk types from textual risk disclosures. Management Science, 60(6):1371–1391.

Bartov, E. (1993). The timing of asset sales and earnings manipulation. Accounting Review, pages 840–855.

Beaver, W. H., McNichols, M. F., and Nelson, K. K. (2003). Management of the loss reserve accrual and the distribution of earnings in the property-casualty insurance industry. Journal of Accounting and Economics, 35(3):347–376.

Bergstresser, D., Desai, M., and Rauh, J. (2006). Earnings manipulation, pension assumptions, and managerial investment decisions. Quarterly Journal of Economics, 121(1):157–195.

Bergstresser, D. and Philippon, T. (2006). CEO incentives and earnings management. Journal of Financial Economics, 80(3):511–529.

Bhojraj, S., Hribar, P., Picconi, M., and McInnis, J. (2009). Making sense of cents: An examination of firms that marginally miss or beat analyst forecasts. Journal of Finance, 64(5):2361–2388.

Billett, M. T. and Qian, Y. (2008). Are overconfident CEOs born or made? Evidence of self-attribution bias from frequent acquirers. Management Science, 54(6):1037–1051.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. Journal of Machine Learning Research, 3(Jan):993–1022.

Bloomfield, R. (2008). Discussion of "Annual report readability, current earnings, and earnings persistence". Journal of Accounting and Economics, 45(2-3):248–252.

Bozanic, Z., Roulstone, D. T., and Van Buskirk, A. (2018). Management earnings forecasts and other forward-looking statements. Journal of Accounting and Economics, 65(1):1–20.

Brown, S. V. and Tucker, J. W. (2011). Large-sample evidence on firms' year-over-year md&a modifications. Journal of Accounting Research, 49(2):309–346.

Burgstahler, D. and Dichev, I. (1997). Earnings management to avoid earnings decreases and losses. Journal of Accounting and Economics, 24(1):99–126.

Cornett, M. M., Marcus, A. J., and Tehranian, H. (2008). Corporate governance and pay-for-performance: The impact of earnings management. Journal of Financial Economics, 87(2):357–373.

Dechow, P., Ge, W., and Schrand, C. (2010). Understanding earnings quality: A review of the proxies, their determinants and their consequences. Journal of Accounting and Economics, 50(2-3):344–401.

Dechow, P. M. and Skinner, D. J. (2000). Earnings management: Reconciling the views of accounting academics, practitioners, and regulators. Accounting Horizons, 14(2):235–250.

Dechow, P. M. and Sloan, R. G. (1991). Executive incentives and the horizon problem: An empirical investigation. Journal of Accounting and Economics, 14(1):51–89.

Degeorge, F., Patel, J., and Zeckhauser, R. (1999). Earnings management to exceed thresholds. Journal of Business, 72(1):1–33.

Ding, K., Lev, B., Peng, X., Sun, T., and Vasarhelyi, M. A. (2020). Machine learning improves accounting estimates: evidence from insurance payments. Review of Accounting Studies, 25:1–37.

Dyer, T., Lang, M., and Stice-Lawrence, L. (2017). The evolution of 10-K textual disclosure: Evidence from Latent Dirichlet Allocation. Journal of Accounting and Economics, 64(2-3):221–245.

Eckles, D. L. and Halek, M. (2010). Insurer reserve error and executive compensation. Journal of Risk and Insurance, 77(2):329–346.

Eckles, D. L., Halek, M., He, E., Sommer, D. W., and Zhang, R. (2011). Earnings smoothing, executive compensation, and corporate governance: Evidence from the property–liability insurance industry. Journal of Risk and Insurance, 78(3):761–790.

Efendi, J., Srivastava, A., and Swanson, E. P. (2007). Why do corporate managers misstate financial statements? The role of option compensation and other factors. Journal of Financial Economics, 85(3):667–708.

Ferri, C., Hernández-Orallo, J., and Flach, P. A. (2011). A coherent interpretation of AUC as a measure of aggregated classification performance. In Proceedings of the 28th International Conference on Machine Learning (ICML-11), pages 657–664.

Fields, T. D., Lys, T. Z., and Vincent, L. (2001). Empirical research on accounting choice. Journal of Accounting and Economics, 31(1-3):255–307.

Folsom, D., Hribar, P., Mergenthaler, R. D., and Peterson, K. (2017). Principles-based standards and earnings attributes. Management Science, 63(8):2592–2615.

Frankel, R., Mayew, W. J., and Sun, Y. (2010). Do pennies matter? Investor relations consequences of small negative earnings surprises. Review of Accounting Studies, 15(1):220–242.

Gaver, J. J. and Paterson, J. S. (2004). Do insurers manipulate loss reserves to mask solvency problems? Journal of Accounting and Economics, 37(3):393–416.

Gentzkow, M., Kelly, B., and Taddy, M. (2019). Text as data. Journal of Economic Literature, 57(3):535–74.

Grace, M. F. and Leverty, J. T. (2010). Political cost incentives for managing the property-liability insurer loss reserve. Journal of Accounting Research, 48(1):21–49.

Grace, M. F. and Leverty, J. T. (2012). Property–liability insurer reserve error: Motive, manipulation, or mistake. Journal of Risk and Insurance, 79(2):351–380.

Graham, J. R., Harvey, C. R., and Rajgopal, S. (2005). The economic implications of corporate financial reporting. Journal of Accounting and Economics, 40(1-3):3–73.

Grice, H. P. (1989). Studies in the Way of Words. Harvard University Press.

Guay, W., Samuels, D., and Taylor, D. (2016). Guiding through the fog: Financial statement complexity and voluntary disclosure. Journal of Accounting and Economics, 62(2-3):234–269.

Hanley, J. A. and McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology, 143(1):29–36.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). The elements of statistical learning: Data mining, inference, and prediction. Springer Science & Business Media.

Hazarika, S., Karpoff, J. M., and Nahata, R. (2012). Internal corporate governance, CEO turnover, and earnings management. Journal of Financial Economics, 104(1):44–69.

Healy, P. M. and Wahlen, J. M. (1999). A review of the earnings management literature and its implications for standard setting. Accounting Horizons, 13(4):365–383.

Hoberg, G. (2016). Discussion of using unstructured and qualitative disclosures to explain accruals. Journal of Accounting and Economics, 62(2):228–233.

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

Hoberg, G. and Maksimovic, V. (2014). Redefining financial constraints: A text-based analysis. Review of Financial Studies, 28(5):1312–1352.

Hoberg, G. and Phillips, G. (2016). Text-based network industries and endogenous product differentiation. Journal of Political Economy, 124(5):1423–1465.

Huang, A. H., Lehavy, R., Zang, A. Y., and Zheng, R. (2018). Analyst information discovery and interpretation roles: A topic modeling approach. Management Science, 64(6):2833–2855.

Humpherys, S. L., Moffitt, K. C., Burns, M. B., Burgoon, J. K., and Felix, W. F. (2011). Identification of fraudulent financial statements using linguistic credibility analysis. Decision Support Systems, 50(3):585–594.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). An introduction to statistical learning, volume 112. Springer Science & Business Media.

Kang, J. S., Kuznetsova, P., Luca, M., and Choi, Y. (2013). Where not to eat? Improving public policy by predicting hygiene inspections using online reviews. In Proceedings of the 2013 conference on empirical methods in natural language processing, pages 1443–1448.

Kothari, S. P. and Sloan, R. G. (1992). Information in prices about future earnings: Implications for earnings response coefficients. Journal of Accounting and Economics, 15(2-3):143–171.

Lang, M. and Lundholm, R. (1993). Cross-sectional determinants of analyst ratings of corporate disclosures. Journal of Accounting Research, 31(2):246–271.

Lang, M. and Stice-Lawrence, L. (2015). Textual analysis and international financial reporting: Large sample evidence. Journal of Accounting and Economics, 60(2-3):110–135.

Larcker, D. F. and Zakolyukina, A. A. (2012). Detecting deceptive discussions in conference calls. Journal of Accounting Research, 50(2):495–540.

Leuz, C., Nanda, D., and Wysocki, P. D. (2003). Earnings management and investor protection: An international comparison. Journal of Financial Economics, 69(3):505–527.

Li, F. (2008). Annual report readability, current earnings, and earnings persistence. Journal of Accounting and Economics, 45(2-3):221–247.

Li, F. (2010). The information content of forward-looking statements in corporate filings—a naïve bayesian machine learning approach. Journal of Accounting Research, 48(5):1049–1102.

Lopez-Lira, A. (2019). Risk factors that matter: Textual analysis of risk disclosures for the cross-section of returns. Available at SSRN 3313663.

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

Louis, H. (2004). Earnings management and the market performance of acquiring firms. Journal of Financial Economics, 74(1):121–148.

McCornack, S. A. (1992). Information manipulation theory. Communications Monographs, 59(1):1–16.

McKinsey & Company (2020). Report on State of property and casualty insurance 2020. `https://www.mckinsey.com/industries/financial-services/our-insights/state-of-property-and-casualty-insurance-2020`, (accessed: 07/15/2020).

