Justus-Liebig-University Gießen
Faculty of Economics
Chair of Statistics and Econometrics
(VWL VII)

Prof. Dr. Peter Winker

A Study of Agency Problem in Companies 10-K Disclosures

Text Mining in the Financial Sector

(Seminar)

presented by:

Lenon Ferreira
lenon.ferreira@wirtschaft.uni-giessen.de

Master of Science

Advisor:
Elena Tönjes

Gießen, January 15, 2024

Contents

1 Introduction

In today's dynamic and ever-evolving financial landscape, the effective dissemination of information plays a pivotal role in shaping investor perceptions, influencing market behavior, and ultimately, determining the allocation of capital. Among the myriad channels through which information flows, corporate disclosures hold a central position, serving as the primary conduit through which companies communicate their performance, prospects, and strategic intent to stakeholders. Traditionally, the majority of studies in the fields of finance and accounting have predominantly focused on explaining company performance through quantitative metrics, relegating qualitative information to limited attention. However, a comprehensive analysis of a company's processes, financial performance, and operational practices demands the inclusion of such data as they provide valuable insights into the underlying factors that generate quantitative metrics, including those related to corporate governance (Siano and Wysocki 2019: 8). The underlying study applies text mining technique to study the behavior of managers in communicating with their stakeholders. Specifically, we focus on the Management's Discussion and Analysis of Financial Condition and Results of Operations (MD&A) of the 10-K disclosures filed by US-public listed companies yearly to the Securities and Exchange Commission (SEC). Based on agency and neoclassical theory of utility maximization, we hypothesize that managers have an incentive to hide relevant information from the market in order to obtain or maintain their favorable position, we further hypothesize that the tone managers use in their MD&A increases the systematic risk of their companies when the tone used in the reports are not in line with their performance.

2 Literature Review and Hypothesis Definition

The availability of precise, reliable and timely information is a cornerstone for well functioning capital markets. In a perfect market the price of a share is informative to investors about the financial soundness of a company and of its expected future performance. Nevertheless, in reality financial markets are not perfect and they are full of frictions due to agency and information costs as well as behavioral biases. When the market fails to absorb relevant informational contents, securities are mis-priced leading to bubbles and financial crises with severe socioeconomic consequences and

even some doubts on the so-called Efficient market hypothesis (Malkiel 2011; Gielnson and Kraakman 2019; Bloomfield 2002: 233).

In the US, the Securities Exchange Act of 1934 requires companies to submit comprehensive financial disclosure to the Securities and Exchange Commission (SEC), aiming to enhance transparency and fairness in capital markets. One of these reports and focus of this research is the so-called 10-K, an annual file that offers a detailed overview of a company's business, risks, and financial performance, allowing managers to report not only financial metrics but also to explain their causes and expectations for the future. Over the years, the SEC has implemented initiatives like the Plain English Rule and the Sarbanes-Oxley Act to improve reporting quality and corporate governance. Evidence suggests these efforts have enhanced the informational content of reports, with studies showing improved risk prediction accuracy (see Kogan et al., 2009: $277 - 278$) and readability (see Bonsall IV and Miller, 2017: 627-631) using data for the period after the actions took place. However, despite regulatory efforts, the amount of and how the information is disclosed in the narrative sections of the 10-K is still at the discretion of the manager, allowing for variability in transparency levels, potentially leading to managers writing lengthy and redundant texts, causing information overload and increasing processing and information costs to stakeholders (Cazier and Pfeiffer, 2016: 10-15).

In an ideal market scenario, managers are incentivized to consistently prioritize the interests of their stakeholders, as failure to do so would result in penalties imposed by market participants, implying their disclosures to be concise and lucid. However, in an imperfect market, neoclassical theory posits that managers prioritize maximizing their own utility, and these, very often, diverge from that of their stakeholders. As exemplified by Alchian and Demsetz (1972: 783), and Jensen and Meckling (1976: 308-310) a firm is formed by a nexus of contracts governed by agency relationships and, thus, source of agency problems. To ensure that the manager (the agent) acts on best behalf of the principal (stakeholders), the latter need to create incentives or a monitoring system, incurring agency costs. A single investor will not spend much effort in monitoring managers, as they can easily move their money elsewhere, hence, monitoring is allocated to institutional investors and the capital market, consequently, it creates an incentive for managers to use their reports to either emphasize the company's fundamentals or obscure poor performance by withholding or burying relevant information in

lengthy and ambiguous texts (Bloomfield 2002: 238). Klueger and Shields (1991) find evidence of managerial opportunism showing that managers of companies in financial distress withhold information from the market by changing non-compliant auditors. The arguments brought so far leads us to ascertain our first hypothesis:

- H_1^1: The MD&A section of poorly performing companies will exhibit a lower degree of readability than otherwise.