McNichols, M. F. (2000). Research design issues in earnings management studies. Journal of Accounting and Public Policy, 19(4-5):313–345.

Modigliani, F. and Miller, M. H. (1958). The cost of capital, corporation finance and the theory of investment. American Economic Review, 48(3):261–297.

Müller, A. C. and Guido, S. (2016). Introduction to Machine Learning with Python: A Guide for Data Scientists. O'Reilly Media, Inc.

Muslu, V., Radhakrishnan, S., Subramanyam, K., and Lim, D. (2014). Forward-looking MD&A disclosures and the information environment. Management Science, 61(5):931–948.

National Association of Insurance Commissioners (NAIC) (2019). RISK-BASED CAPITAL. `https://content.naic.org/cipr_topics/topic_risk_based_capital.htm`, (accessed: 05/25/2020).

Petroni, K. R. (1992). Optimistic reporting in the property-casualty insurance industry. Journal of Accounting and Economics, 15(4):485–508.

Rabinowicz, A. and Rosset, S. (2020). Cross-validation for correlated data. Journal of the American Statistical Association, (just-accepted):1–38.

Roychowdhury, S. (2006). Earnings management through real activities manipulation. Journal of Accounting and Economics, 42(3):335–370.

Salton, G. and McGill, M. J. (1986). Introduction to Modern Information Retrieval. McGraw-Hill, Inc.

Schipper, K. (1989). Earnings management. Accounting Horizons, 3(4):91.

Scholes, M. S., Wilson, G. P., and Wolfson, M. A. (1990). Tax planning, regulatory capital planning, and financial reporting strategy for commercial banks. Review of Financial Studies, 3(4):625–650.

Schrand, C. M. and Walther, B. R. (2000). Strategic benchmarks in earnings announcements: The selective disclosure of prior-period earnings components. Accounting Review, 75(2):151–177.

Schütze, H., Manning, C. D., and Raghavan, P. (2008). Introduction to information retrieval, volume 39. Cambridge University Press Cambridge.

SEC (2003). Interpretation: Commission guidance regarding management's discussion and analysis of financial condition and results of operations. Technical report, Securities and Exchange Commission. `http://www.sec.gov/rules/interp/33-8350.htm`, (accessed: 05/25/2020).

SEC (2013). Report on Review of Disclosure Requirements in Regulation S-K. Technical report, Securities and Exchange Commission. `https://www.sec.gov/news/studies/2013/reg-sk-disclosure-requirements-review.pdf`, (accessed: 05/25/2020).

Shannon, C. E. (1948). A mathematical theory of communication. Bell System Technical Journal, 27(3):379–423.

Shrieves, R. E. and Dahl, D. (2003). Discretionary accounting and the behavior of japanese banks under financial duress. Journal of Banking & Finance, 27(7):1219–1243.

Teoh, S. H., Welch, I., and Wong, T. J. (1998a). Earnings management and the long-run market performance of initial public offerings. Journal of Finance, 53(6):1935–1974.

Teoh, S. H., Welch, I., and Wong, T. J. (1998b). Earnings management and the underperformance of seasoned equity offerings. Journal of Financial Economics, 50(1):63–99.

Tetlock, P. C. (2007). Giving content to investor sentiment: The role of media in the stock market. Journal of Finance, 62(3):1139–1168.

A Internet appendix

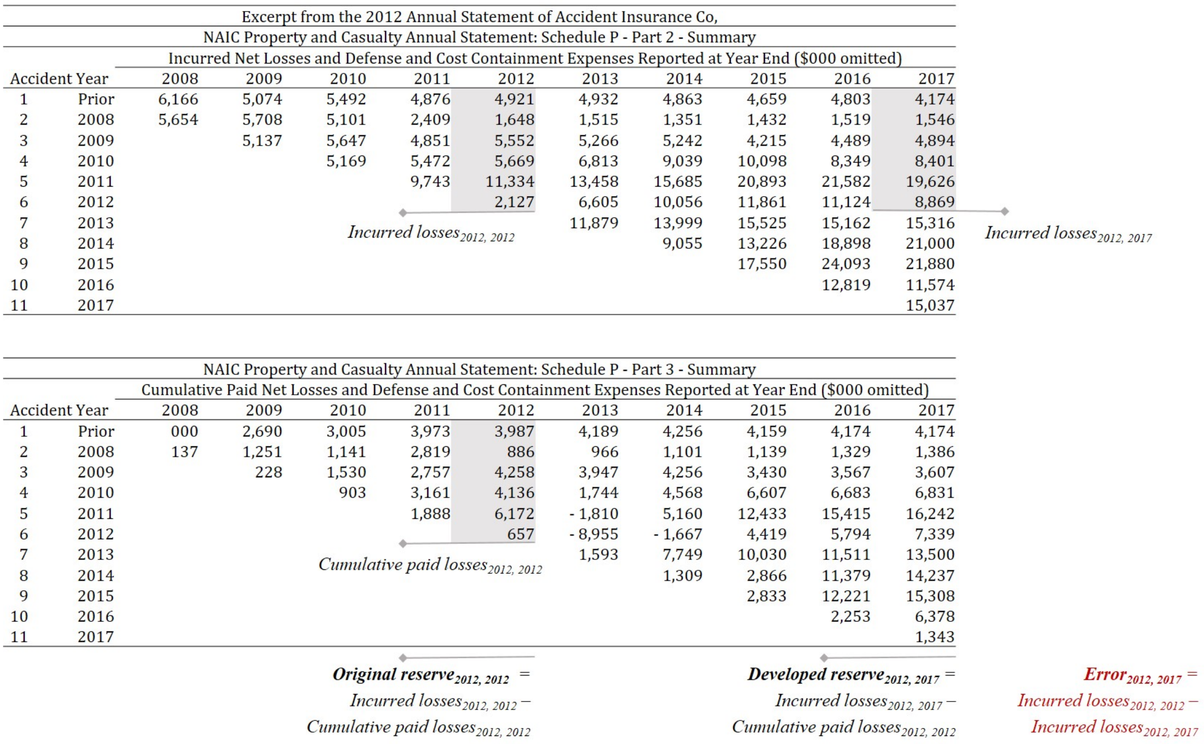
A.1 Actuarial background on loss reserve

The original loss reserve for a specific year is defined as the estimated total incurred losses in the calendar year less the cumulated paid losses in that year. The corresponding developed reserve describes the revision of the total incurred losses from the original calendar year after a development window minus the cumulated paid losses in the original calendar year. Both the original and the developed loss reserve are reported to the NAIC in the Schedule P - Part 2 and 3 of the company's annual statements. An example for a Schedule P with annotations is shown in Figure A.1. The loss reserve error is the difference between original and developed reserve. Thus, the loss reserve error for a company i is calculated as the total incurred losses in a calendar year t minus the revised estimate of the incurred losses in t reported in the calendar year t+n:

$$Error_{i,t} = Incurred\ losses_{i,t,t} - Incurred\ losses_{i,t,t+n} \tag{2}$$

An insurer underreserved in case the original loss reserve was less than the developed reserve. Vice versa, the insurer overreserved in case the original loss reserve was greater than the developed reserve.

Figure A.1: Example for Annual Statement Schedule P



Excerpt from the 2012 Annual Statement of Accident Insurance Co,

NAIC Property and Casualty Annual Statement: Schedule P - Part 2 - Summary

Accident Year		Incurred Net Losses and Defense and Cost Containment Expenses Reported at Year End ($000 omitted)									
		2008	2009	2010	2011	2012	2013	2014	2015	2016	2017
1	Prior	6,166	5,074	5,492	4,876	4,921	4,932	4,863	4,659	4,803	4,174
2	2008	5,654	5,708	5,101	2,409	1,648	1,515	1,351	1,432	1,519	1,546
3	2009		5,137	5,647	4,851	5,552	5,266	5,242	4,215	4,489	4,894
4	2010			5,169	5,472	5,669	6,813	9,039	10,098	8,349	8,401
5	2011				9,743	11,334	13,458	15,685	20,893	21,582	19,626
6	2012					2,127	6,605	10,056	11,861	11,124	8,869
7	2013						11,879	13,999	15,525	15,162	15,316
8	2014							9,055	13,226	18,898	21,000
9	2015								17,550	24,093	21,880
10	2016									12,819	11,574
11	2017										15,037

Incurred losses$_{2012,\ 2012}$

Incurred losses$_{2012,\ 2017}$

NAIC Property and Casualty Annual Statement: Schedule P - Part 3 - Summary

Accident Year		Cumulative Paid Net Losses and Defense and Cost Containment Expenses Reported at Year End ($000 omitted)									
		2008	2009	2010	2011	2012	2013	2014	2015	2016	2017
1	Prior	000	2,690	3,005	3,973	3,987	4,189	4,256	4,159	4,174	4,174
2	2008	137	1,251	1,141	2,819	886	966	1,101	1,139	1,329	1,386
3	2009		228	1,530	2,757	4,258	3,947	4,256	3,430	3,567	3,607
4	2010			903	3,161	4,136	1,744	4,568	6,607	6,683	6,831
5	2011				1,888	6,172	- 1,810	5,160	12,433	15,415	16,242
6	2012					657	- 8,955	- 1,667	4,419	5,794	7,339
7	2013						1,593	7,749	10,030	11,511	13,500
8	2014							1,309	2,866	11,379	14,237
9	2015								2,833	12,221	15,308
10	2016									2,253	6,378
11	2017										1,343

Cumulative paid losses$_{2012,\ 2012}$

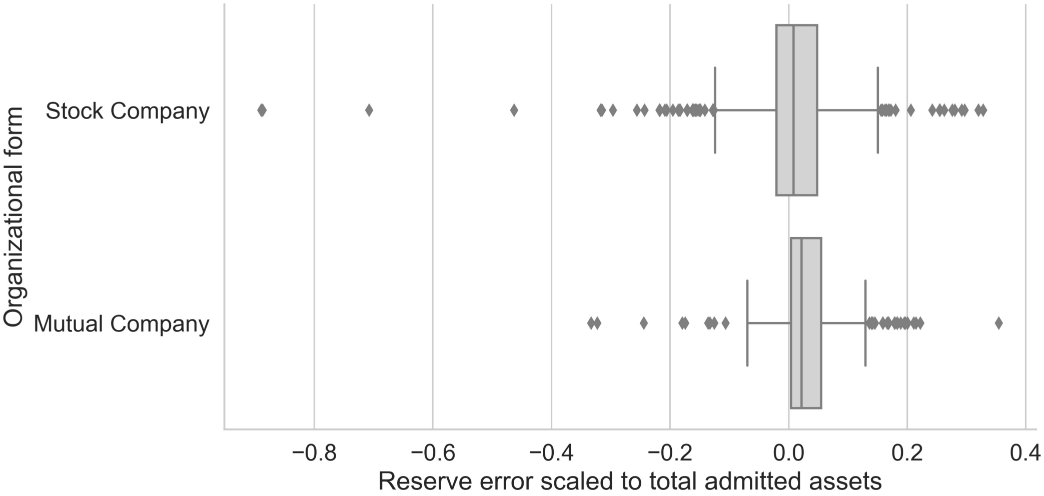
Original reserve$_{2012,\ 2012}$ =
Incurred losses$_{2012,\ 2012}$ –
Cumulative paid losses$_{2012,\ 2012}$

Developed reserve$_{2012,\ 2017}$ =
Incurred losses$_{2012,\ 2017}$ –
Cumulative paid losses$_{2012,\ 2012}$

Error$_{2012,\ 2017}$ =
Incurred losses$_{2012,\ 2012}$ –
Incurred losses$_{2012,\ 2017}$

A.2 Descriptive statistics for stock and mutual companies

Figure A.2: Comparison of the reserve error distribution for stock and mutual companies



The figure compares the reserve error scaled to total assets for stock and mutual companies. The boxplot indicates the 25%-, 50%-, and 75%- percentile. The light lines are whiskers indicating the data ranging within 1.5 times interquantile range past the low and high quartiles. Points outside this range are identified as outliers.

Table A.1: Groupwise summary statistics

		Count	**Mean**	**SD**	**25%**	**50%**	**75%**
Scaled reserve error	Total	722	0.0149	0.1006	-0.0061	0.0151	0.0514
	Mutual	293	0.0324	0.0698	0.0038	0.0214	0.0547
	Stock	429	0.0028	0.1156	-0.0208	0.0082	0.0483
Reported reserve	Total	722	734.2433	3,868.8758	2.4380	16.0270	116.4905
	Mutual	293	316.0753	1,973.6405	1.8010	17.0510	99.8590
	Stock	429	1,019.8452	4,728.2283	2.6810	14.9980	130.2620
Developed reserve	Total	722	702.3914	3,790.7891	2.1318	13.4585	109.8980
	Mutual	293	289.7807	1827.4979	1.3870	12.5200	87.9450
	Stock	429	984.1978	4661.8533	2.7370	13.7670	125.2510
Total assets	Total	722	2,048.1794	11,574.8734	17.8359	80.7031	445.2465
	Mutual	293	1,187.8026	8,796.4823	14.9741	72.6196	402.2610
	Stock	429	2,635.8026	13,116.8798	18.8117	83.0273	522.3763
Net income	Total	721	49.0255	403.9909	0.0229	1.1751	8.7625
	Mutual	293	22.2802	163.3603	0.0571	1.1013	8.4124
	Stock	428	67.3347	506.0863	-0.0476	1.2581	8.8335
Direct premiums written	Total	722	631.7138	3,193.1869	7.2752	36.7996	186.5501
	Mutual	293	425.3609	3,307.0738	5.9046	29.7026	155.2256
	Stock	429	772.6495	3,109.0426	8.3644	39.4052	231.9189
Growth of direct premiums written	Total	691	12.2079	46.2521	-0.0071	5.7713	13.0655
	Mutual	292	6.3113	16.8378	0.3787	5.1518	9.7688
	Stock	399	16.5232	58.7987	-0.3767	6.6728	17.0965
Business concentration index	Total	695	0.5480	0.3309	0.2476	0.4753	0.9813
	Mutual	291	0.5059	0.3273	0.2271	0.3794	0.9258
	Stock	404	0.5783	0.3305	0.2802	0.5198	0.9967
RBC ratio	Total	683	1,297.1244	2,914.9587	496.2897	826.6821	1,281.0761
	Mutual	271	1,227.7156	1,010.9459	668.6531	973.1861	1,507.3788
	Stock	412	1,342.7792	3,663.7355	451.2245	707.9099	1,166.5369
MD&A characters	Total	722	32.1458	82.7371	8.3875	13.1140	20.6820
	Mutual	293	17.5471	40.8568	8.1660	12.9210	18.7280
	Stock	429	42.1164	100.7336	8.4900	13.4460	22.3990

All numbers are taken from the companies' respective NAIC annual statements for 2012. *Reported reserve, developed reserve, total admitted assets, net income, and direct premiums written* are presented in million US-Dollars. *Growth of direct premiums written* and *RBC ratio* are expressed in percentage rates. *Reserve error* is scaled to total admitted assets. *Developed reserve* is calculated for a five-year development window, i.e., it is measured in 2017. *Premiums growth* measures the annual growth rate of direct premiums written from 2011 to 2012. *Business concentration index* is a Herfindahl Index across the property and casualty lines of business and is based on direct premiums earned. *RBC ratio* indicates the ratio of adjusted capital to the authorized control level. *MD&A characters* denote the count of characters (in 1,000) in the MD&A section of a company.

A.3 MD&A collection procedure

Document retrieval

We search for MD&A sections for all companies fulfilling the financial sample selection criteria described in section 3. For companies that have been publicly traded in 2012, we collect the 10-K filing for the reporting year 2012. We use a Python script to cut out the MD&A section on the basis of the beginning and ending of the section. We identified the beginning based on the header "Item 7." and the ending based on the header of the next section "Item 8.". We account for possible spelling variants, such as "Item 7", "Item7", "item.7". For all remaining private companies, stock as well as mutual, we collect the MD&A section as part of the annual statement filing with the NAIC. All filings are accessed via the S&P Global Market Intelligence Documents & Filings Search. If entities of a company group have individual filings, the filing of the largest US property and casualty entity in terms of reported capital and surplus is collected.

Transformation to raw text

All filings are stored as a PDF document and need to be changed to raw text for the textual analysis. We use the tool pdftotext.com for the conversion. In case of a technical error of the pdf text conversion, we assessed whether an alternative, pdf2go.com, is successful. However, in some cases the pdf consists of an image of the MD&A section and we need to use an adequate image conversion tool, www.convertimagetotext.net. In case of a technical error during the image conversion, we use www.newocr.com.

Merge with data set

Lastly, we merge the MD&A raw text sections with the financial dataset. Any further text cleaning, such as lower casing or deletion of numbers, is done by the TF-IDF vectorizer itself.

A.4 Machine learning models

This section introdues the machine learning models that we have used to classify the sign of the reserve error. We consider twelve different models that belong to different model families. During the model selection, we use all unigrams of lemmatized tokens in the MD&A sections as feature set. We fit the models in a cross-validation approach to maximize the out-of-sample power. A detailed presentation is beyond the scope of this paper and we refer to (Hastie et al., 2009; James et al., 2013).

K-Nearest Neigbors model belongs to instance-based learning. It classifies a new data point based on the closest data points in the training data, the data point's nearest neighbors. It is conducted as a simple majority vote of the k nearest neighbors of the data point, whereby k is an integer number chosen by the researcher. The data point is allocated to the group with the most representatives in its neighborhood.

We use Python scikit-learn package "KNeighborsClassifier" to fit the k-nearest neighbors model with the default number of neighbors that equals five neighbors.