The hypothesis finds support in the study of Kim, Wang and Zhang (2019) that show that stocks of companies with less readable reports have a higher stock price crash risk, and the study of Li (2008: 234-237) finds that companies with less "foggy" reports present higher current and future earnings than those of companies with less readable reports. Following those authors, this study uses the Gunning Fog Index (GFI) as a measure of readability for the MD&A sections of the 10-K reports. Given a document i at fiscal year t, the index is calculated as:

$$GFI_{i,t} = \left(\frac{\text{words}_{i,t}}{\text{sentences}_{i,t}} + \frac{\text{complex words}_{i,t}}{\text{words}_{i,t}} * 100 \right) 0.4 \tag{2.1}$$

where a complex word is any word with more than 2 syllables. One of the major criticisms posed on the GFI (see Loghran and McDonald, 2014: 16-45) is that many words considered as complex are not complex for the average reader of financial reports, and other proxies like the file-size or the BOG-index as proposed by Bonall IV, and Miller (2017) may be more appropriate. The file size itself has its own shortcoming (see Bonall IV, Leon and Miller, 2017: 344-345), and we do not have access to the BOG index. Fig. 1 illustrates a word-cloud created using the most frequent complex words found in the MD&A-sections of the 10Ks, words like financial, operating, December and company, are considered complex but of easy readability even for a person with no financial background. Performance is evaluated using Jensen's alpha, which measures the deviation between a company's actual return and the return expected by the market at a given time. A positive (negative) Jensen's alpha is a sign of good (poor) performance (see Jensen (1968: 393) for computation).

Fig. 1: Word Cloud: Complex Words

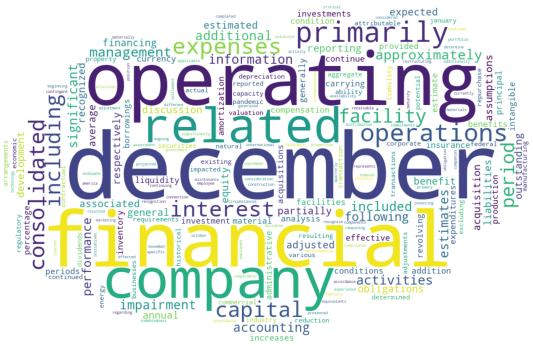

Figure illustrates the complex words as defined by the Fog-Index found in the MD&A of the companies in the data-set. To construct the word-cloud a list of complex words was created and then vectorized using a BAG-of-Words approach. The larger a word is pictured, the more often this word appeared in the corpus.

According to Luo and Zhou (2020: 105-106) managerial opportunism can also be reflected in the tone managers use in writing their reports, i.e., how positively or negatively managers explain the events happening to their companies. Kang, Park and Han (2018: 376-378) show in their study that managers using excessively positive tone in the narrative sections of their 10-K reports, i.e. not aligned with current performance, cannot guarantee the same level of future performance in future period, more precisely, for that group of companies, a positive tone in current period means a decrease in earnings in the subsequent year, indicating opportunistic behavior of the management. Such behavior may help the management of the companies in the short but is unlikely to be beneficial in the long-run, Rogers, van Buskirk and Zechman (2011: 2170-2173) demonstrate in their study that, ceteris paribus, companies facing a lawsuit have plaintiffs quoting passages with "good tone" that the sued companies made to motivate the damage to their clients, associating optimism in reports with higher litigation risk. Further, Bushee, Taylor and Zhu (2023) find evidence that managers increase their voluntary disclosure before conferences, and that these are especially positive. This behavior, in conjunction with insider trading, suggests that the management of the companies are manipulating their stock prices to obtain self-advantage, however, such a doing increases the probability of future lawsuits (p. 51-52) by the those companies. In summary, poor performance in conjunction with positive tones in reports

make a company riskier in relation to other companies and the market, resulting in the second hypothesis of this study:

- H_1^2: The positive tone used by managers in their MD&A-sections will result in an increase of future systematic risk, given that the company performed poorly in the period of report.