Logistic Regression, also known as logit regression of log-linear classifier, is a linear classification model. Despite its name, it is a classification algorithm and not a regression algorithm. It uses a logistic function to model the probability that the data point's value of the target variable belongs to a certain target variable class given the datapoint's features. The logistic function ensures that the predicted probability ranges between 0 and 1. It thereby compensates the inability of linear regression models to appropriately investigate influences on discrete variables. For any data point, a prediction of the target variable' value can be made by defining a decision boundary. It uses the maximum likelihood to estimate the features' coefficients. If the probability to belong to a certain class is above the defined threshold, the target variable is predicted to belong to this particular class.

We use Python scikit-learn package "LogisticRegression" to fit the logistic regression model with the default L2 penalty. Additionally, we set a random initial starting value to be able to replicate the results.

Linear Discriminant Analysis is a classifier that assumes that the sample is drawn from a multivariate Gaussian distribution with a class-specific mean vector and a common covariance matrix. By projecting input data into a linear subspace whose directions maximize class separation, it can be applied to reduce dimensionality in a supervised manner.The Linear Discriminant Analysis is closely connected to the Logistic Regression model: Both approaches produce linear decision boundaries, but the models differ in the way the coefficients are computed.

We use Python scikit-learn package "LinearDiscriminantAnalysis". Since a Linear Discriminant Analysis requires dense data, but the TF-IDF-matrix is sparse, we add a FunctionTransformer to the pipeline to generate dense matrix.

Naive Bayes Classifier belongs to the supervised learning algorithms that apply the Bayes theorem. The underlying assumption is that feature pairs are conditionally independent from each other given the class variable value. While the assumption does not hold true in general, it considerably facilitates the estimation. The Multinomial Naive Bayes Classifier is one variant of the Naive Bayes method and it applies the Naive Bayes algorithm to data that is multinomially distributed. The Multinomial Naive Bayes Classifier is suitable for classification with discrete features (e.g., word counts for text classification) (Schütze et al., 2008).

We use Python scikit-learn package "MultinomialNB" to fit the Mulitnomial Naive Bayes Classification model with the default additive smoothing parameter of 1.0.

Decision Tree models are part of the non-parametric supervised learning methods. As such, they find use in both regressions and classifications. In essence, they learn a hierarchy of True/False questions, that lead to a decision. Decision Trees take data features and derive decision-making rules from them to devise a model that can predict a given target variable's value. In more detail, they split the feature space into rectangles and subsequently fit a model for each one. Decision Trees can be used with both categorical and numerical data.

We use Python scikit-learn package "DecisionTreeClassifier" to fit the decision tree model. To avoid overfitting, we restrict the maximum depth of the tree to ten (Müller and Guido, 2016). Additionally, we set a random initial starting value to be able to replicate the results.

Random Forest belongs to the group of ensemble methods, and within that to the subgroup of averaging methods. As such, it constructs a vast array of uncorrelated decision trees that are built on bootstrapped training samples. The resulting element of randomness allows for trees with prediction errors, that can cancel each other out upon averaging them. This reduces the amount of overfitting. Consequently, compared to a single Decision Tree, Random Forests can often yield a significantly lower variance at the expense of only a little increase in bias, and thus a better performing model.

We use Python scikit-learn package "RandomForestClassifier" to grow a random forest of 100 decision tree model. We restrict the maximum depth of any decision tree to ten (Müller and Guido, 2016). Additionally, we set a random initial starting value to be able to replicate the results.

Gradient Boosting Machine is another ensemble method that combines several decision trees. In contrast to the Random Forest model, gradient boosting builds trees sequentially, where each tree tries to correct the errors of the previous one. Thereby, several weaker models can be synthesized into a more powerful one. The method particularly stands out due to its robustness regarding outliers.

We use Python scikit-learn package "GradientBoostingClassifier" with 100 boosting stages to perform. We restrict the maximum depth of any decision tree to three nodes (Müller and Guido, 2016). The contribution of each new tree shrinks with the learning rate, which is set to 0.1. Additionally, we set a random initial starting value to be able to replicate the results.

Support Vector Machine is a representation of the examples as points in space, mapped in way such that the examples of the separate categories are divided by a clear gap that is as wide as possible. The model classifies samples to find the decision boundary with the largest margin. New examples are then mapped into that same space and predicted to belong to a category based on the side of the gap on which they fall. Support Vector Machines are soft-margin classifiers: In the interest of better classification of most of the training set and greater robustness to new examples, the maximum margin does not need to perfectly separate the two classes. There are several variants of Support Vector Machines. The linear Support Vector Classifier works with a linear class boundary, whereas Support Vector Machines are extensions of the Support Vector Classifier and accommodate non-linear class boundaries. Support Vector Machines result results from enlarging the feature space using kernels (James et al., 2013).

We assess a linear Support Vector Classifier by using the scikit-learn package "LinearSVC". Moreover, we use the Python scikit-learn package "SVC" to make classifications with a Support Vector Machine. We investigate to two different kernel function: Radial basis function kernel ("rbf") and a linear kernel. Gamma, the kernel coefficient for "rbf", is set to 'scale', which uses the value 1/(number of features * X.var()). In all three models, we set a random initial starting value to be able to replicate the results.

Stochastic Gradient Descent is a method associated with discriminant learning of linear classifiers under convex loss functions. The method recently rose to prominence along with some of its major fields of application that include natural language processing or text classification. Being comparatively efficient and simple, it particularly proved useful in tasks characterized by large scale and scarce data. A Gradient Descent measures the local gradient of the cost function with regards to the parameter vector and it moves in the direction of the descending gradient. Once the gradient is zero, it reaches a minimum. The gradient vector contains all the partial derivatives of the cost function for the complete training set. A Stochastic Gradient Descent is an efficient learning approach for linear classifiers with a convex cost function. The gradients are computed based on that single instance rather than on full training set.

We use Python scikit-learn package "SGDClassifier" with a hinge loss function. The hinge loss function describes a soft-margin linear Support Vector Machine, which is equivalent to a SVC with a linear kernel. We set a random initial starting value to be able to replicate the results. We use balanced class weights to adjust weights inversely proportional to class frequencies. A class denotes a specific value of the target variable).

Neural Network comprise a selection of numerous non-linear statistical models. In general, a neural network can be understood as a regression or classification model with two steps. In step one, the idea is to arrive at features that are derived from linear input combinations. In step two, the features serve to model the target as a function of those features. Neural Networks have proven powerful in many ways and thus have found broad application in diverse fields.

We use Python scikit-learn package "MLPClassifier" for the Multi-Layer Perceptron Network with 100 hidden layers. We set a random initial starting value to be able to replicate the results.

A.5 Parameter fine-tuning of the most promising models of the model selection

A.5.1 Parameter fine-tuning of the Gradient Boosting model

Table A.2: Tuning of number of trees

	Number of trees	**Mean test score**	**SD test score**	**Rank test score**
1	20	0.6268	0.0364	8
2	30	0.6213	0.0392	9
3	40	0.6314	0.0433	6
4	50	0.6301	0.0432	7
5	60	0.6369	0.0473	3
6	70	0.6355	0.0483	5
7	80	0.6427	0.0445	1
8	90	0.6390	0.0391	2
9	100	0.6364	0.0397	4

The table shows the grid search results for the first-level parameter fine-tuning of the Gradient Boosting model. The first-level parameter fine-tuning identifies the number of trees with the highest test score among the range from 20 to 100. The table shows the mean, the standard deviation (SD), and the rank of the AUC test scores of the five-fold cross-validation.

Table A.3: First tree-specific tuning level: Maximal depth and minimum of samples for a split

	Maximal depth	**Minimum samples for split**	**Mean test score**	**SD test score**	**Rank test score**
1	3	2	0.6427	0.0445	3
2	3	5	0.6354	0.0259	6
3	3	7	0.6432	0.0318	2
4	3	10	0.6338	0.0190	8
5	3	15	0.6314	0.0067	10
6	3	20	0.6324	0.0223	9
7	3	30	0.6362	0.0308	5
8	5	2	0.6148	0.0283	24
9	5	5	0.6018	0.0510	35
10	5	7	0.6021	0.0321	34
11	5	10	0.6077	0.0379	31
12	5	15	0.6196	0.0293	19
13	5	20	0.6185	0.0213	20
14	5	30	0.6376	0.0356	4
15	7	2	0.6115	0.0540	28

Table A.3: First tree-specific tuning level: Maximal depth and minimum of samples for a split (continued)

	Maximal depth	**Minimum samples for split**	**Mean test score**	**SD test score**	**Rank test score**
16	7	5	0.6168	0.0373	21
17	7	7	0.6160	0.0662	22
18	7	10	0.6207	0.0450	18
19	7	15	0.6271	0.0533	15
20	7	20	0.6350	0.0444	7
21	7	30	0.6140	0.0229	26
22	10	2	0.5994	0.0446	36
23	10	5	0.5951	0.0517	38
24	10	7	0.6071	0.0694	32
25	10	10	0.6148	0.0511	25
26	10	15	0.6291	0.0410	12
27	10	20	0.6493	0.0555	1
28	10	30	0.6302	0.0474	11
29	12	2	0.6156	0.0579	23
30	12	5	0.5804	0.0598	41
31	12	7	0.6132	0.0589	27
32	12	10	0.6069	0.0469	33
33	12	15	0.6253	0.0531	16
34	12	20	0.6247	0.0497	17
35	12	30	0.6274	0.0382	14
36	15	2	0.5993	0.0449	37
37	15	5	0.5815	0.0597	40
38	15	7	0.5789	0.0480	42
39	15	10	0.5859	0.0535	39
40	15	15	0.6085	0.0522	29
41	15	20	0.6082	0.0508	30
42	15	30	0.6289	0.0277	13

The table shows the grid search results for the first tree-specific tuning level, which consider a range of maximal depth from 3 to 15 and a range of minimum of samples required for a split from 2 to 30. The Gradient Boosting model's number of trees is 80, which has been identified in the first-level parameter fine-tuning of the Gradient Boosting model. The table shows the mean, the standard deviation (SD), and the rank of the AUC test scores of the five-fold cross-validation.