The subsequent section details the process of data collection and the variables in the data-set. Section 4 provides information on the methodology used to compute management's tone. Within Section 5, regression analysis is conducted to assess the hypotheses outlined. Specifically, H_1^1 is examined by regressing the Gunning Fog Index (GFI) on a performance indicator while controlling for relevant variables that could influence the GFI. To test H_1^2, the beta of the company, calculated using 52-week returns at the end of the reporting year, is regressed against the interaction of the performance indicator with a proxy for management's tone. This regression also includes controls for variables that may impact the company's risk.

3 Data Description

In order to compute the variables for testing the hypotheses stated in Section 2, a dataset containing information from multiple sources was constructed. The 10-K reports were obtained from the SEC website; the SEC provides an API and a list of ticker that can be used to acquire metadata on the respective companies. The list contained $10,713$ companies from which 394 remained in the final dataset. A web scraper was created to download and store the files locally in html format, totaling $3,152$ reports for the fiscal years between 2015-2022. In a second step the class provided by (Loukas et. al (2021) was applied to the corpus to remove the html-tags and numerical tables from the files, and store them in json-format with each key of a file carrying a 10-K section as value. A company remained in the dataset if it consistently submitted its 10-K between the dates $02/15$–$03/15$, reliable financial data could be queried using CapitalIQ excel Add-in, its MD&A sections had at least 3 sentences and its daily share prices, used to compute Jensen's alpha, was available in Yahoo-Finance API. The estimation of Jensen's alpha requires the systematic risk of the company, obtained directly from Capital IQ. This beta is computed using the 52-week returns of the companies, with the S&P500 used as a benchmark for the market portfolio for US companies. Ensuring consistency, the

S&P500 is also employed in estimating Jensen's alpha and the US 10-year bond yield at constant maturity, obtained from the website of the Federal Reserve of St. Louis is used as a proxy for the risk-free rate of return.

Tab. 1 presents descriptive statistics on the variables employed in the calculation of the GFI, laying particular focus on the MD&A-sections of the reports. On average, the MD&A-sections in the corpus consist of $10,877$ words, with $2,395$ (22.02%) being considered complex. These words are distributed across 379 sentences, resulting in a GFI of 20.28 points. According to Li (2008: 226), any text with a GFI above 18 is deemed unreadable. Notably, even the 25th percentile of the variable in the corpus exceeds this threshold, indicating a high level of complexity. Interestingly, while all the components of the GFI has shown a downward trend since the fiscal year 2015, the GFI itself has steadily increased as illustrated in Fig. 2. Comparing the files submitted in the year 2023 with those in 2016, the average number of words fell by 21.07%, the number of sentences by 10.89%, and the total number of complex words by 18.79%, not enough to make the reports more readable according to the GFI.

Used as control-variable to isolate accounting from market performance in the regression analysis, Almans' z–scores for the companies were computed using the financial data from CapitalIQ. The z-score was chosen as an aggregated measure of accounting performance in order to avoid the multicollinearity problem inherent to financial ratios. The z-score is a measure created by Altman (1968) to discriminate between companies facing high risk of bankruptcy from sound ones (see Altman and Hotchkiss, 2006: 241 for further details on its computation.) We further calculate the annualized 52-week volatility of the stock returns for each fiscal year t and employ them as control to test our first hypothesis. Additionally, we use the logarithmic value of the market capitalization as a proxy for company complexity and a cosine-similarity-score between the MD&A of company i for the years t and $t-1$, vectorized by first training a doc-to-vec model; the computation of similarity score lead us to drop the first fiscal year, i.e., 2015. We also account for text richness, defined as the ratio of the total number of unique words to the total number of words. This measure, as the similarity-score, ranges between 0 and 1, with one representing the highest achievable richness in the text.

Tab. 1: Descriptive Statistics for MD&A Sections

	GFI							Words						
Filing Year	mean	std	min	25%	50%	75%	max	mean	std	min	25%	50%	75%	max
2016	19.97	1.48	14.4	19.03	19.84	20.89	26.31	12003.52	5516.73	39.0	8691.0	11024.5	14468.0	44955.0
2017	19.98	1.46	14.59	19.02	19.91	20.91	27.16	11659.05	5185.0	39.0	8597.5	11076.0	14034.0	31132.0
2018	20.08	1.49	14.44	18.99	20.03	21.01	25.71	11647.57	5091.24	39.0	8402.25	11104.0	14234.0	36465.0
2019	20.12	1.51	14.41	19.11	20.07	21.07	24.98	11410.77	5286.66	39.0	7811.5	10681.5	13949.75	44268.0
2020	20.26	1.47	16.5	19.31	20.15	21.2	25.47	10116.99	4945.83	39.0	6775.25	9389.5	12463.75	38516.0
2021	20.56	1.39	16.95	19.68	20.53	21.43	25.14	10683.86	5238.84	39.0	7092.0	9682.5	13305.5	38487.0
2022	20.64	1.43	17.07	19.65	20.59	21.47	26.17	10015.48	5076.65	39.0	6654.0	9275.0	12253.25	37861.0
2023	20.65	1.48	16.59	19.64	20.62	21.59	26.78	9473.87	4965.39	39.0	6280.5	8522.0	11493.5	33346.0
mean (all years)	20.28	1.46	15.62	19.3	20.22	21.2	25.97	10876.39	5163.29	39.0	7538.0	10094.38	13275.22	38128.75
total (all years)	–							34,282,380						
	Compex Words							**Sentences**						
Filing Year	mean	std	min	25%	50%	75%	max	mean	std	min	25%	50%	75%	max
2016	2615.46	1264.85	14.0	1858.0	2401.0	3149.0	10683.0	424.35	185.99	3.0	313.5	396.0	494.0	1570.0
2017	2534.94	1173.98	14.0	1860.5	2348.5	3047.0	7393.0	410.69	173.8	3.0	307.25	394.0	486.0	1100.0
2018	2533.74	1152.21	14.0	1782.0	2394.5	3114.0	8674.0	407.85	171.63	3.0	299.5	393.5	489.0	1274.0
2019	2489.42	1209.44	14.0	1701.75	2304.0	2991.0	10711.0	398.71	177.94	3.0	284.0	379.5	487.75	1491.0
2020	2240.2	1137.77	14.0	1481.5	2056.5	2714.5	9482.0	352.91	164.87	3.0	242.5	331.5	423.0	1342.0
2021	2380.98	1192.53	14.0	1569.0	2147.0	2914.25	9296.0	365.21	172.34	3.0	249.0	335.5	443.25	1256.0
2022	2237.89	1152.24	14.0	1493.5	2030.0	2721.75	9131.0	341.64	167.45	3.0	232.5	313.5	408.75	1260.0
2023	2123.93	1133.47	14.0	1388.75	1895.5	2517.0	8122.0	323.57	164.54	3.0	223.0	289.0	389.5	1122.0
mean (all years)	2394.57	1177.06	14.0	1641.88	2197.12	2896.06	9186.5	378.12	172.32	3.0	268.91	354.06	452.66	1301.88
total (all years)	7,547,686							1,191,823						

Table provides descriptive statistics for the the textual features of the MD&A sections of the 10-Ks disclosures submitted to the SEC in the period between 2016-2023 by the 394 companies included in the dataset. Accounting all years the MDAs have a total 34,282,380 from which 7,547,686 are complex, distributed among 1,1191,823 sentences. Despite the decreasing trend in the mean number of words, sentences, and complex words over the years, the Gunning Fog index has steadily risen, indicating a growing complexity in the managers' writing.