Table A.4: Second tree-specific tuning level: Minimum of samples at a leaf

	Minimum samples at leaf	**Mean test score**	**SD test score**	**Rank test score**
1	1	0.6493	0.0555	2
2	2	0.6228	0.0431	5
3	5	0.6142	0.0212	6
4	7	0.6440	0.0268	4
5	10	0.6602	0.0183	1
6	15	0.6463	0.0263	3

The table shows the grid search results for the second tree-specific tuning level, which considers a range of minimum of samples at a leaf from 1 to 15. The Gradient Boosting model has the following parameters, which have been identified in the first-level parameter fine-tuning and the first tree-specific tuning: the number of trees is 80, the maximal depth of a tree is 10, and the minimum of samples for a split is 20. The table shows the mean, the standard deviation (SD), and the rank of the AUC test scores of the five-fold cross-validation.

Table A.5: Third tree-specific tuning level: Maximal number of features considered for split at a node

	Maximal features	**Mean test score**	**SD test score**	**Rank test score**
1	10	0.5696	0.0224	6
2	20	0.5848	0.0384	5
3	50	0.6264	0.0649	2
4	100	0.6019	0.0240	3
5	Squareroot(N features)	0.5950	0.0598	4
6	N features	0.6602	0.0183	1

The table shows the grid search results for the third tree-specific tuning level, which considers a range of maximal features considered at each node from 10 to number of all features. The number of features is 18600 and the square root of the number of features is 138. The Gradient Boosting model has the following parameters, which have been identified in the first-level parameter fine-tuning and the first and second tree-specific tuning: the number of trees is 80, the maximal depth of a tree is 10, the minimum of samples for a split is 20, and the minimum samples at a leaf are 10. The table shows the mean, the standard deviation (SD), and the rank of the AUC test scores of the five-fold cross-validation.

Table A.6: Text preprocessing parameter tuning for the Gradient Boosting model

	Maximal share of documents	**Minimal number of documents**	**Mean test score**	**SD test score**	**Rank test score**
1	1	1	0.6602	0.0183	1
2	1	5	0.6349	0.0405	3
3	1	10	0.6400	0.0270	2
4	0.9	1	0.6130	0.0381	7
5	0.9	5	0.6168	0.0424	5
6	0.9	10	0.6096	0.0287	9
7	0.8	1	0.6189	0.0361	4
8	0.8	5	0.6159	0.0464	6
9	0.8	10	0.6110	0.0528	8

The table shows the grid search results for a range of the maximum share of documents for a token from 0.8 and 1.0 in combination with a range of the minimum number of documents for a token from 1 to 10. The Gradient Boosting model has the following parameters, which have been identified in the first-level parameter fine-tuning and the tree-specific tuning: the number of trees is 80, the maximal depth of a tree is 10, the minimum of samples for a split is 20, the minimum samples at a leaf are 10, and all features are considered for a split. The table shows the mean, the standard deviation (SD), and the rank of the AUC test scores of the five-fold cross-validation.

A.5.2 Parameter fine-tuning of the Stochastic Gradient Descent model

Table A.7: Parameter fine-tuning for the Stochastic Gradient Descent model

	Alpha	L1 ratio	Mean test score	SD test score	Rank test score
1	0.0001	0	0.6148	0.0336	10
2	0.0001	0.1	0.6139	0.0347	14
3	0.0001	0.15	0.6141	0.0335	13
4	0.0001	0.2	0.6133	0.0330	16
5	0.0001	0.3	0.6137	0.0321	15
6	0.0001	0.4	0.6145	0.0364	12
7	0.0001	0.5	0.6119	0.0369	18
8	0.0001	0.6	0.6103	0.0404	20
9	0.0001	0.7	0.6111	0.0413	19
10	0.0001	0.8	0.6099	0.0432	21
11	0.0001	0.9	0.6076	0.0442	24
12	0.0001	1	0.6030	0.0465	26
13	0.001	0	0.6219	0.0389	1
14	0.001	0.1	0.6186	0.0407	3
15	0.001	0.15	0.6189	0.0431	2
16	0.001	0.2	0.6156	0.0452	7
17	0.001	0.3	0.6146	0.0481	11
18	0.001	0.4	0.6157	0.0507	6
19	0.001	0.5	0.6168	0.0538	5
20	0.001	0.6	0.6154	0.0561	8
21	0.001	0.7	0.6149	0.0565	9
22	0.001	0.8	0.6128	0.0537	17
23	0.001	0.9	0.6091	0.0532	22
24	0.001	1	0.6083	0.0502	23
25	0.01	0	0.5899	0.0311	28
26	0.01	0.1	0.5734	0.0310	31
27	0.01	0.15	0.5638	0.0343	32
28	0.01	0.2	0.5620	0.0437	33
29	0.01	0.3	0.5540	0.0557	34
30	0.01	0.4	0.5821	0.0695	30
31	0.01	0.5	0.5934	0.0645	27
32	0.01	0.6	0.6058	0.0291	25
33	0.01	0.7	0.6176	0.0297	4
34	0.01	0.8	0.5383	0.0485	35
35	0.01	0.9	0.5231	0.0464	36
36	0.01	1	0.5000	0.0000	37
37	0.1	0	0.5884	0.0319	29

Table A.7: Parameter fine-tuning for the Stochastic Gradient Descent model (continued)

	Alpha	**L1 ratio**	**Mean test score**	**SD test score**	**Rank test score**
38	0.1	0.1	0.5000	0.0000	37
39	0.1	0.15	0.5000	0.0000	37
40	0.1	0.2	0.5000	0.0000	37
41	0.1	0.3	0.5000	0.0000	37
42	0.1	0.4	0.5000	0.0000	37
43	0.1	0.5	0.5000	0.0000	37
44	0.1	0.6	0.5000	0.0000	37
45	0.1	0.7	0.5000	0.0000	37
46	0.1	0.8	0.5000	0.0000	37
47	0.1	0.9	0.5000	0.0000	37
48	0.1	1	0.5000	0.0000	37

The table shows the grid search results for a range of the learning rate alpha from 0.0001 to 0.1 in combination with a range of the L1 ratio from 0 to 1. The L1 ratio describes the form of the penalty. It can be the squared euclidean norm (L2, which corresponds to the value of the L1 ratio 0) or the absolute norm (L1, which corresponds to the value of the L1 ratio1) or combinations of both (L1 ratio between 0 and 1). The Stochastic Gradient Descent model has a hinge loss function and balanced class weights. The table shows the mean, the standard deviation (SD), and the rank of the AUC test scores of the five-fold cross-validation.

Table A.8: Text preprocessing parameter tuning for the Stochastic Gradient Descent model

	Maximal share of documents	**Minimal number of documents**	**Mean test score**	**SD test score**	**Rank test score**
1	1	1	0.6219	0.0389	7
2	1	5	0.6179	0.0337	8
3	1	10	0.6156	0.0354	9
4	0.9	1	0.6498	0.0404	4
5	0.9	5	0.6470	0.0282	6
6	0.9	10	0.6496	0.0321	5
7	0.8	1	0.6531	0.0505	2
8	0.8	5	0.6499	0.0434	3
9	0.8	10	0.6556	0.0399	1

The table shows the grid search results for a range of the maximum share of documents for a token from 0.8 and 1.0 in combination with a range of the minimum number of documents for a token from 1 to 10. The Stochastic Gradient Descent model has a L2 penalty, a learning rate of 0.001, a hinge loss function and balanced class weights. The table shows the mean, the standard deviation (SD), and the rank of the AUC test scores of the five-fold cross-validation.