Fig. 2: Evolution of Amount of Words and GFI In 10-K Reports

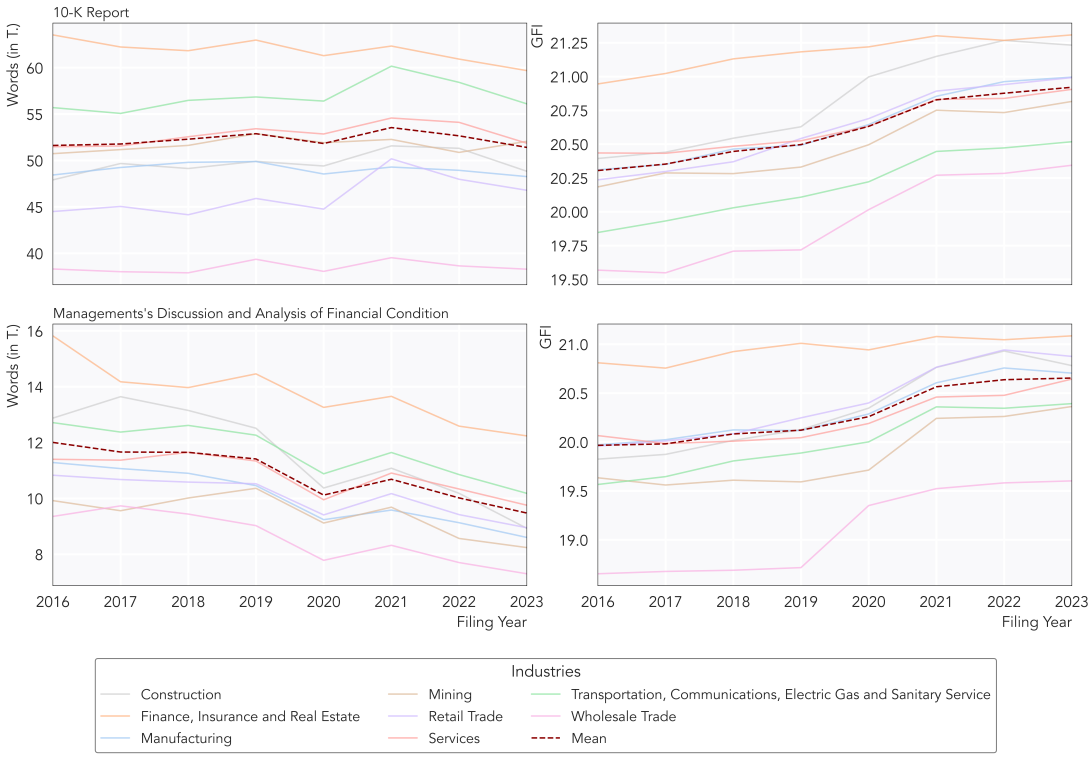

Figure illustrates the evolution of words and of the Gunning Fog Index (GFI) over time. The upper row considers the whole 10-K report, while the bottom row focus on the MD&A section. While a decreasing number of words could be observed for the MD&A section for the average report, the GFI index has continuously increased since 2016.

4 Computing Management's Tone

Management's tone, as elucidated by Kang, Park and Han (2018: 371) relates to how managers portray their company's performance and future prospects in the narrative section of their reports. Unlike standard sentiment analysis, which categorizes sentiment as being either positive or negative, i.e. a binary variable, tone-measurement, here referred to as polarity, is a continuous variable calculated in a similar fashion employed by Buschee, Taylor and Zhu (2023: 39). However, instead of relying on the count of words, the present study predicts the tone at a sentence level. Given a MD&A section of company i on fiscal year t, polarity is computed as:

$$polarity_{i,t} = \frac{\text{positive sentences} - \text{negative sentences}}{\text{positive sentences} + \text{negative sentences}} * 100 \qquad (4.1)$$

is a value in the interval $[-100, 100]$.

Sentiment analysis was initially applied to assess product reviews, given the abundance of data available, but is nowdays applied in a variety of fields (Medthat et al., 2014:

1093). One of the greatest challenges in extracting sentiment from texts in the financial domain it its specialized jargon. At its inception, the method relied on lexica and dictionaries to count or assign a score to positive and negative words in a document to determine its polarity (Hardeniya, Borikar 2016: 318), nevertheless, such approach failed in dealing with the evolution of the language and in capturing more complex structures. Further, dictionaries are specific to their own domain, Loughran and McDonald (2011: 46-48) address the problem in using standard dictionaries in financial report analysis, like the Harvard-IV-4 TagNeg (H4N) from psychology field and show that H4N misclassifies words like "cancer", "mine", and "capital" by assigning sentiment to them, although they are mostly merely descriptive regarding the companies business and do not carry any sentiment in them. To address this limitation, the authors developed a financial-specific dictionary tailored for 10-Ks and other financial reports. Over the years, sentiment analysis has departed from dictionary based approaches, evolving through a phase based on machine learning methods until it reached its state of the art by making use of transformers (Kheiri and Karimi, 2023: 2).

Before selecting a model to apply to MD&A sections in our corpus, we compared the performance of 4 different models: the Dictionary produced by Loughran and McDonald (2011), Support Vector Machine (SVM), XGBoost (XGB) and the transformer finBert trained by Juang, Wang and Yang (2020). Due to the lack of labeled data in our corpus, XGBoost and SVM were trained on the annotated data-set created by Malo et al. (2013: 9), publicly available on HuggingFace-website in different shapes according to annotators' agreement level. The threshold of 50%-agreement was chosen in order to increase sample size. The data-set used to train the model contains $4,846$ sentences from financial news on all in the OMX Helsinki listed companies, from which 604 (12.46%) are negative, $2,879$ (59.40%) neutral, and $1,363$ (28.12%) positive sentences.

XGBoost is a decision tree-based method. A decision tree classifier works by dividing the feature space using a series of splitting rules. These rules are determined based on the values of input features, ultimately leading to a prediction for each observation in the dataset. The goal is to find the best split at each step in order to minimize a specified loss function. A major drawback of the classic decision tree classifier method is that it easily overfits the data, meaning that it performs well in the training data but fails to generalize in data never seen (James et al. 2023: 333-343). An approach to improve

out-of-sample performance is called gradient boosting, consisting of an ensemble of models, where each prediction is made to correcting errors of past predictions, i.e., instead of trying to predict the outcome itself, after an initial prediction the algorithm uses the residual as input in building the trees in order to improve the model. For each set of residuals, a new tree is built (known as weak learner) and a new prediction is made, the residuals are again computed. The output values are added to the last prediction updating the function. The process is repeated until the model performance does not increase or the number of trees used is reached (see Chen and Guestrin: 2016 for further details).

In its most basic form Support Vector Machine uses a hyperplane that separates data points into different classes while maximizing the margin, which is the distance between the hyperplane and the nearest data point from each class. As a change in the position of these nearest points also causes the position of maximal-margin hyperplane to change, they are known as support vectors. The method described assumes that the boundaries of the classes are linearly separable, when this is not the case, the application of other kernels to transform the data in a linearly separable structure is needed (James et al. 2023: 370-372).

FinBert is a deep-learning model based Google's Bert architecture. Its training process consisted into two phases, the first one the model learned textual representations by using a masked model and next sentence prediction on a myriad of financial texts, among others 10-K reports. In a second step the model is trained to fit the purpose of a specific NLP task, in this case, sentiment analysis. One of the major advantages of finBert over the models presented so far, is that it is able to understand contextual, syntactic and semantic features of a document, contrasting to bag-of-words and one-hot- encode vectorization approaches that disregard the order and the semantic connection between words (Huang, Wang and Yang 2023: 1, 4, 10-13).

The training of the XGB and SVM models involved partitioning the data into training (80%) and testing (20%) splits; due to its imbalance, stratification was applied to better reflect the structure of the data in the splits. To optimize model performance, 5-fold cross-validation with grid search was applied to choose from a set of hyperparameters and vectorization methods that yielded the best macro-average F_1-score. The F_1 score is computed as the harmonic mean of precision and recall measures, and its macro-version was chosen because it gives more weight for rare categories, in the

case positive and negative sentences (Tan, 2005: 669). We tested two types of vectorization methods, namely, Term Frequency and Term Frequency-Inverse Document (TFIDF) allowing for bags of up to 2 grams and compared models on cleaned (using only alphanumeric characters, removing stop-words and applying lemmatization) and raw sentences, and chose the two best models, one for XGBoost and one for SVM for comparison with finBert and LM-dictionary. For both models, the best performance was achieved using raw-text and the vectorization using TF-IDF, SVM achieved the best performance (macro avg. $F_1 : 0.72$) with a linear kernel and bag of up to 2 grams, while XGB used only 1–gram (macro avg. $F_1 : 0.711$).