A.6 Evaluation of alternative text representation

A.6.1 Tuning of text representation refinements of the Stochastic Gradient Descent model

Table A.9: Text representation refinement

	Ngram range	**Stop words**	**Mean test score**	**SD test score**	**Rank test score**
1	(1, 1)		0.6618	0.0411	10
2	(1, 1)	Abbreviations	0.6538	0.0418	28
3	(1, 1)	Numbers	0.6682	0.0422	3
4	(1, 1)	Locations	0.6465	0.0452	41
5	(1, 1)	Names	0.6609	0.0396	15
6	(1, 1)	Abbreviations, Numbers	0.6584	0.0414	20
7	(1, 1)	Abbreviations, Locations	0.6401	0.0426	46
8	(1, 1)	Abbreviations, Names	0.6547	0.0430	27
9	(1, 1)	Numbers, Locations	0.6560	0.0400	25
10	(1, 1)	Numbers, Names	0.6675	0.0374	4
11	(1, 1)	Locations, Names	0.6496	0.0413	34
12	(1, 1)	Abbreviations, Numbers, Locations	0.6456	0.0389	42
13	(1, 1)	Abbreviations, Numbers, Names	0.6615	0.0408	12

Table A.9: Text representation refinement (continued)

	Ngram range	Stop words	Mean test score	SD test score	Rank test score
14	(1, 1)	Abbreviations, Locations, Names	0.6415	0.0422	44
15	(1, 1)	Numbers, Locations, Names	0.6588	0.0399	19
16	(1, 1)	All	0.6486	0.0361	36
17	(1, 2)		0.6675	0.0228	5
18	(1, 2)	Abbreviations	0.6600	0.0293	17
19	(1, 2)	Numbers	0.6726	0.0222	2
20	(1, 2)	Locations	0.6548	0.0302	26
21	(1, 2)	Names	0.6671	0.0223	6
22	(1, 2)	Abbreviations, Numbers	0.6617	0.0309	11
23	(1, 2)	Abbreviations, Locations	0.6470	0.0383	40
24	(1, 2)	Abbreviations, Names	0.6595	0.0277	18
25	(1, 2)	Numbers, Locations	0.6574	0.0344	22
26	(1, 2)	Numbers, Names	0.6732	0.0221	1
27	(1, 2)	Locations, Names	0.6562	0.0300	24

Table A.9: Text representation refinement (continued)

	Ngram range	Stop words	Mean test score	SD test score	Rank test score
28	(1, 2)	Abbreviations, Numbers, Locations	0.6481	0.0419	39
29	(1, 2)	Abbreviations, Numbers, Names	0.6645	0.0279	9
30	(1, 2)	Abbreviations, Locations, Names	0.6481	0.0374	38
31	(1, 2)	Numbers, Locations, Names	0.6610	0.0324	14
32	(1, 2)	All	0.6496	0.0426	35
33	(1, 3)		0.6607	0.0298	16
34	(1, 3)	Abbreviations	0.6503	0.0333	31
35	(1, 3)	Numbers	0.6651	0.0273	7
36	(1, 3)	Locations	0.6499	0.0310	32
37	(1, 3)	Names	0.6614	0.0294	13
38	(1, 3)	Abbreviations, Numbers	0.6569	0.0312	23
39	(1, 3)	Abbreviations, Locations	0.6396	0.0363	47
40	(1, 3)	Abbreviations, Names	0.6512	0.0320	29
41	(1, 3)	Numbers, Locations	0.6508	0.0338	30

Table A.9: Text representation refinement (continued)

	Ngram range	**Stop words**	**Mean test score**	**SD test score**	**Rank test score**
42	(1, 3)	Numbers, Names	0.6649	0.0252	8
43	(1, 3)	Locations, Names	0.6483	0.0313	37
44	(1, 3)	Abbreviations, Numbers, Locations	0.6430	0.0366	43
45	(1, 3)	Abbreviations, Numbers, Names	0.6578	0.0317	21
46	(1, 3)	Abbreviations, Locations, Names	0.6378	0.0382	48
47	(1, 3)	Numbers, Locations, Names	0.6497	0.0329	33
48	(1, 3)	All	0.6404	0.0389	45

The table shows the grid search results for ngram range and stopword compositions of the Stochastic Gradient Descent model. The ngram range describes whether unigrams, uni- and bigrams, or uni-,bi-, and trigrams should be used by the classification model. Four stop word lists are defined: Abbreviations, verbally written numbers, locations, and names. The stop word lists are shown in Section A.6.3. The classification model is a Stochastic Gradient Descent model with a L2 penalty, a learning rate of 0.001, a hinge loss function and balanced class weights. The text preprocessing ignores lemmatized tokens of the MD&A section that appear either in more than 80 percent of the documents or in less than 10 documents. The table shows the mean, the standard deviation (SD), and the rank of the AUC test scores of the five-fold cross-validation.

A.6.2 Tuning of text representation refinements of the Gradient Boosting model

Table A.10: Text representation refinement

	Ngram range	**Stop words**	**Mean test score**	**SD test score**	**Rank test score**
1	(1, 1)		0.6602	0.0183	1
2	(1, 1)	Abbreviations	0.6376	0.0398	3
3	(1, 1)	Numbers	0.6280	0.0144	5
4	(1, 1)	Locations	0.6305	0.0307	4
5	(1, 1)	Names	0.6136	0.0325	11
6	(1, 1)	Abbreviations, Numbers	0.6020	0.0303	18
7	(1, 1)	Abbreviations, Locations	0.5936	0.0299	35
8	(1, 1)	Abbreviations, Names	0.6234	0.0360	6
9	(1, 1)	Numbers, Locations	0.6061	0.0363	14
10	(1, 1)	Numbers, Names	0.6383	0.0215	2
11	(1, 1)	Locations, Names	0.6223	0.0292	7
12	(1, 1)	Abbreviations, Numbers, Locations	0.5946	0.0359	34
13	(1, 1)	Abbreviations, Numbers, Names	0.6218	0.0295	8

Table A.10: Text representation refinement (continued)

	Ngram range	**Stop words**	**Mean test score**	**SD test score**	**Rank test score**
14	(1, 1)	Abbreviations, Locations, Names	0.5930	0.0320	36
15	(1, 1)	Numbers, Locations, Names	0.6144	0.0210	10
16	(1, 1)	All	0.6049	0.0261	16
17	(1, 2)		0.6188	0.0444	9
18	(1, 2)	Abbreviations	0.6000	0.0395	22
19	(1, 2)	Numbers	0.5955	0.0469	31
20	(1, 2)	Locations	0.6122	0.0554	12
21	(1, 2)	Names	0.6061	0.0307	15
22	(1, 2)	Abbreviations, Numbers	0.6003	0.0490	21
23	(1, 2)	Abbreviations, Locations	0.5985	0.0645	26
24	(1, 2)	Abbreviations, Names	0.5970	0.0437	28
25	(1, 2)	Numbers, Locations	0.6082	0.0548	13
26	(1, 2)	Numbers, Names	0.6009	0.0417	20
27	(1, 2)	Locations, Names	0.5974	0.0433	27

Table A.10: Text representation refinement (continued)

	Ngram range	**Stop words**	**Mean test score**	**SD test score**	**Rank test score**
28	(1, 2)	Abbreviations, Numbers, Locations	0.5997	0.0663	23
29	(1, 2)	Abbreviations, Numbers, Names	0.5957	0.0315	30
30	(1, 2)	Abbreviations, Locations, Names	0.6024	0.0603	17
31	(1, 2)	Numbers, Locations, Names	0.5915	0.0460	37
32	(1, 2)	All	0.5739	0.0477	46
33	(1, 3)		0.5685	0.0269	47
34	(1, 3)	Abbreviations	0.5863	0.0287	40
35	(1, 3)	Numbers	0.5948	0.0522	33
36	(1, 3)	Locations	0.5968	0.0395	29
37	(1, 3)	Names	0.5764	0.0340	45
38	(1, 3)	Abbreviations, Numbers	0.5624	0.0268	48
39	(1, 3)	Abbreviations, Locations	0.5846	0.0118	42
40	(1, 3)	Abbreviations, Names	0.5765	0.0497	44
41	(1, 3)	Numbers, Locations	0.5950	0.0411	32

Table A.10: Text representation refinement (continued)

	Ngram range	Stop words	Mean test score	SD test score	Rank test score
42	(1, 3)	Numbers, Names	0.5848	0.0291	41
43	(1, 3)	Locations, Names	0.5885	0.0343	38
44	(1, 3)	Abbreviations, Numbers, Locations	0.5881	0.0439	39
45	(1, 3)	Abbreviations, Numbers, Names	0.5831	0.0441	43
46	(1, 3)	Abbreviations, Locations, Names	0.5995	0.0261	25
47	(1, 3)	Numbers, Locations, Names	0.5997	0.0444	24
48	(1, 3)	All	0.6015	0.0498	19

The table shows the grid search results for ngram range and stopword compositions of the Gradient Boosting model. The ngram range describes whether unigrams, uni- and bigrams, or uni-,bi-, and trigrams should be used by the classification model. Four stop word lists are defined: Abbreviations, verbally written numbers, locations, and names. The stop word lists are shown in Section A.6.3. The classification model is a Gradient Boosting model with the following parameters: the number of trees is 80, the maximal depth of a tree is 10, the minimum of samples for a split is 20, the minimum samples at a leaf are 10, and all features are considered for a split. The table shows the mean, the standard deviation (SD), and the rank of the AUC test scores of the five-fold cross-validation.