To ensure comparability between finBert, SVM, XGBoost and LM-dictionary, the same test set was used to assess their performance. When applying the dictionary, the polarity score was first computed at a word level as in Buschee, Taylor and Zhu (2023: 39); given a score greater (smaller) than 0.5 (-0.5), the sentence was assigned the class positive (negative), and neutral otherwise. Fig. 3 provides the confusion matrices for the out-of-sample classifications and a table with the F_1 achieved by each one of the models considered; finBert achieved the highest score (0.773), followed by SVM (0.733), XGBoost (0.718), and the poorest F_1 score was achieved using the LM-Dictionary (0.463).

Fig. 3: Performance Results: Models for Sentiment Classification

	LM-Dictionary	XGBoost	SVM	finBert
(Macro avg.) F1-Score	0.463	0.718	0.733	0.773

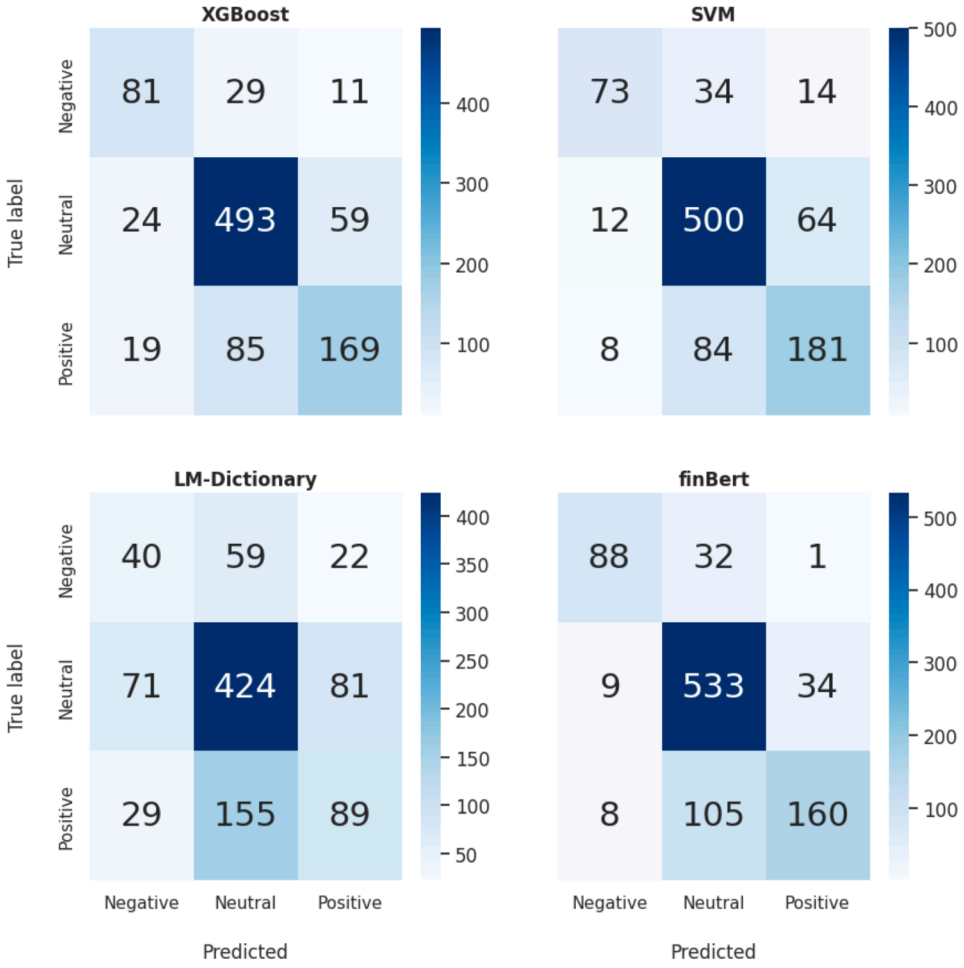

Figure illustrates the confusion matrices computed from the results of the classification models applied to the test set using the financial-phrasebanks dataset created by Malo et al. (2023). In the training phase, grid-search was applied to find the best set of hyper-parameters that maximizes the macro-avg. F_1 score for both the XGBoost and Support Vector Machine models. The best out-of-sample performance was achieved by finBert with a macro-avg. F_1 score of 0.773, followed by SVM with a score of 0.733, XGBoost with a score of 0.718 and the poorest performance was achieved by the Loughran-McDonald dictionary, with a score of 0.463.