A.6.3 Stop words lists

Table A.11: Stop word list: Abbreviations

aa
aaa
ab
abs
ac
ace
ach
acl
acre
ad
adj
afs
ag
ak
al
alae
amt
ann
aoci
ar
asc
asi
asic
asu
ay
az
ba
baa
bb
bbb
bi
bma
bona
bop
bp
bps
ca
caa
cc
co
col
coli
comp
cpa
cpcu
cpp
csi
ct
da
dac
db
dc
dcc
ddr
de
dec
des
dfs
dl
doi
dpac
edp
eea
eft
en
ent
epli
er
es
et
etc
ex
exp
fa
fas
fasb
fcas
fcic
gic
gl
gmdb
gmib
gnma
gse
guam
hcc
hi
hm
ho
ia
ibnr
id
ii
iii
iine
ike
il
inc
inure
inv
inve
iong
ioss
irs
isda
iv
ix
ks
ky
le
lee
li
liab
liabili
llc
llp
loc
mga
mgas
mi
mic
mm
mn
mo
mpci
mrb
ms
mt
na
nalc
nc
ncci
nd
ne
nh
nj
nm
nol
nols
nr
nt
nv
nys
occ
oci
oh
ooo
otc
otti
pa
pcs
pd
phs
pip
pl
plc
ri
rli
rm
rmbs
rrg
sa
sba
sc
sch
sd
se
sr
ss
ssap
st
ste
sti
su
svo
tac
te
tenn
tn
toa
too
top
toss
tpa
tri
tria
tx
ulae
un
va
vi
vie
vii
viii
vp

Table A.11: Stop word list: Abbreviations (continued)

ccc
cd
cdo
cdos
cds
ce
cfc
cfo
cfpb
chg
cic
cio
clo
cmbs
cmp
fdic
fhcf
fhlb
fhlmc
fico
fide
fio
fl
forma
frb
fsa
fsb
fsoc
ft
ga
los
lp
lpt
ltd
ltv
lus
ma
maaa
mac
maj
map
mbia
mbs
md
mda
pml
po
ppa
pr
pra
pre
prem
proj
pt
qbe
rata
rd
rec
reins
relie
vs
vt
wa
wc
wi
wm
wv
xi
xol
xxx
yr

Table A.12: Stop word list: Company names and names

aic
aig
alexander
allianz
allstate
ally
andrew
anthony
antonio
aon
arthur
barclays
benjamin
berkley
berkshire
bernard
bradley
brian
bruce
carlson
catlin
christopher
chubb
citigroup
clarendon
columbus
craig
dale
daniel
david
davis
dennis
deutsche
donald
douglas
edward
edwin
eric
erisa
everest
fannie
ferguson
francis
frank
franklin
freddie
frederick
gary
george
goldman
gregory
hampshire
hancock
harold
harris
hathaway
henry
howard
irene
iris
isaac
james
jean
jefferson
jeffrey
jim
john
johnson
jones
joseph
jp
jpmorgan
katrina
keith
kelly
kenneth
kevin
kpmg
larry
lehman
levy
lewis
lisa
lloyd
lloyds
logan
louis
lynn
mae
maiden
marie
mark
marsh
martin
mary
matthew
mesothelioma
michael
moody
moore
morgan
morris
nationwide
obama
odyssey
oliver
onebeacon
patrick
paul
penn
peter
pimco
ransom
republic
richard
rita
robert
roger
ronald
russell
sachs
samuel
san
sandy
sap
scor
scott
scottsdale
smith
statewide
stephen
sterling
susan
terry
thomas
thompson
timothy
vincent
watson
wayne
wells
william
willis
wilson
wyman

Table A.13: Stop word list: Locations

africa
alabama
alaska
america
american
americas
angeles
arizona
arkansas
asia
atlanta
atlantic
austin
australia
australian
baltimore
belgium
bermuda
bermudian
bermudians
boston
brazil
british
california
canada
canadian
caribbean
carolina
cayman
chicago
chile
china
cincinnati
colorado
columbia
connecticut
dakota
dallas
doddfrank
ecuador
england
eu
euro
europe
european
eurozone
florida
francisco
georgia
germany
greece
hannover
harrisburg
hartford
hawaii
houston
idaho
illinois
indiana
iowa
ireland
irish
island
islands
italy
japan
japanese
jersey
kansas
kentucky
kingdom
kong
la
latin
louisiana
luxembourg
madison
madrid
maine
markel
maryland
mass
massachusetts
mexico
miami
michigan
midwest
midwestern
milwaukee
minnesota
mississippi
missouri
moines
montana
montpelier
munich
nebraska
netherlands
nevada
north
northeast
northeastern
northern
northwest
norway
ny
ohio
oklahoma
ontario
oregon
pennsylvania
philadelphia
pittsburgh
portugal
puerto
rhode
richmond
rico
silica
singapore
southeast
southeastern
southern
southwest
spain
sweden
swiss
switzerland
tennessee
texas
thailand
tokio
transatlantic
uk
usa
utah
vermont
virginia
washington
wellington
west
western
wilmington
wisconsin
wyoming
york
zealand
zurich

Table A.14: Stop word list: Verbally expressed numbers

eight
eighteen
eighth
eighty
eleven
eleventh
fifteen
fifteenth
fifth
fifty
first
five
fiveyear
four
fourteen
fourth
hundred
million
nine
ninety
ninth
seven
seventeen
seventh
seventy
six
sixteen
sixth
sixty
ten
tenth
thousand
three
twelfth
twelve
twenty
twice
two
frst
fve

A.7 Evaluation of alternative text representation with unsupervised learning model

Table A.15: Topic model for text representation: Selection of model with highest log-likelihood

	Number of topics	**Mean test score**	**SD test score**	**Rank test score**
1	1	-1530471.2433	598967.3572	6
2	2	-1522172.4140	593308.8462	2
3	3	-1522006.3645	592921.8627	1
4	4	-1524864.9429	591504.7287	3
5	5	-1525066.7913	584522.1268	4
6	6	-1529305.7810	584581.6824	5
7	7	-1530918.5237	582089.6876	7
8	8	-1533400.0806	580285.3188	8
9	9	-1537138.7123	579939.5564	9
10	10	-1541156.5680	579141.6590	10
11	25	-1586107.7619	574894.7218	11
12	50	-1639754.8785	574702.5288	12
13	75	-1667214.5548	581008.9054	13
14	100	-1699796.8030	578719.6818	14
15	125	-1716609.9218	577471.3137	15
16	150	-1744294.8105	575696.9019	16
17	175	-1762914.5238	574048.3657	17
18	200	-1784962.2406	579433.2413	18
19	225	-1804446.8797	569826.6527	19
20	250	-1813993.3589	575374.9190	20

The table shows the grid search results for the number of topics of an LDA topic model. The text preprocessing ignores lemmatized tokens of the MD&A section that appear either in more than 80 percent of the documents or in less than 10 documents. Words that are names or verbally expressed numbers are also ignored. The table shows the mean, the standard deviation (SD), and the rank of the log-likelihood test scores of the five-fold cross-validation.

Table A.16: Topic model for text representation: Performance of classification model combined with best identifying LDA model

Accuracy	**Precision**	**F1**	**AUC**
0.4950 (0.1547)	0.4141 (0.3381)	0.4407 (0.3633)	0.4882 (0.0269)

The table summarizes the performance scores of the cross-validation prediction of a Stochastic Gradient Descent classification model based on the topics of the best identifying LDA model. The Stochastic Gradient Descent model is implemented with a L2 penalty, a learning rate of 0.001, a hinge loss function and balanced class weights. The text preprocessing ignores lemmatized tokens of the MD&A section that appear either in more than 80 percent of the documents or in less than 10 documents. Words that are names or verbally expressed numbers are also ignored. An LDA topic model with 3 topics is used as second step of the text preprocessing. The table reports mean and standard deviation over the five cross-validation folds. Accuracy is defined as the share of correct predictions. Precision measures the share of true positives (that is, the case of overreserving in our model) among all predicted positives. The F1 score is calculated as the harmonic mean of precision and recall, which measures the share of predicted positives among all actual positives. The AUC estimates the probability that a random positive is ranked before a random negative (in our model, the case of underreserving), without specifying a particular decision threshold.

Table A.17: Topic model for text representation as part of overall classification model

	Number of topics	**Mean test score**	**SD test score**	**Rank test score**
1	5	0.5051	0.0560	16
2	6	0.5114	0.0484	15
3	7	0.5389	0.0464	11
4	8	0.5277	0.0474	13
5	9	0.5248	0.0307	14
6	10	0.5385	0.0509	12
7	25	0.5630	0.0694	9
8	50	0.5822	0.0518	7
9	75	0.5874	0.0720	6
10	100	0.5898	0.0511	5
11	125	0.6006	0.0655	3
12	150	0.6134	0.0223	1
13	175	0.5942	0.0455	4
14	200	0.5809	0.0368	8
15	225	0.5547	0.0663	10
16	250	0.6076	0.0282	2

The table shows the grid search results for the number of topics of a classification model that combines an LDA topic model with a Stochastic Gradient Descent classification model. The classification model is a Stochastic Gradient Descent model with a L2 penalty, a learning rate of 0.001, a hinge loss function and balanced class weights. The text preprocessing ignores lemmatized tokens of the MD&A section that appear either in more than 80 percent of the documents or in less than 10 documents. Words that are names or verbally expressed numbers are also ignored. An LDA topic model is used as second step of the text preprocessing. The table shows the mean, the standard deviation (SD), and the rank of the AUC test scores of the five-fold cross-validation.

A.8 Integration of financial information in Gradient Boosting model

Table A.18: Selection of financial information model

		Profit	**Growth (Positive)**	**Growth (Median)**	**Concentration**	**RBC ratio (Action)**	**RBC ratio (Median)**	**Average AUC score**	**Rank AUC score**
1.	**1.**	x						0.5964	25
	2.		x					0.6019	16
	3.			x				0.5881	32
	4.				x			0.5933	27
	5.					x		0.6081	10
	6.						x	0.5794	36
2.	**1.**	x	x					0.6100	9
	2.	x		x				0.6113	7
	3.	x			x			0.5985	22
	4.	x				x		0.6072	12
	5.	x					x	0.5931	30
	6.		x		x			0.5976	23
	7.		x			x		0.6015	18
	8.		x				x	0.6125	6
	9.			x	x			0.6018	17
	10.			x		x		0.6013	19
	11.			x			x	0.6003	20
	12.				x	x		0.5933	28
	13.				x		x	0.5873	33
3.	**1.**	x	x		x			0.5892	31
	2.	x	x			x		0.6052	14
	3.	x	x				x	0.5991	21
	4.	x		x	x			0.5863	34
	5.	x		x		x		0.6176	3
	6.	x		x			x	0.6040	15
	7.	x			x	x		0.6167	4
	8.	x			x		x	0.6068	13
	9.		x		x	x		0.6159	5
	10.		x		x		x	0.6080	11
	11.			x	x	x		0.6102	8
	12.			x	x		x	0.5971	24

Table A.18: Selection of financial information model (continued)

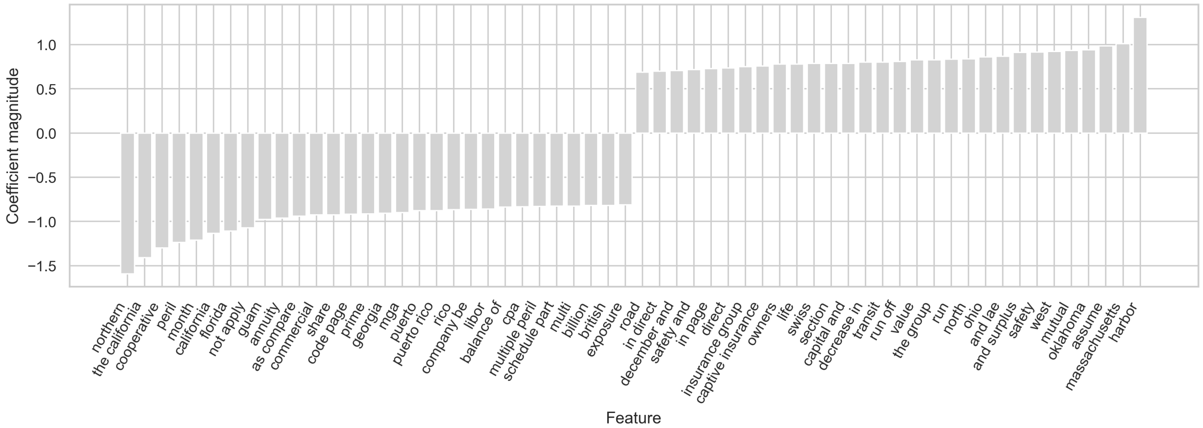
		Profit	**Growth (Positive)**	**Growth (Median)**	**Concentration**	**RBC ratio (Action)**	**RBC ratio (Median)**	**Average test score**	**Rank test score**
4.	**1.**	x	x		x	x		0.5858	35
	2.	x	x		x		x	0.6251	1
	3.	x		x	x	x		0.5932	29
	4.	x		x	x		x	0.6197	2
Baseline								0.5937	26

The table summarizes the AUC scores of the ten-fold cross-validation prediction of the alternative classification models with TF-IDF unigram word tokens of the MD&A sections. The column *Average AUC score* shows the mean AUC test score of the cross-validation and the column *Rank AUC score* indicates the rank of a specific average AUC score. The AUC estimates the probability that a random positive is ranked before a random negative (in our model, the case of underreserving), without specifying a particular decision threshold. The columns *Profit* to *RBC ratio (Median)* indicate which financial information is included. Compared to the Stochastic Gradient Descent model, interaction terms are not explicitly constructed, since the model is based on decision trees, which already consider the combination of different features at each decision node. The *Baseline* category shows the results without financial information.

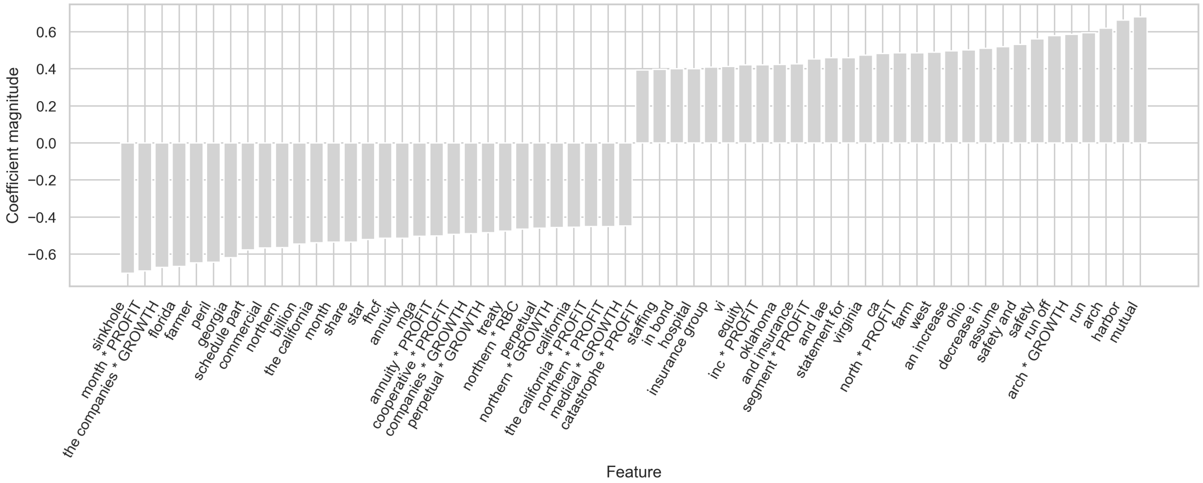
A.9 Strongest features including locations and abbreviations

Figure A.3: Most important feature coefficients in the prediction model with MD&A and financial indicators

(a) Model with MD&A and financial indicators



(b) Model with MD&A, financial indicators, and interaction terms



The figure shows features with the 30 largest and the 30 smallest coefficients for two classification models for over- and underreserving. Both models use a Stochastic Gradient Descent classifier with TF-IDF uni- and bigram word. Models include financial indicators for profit, growth (Median), and RBC ratio (Median). Word features that contain information on locations and abbreviations are shown in the figure.

A.10 Results for Gradient Boosting model

Table A.19: Classification results for a model integrating financial information on test set

	Precision	**Recall**	**F1**	**Accuracy**	**AUC**
0	0.5385	0.1750	0.2642		
1	0.7179	0.9333	0.8116	0.7000	0.5542
Weighted average	0.6627	0.7000	0.6432		

The table summarizes the performance scores of predictions on the hold-out test set of the alternative classification models. Accuracy is defined as the share of correct predictions. Precision measures the share of true positives (that is, the case of overreserving in our model) among all predicted positives. The F1 score is calculated as the harmonic mean of precision and recall, which measures the share of predicted positives among all actual positives. The AUC estimates the probability that a random positive is ranked before a random negative (in our model, the case of underreserving), without specifying a particular decision threshold. The weighted average calculates the support-weighted mean per label. The prediction model is a Gradient Boosting classifier with TF-IDF unigrams of MD&A sections and profit, growth (Positive), concentration and RBC ratio (Median) indicators.