Due to its superior performance, finBert was the chosen model for classifying the MD&A-sentences of the 10-K files. In total, the model classified $1,191,779$ sentences, from which 11.41% were positive, 9.3% negative, and 79.29% neutral. In a second step, we computed polarity as defined as the beginning of this section for the $3,152$ documents in the corpus. The mean polarity score across all documents was 11.79 points, with a standard deviation of 31.66, indicating a tendency to positiveness in the narrative

section, the shape of its distribution as well as a histogram of the classification of the sentences are illustrated in Fig. 4.

Fig. 4: Distribution of Sentiment and Polarity (MD&A-Section)

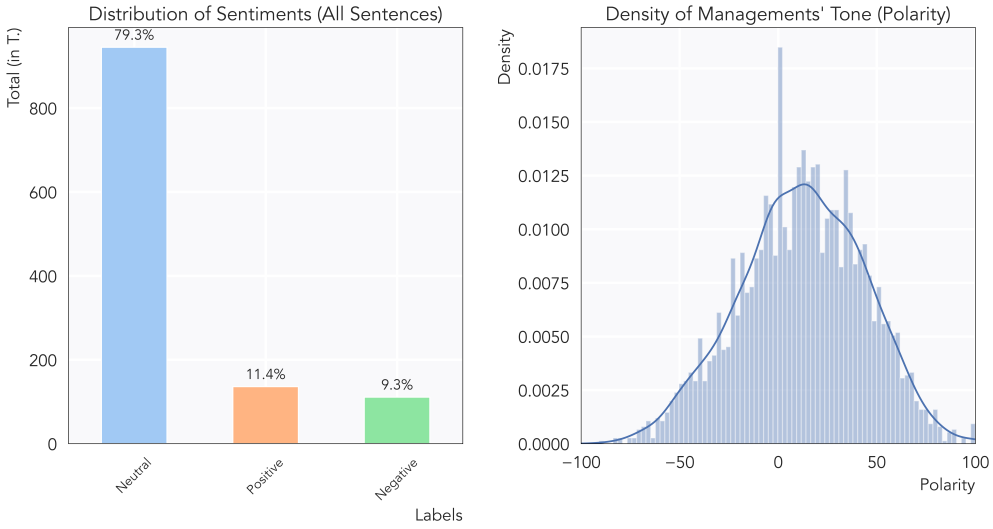

Figure illustrates the distribution of sentences in the corpus classified by finBert into negative, neutral and positive. Polarity, proxy used for the management's tone in the MDA of the 10-K files, is computed as the net count of positive and negative sentences divided by their sum and multiplied by 100, so that its value is bounded in the interval $[-100, 100]$. The great majority of sentences were predicted to be neutral, accounting for 79.32% of all sentences considered. 11.39% of the sentences were positive and 9.29% were estimated to have a negative connotation. In total, $1,202,109$ sentences were evaluated by the model, of which $953,613$ were neutral, $136,861$ positive and $111,634$ negative sentences.

5 Regression Estimation and Results

To assess the first hypothesis, which states that managers of poorly performing companies hide information from investors by writing less readable reports, we fitted the following fixed effects model:

$$\begin{aligned} GFI_{i,t} &= \alpha_i + \beta_1 poor\_perf_{i,t} + \beta_2 \log MC_{i,t} + \beta_3 lexdiv_{i,t} + \beta_4 z_{score(i,t)} \\ &\quad + \beta_5 sim_{i,t}^{(t,t-1)} + \beta_6 \sigma_{i,t} + \varepsilon_{i,t} \end{aligned} \tag{5.1}$$

where $poor\_perf_{i,t}$ equals 1 if company i had a negative Jensen's alpha in fiscal year t, and it is the variable of primary interest. If poor performance causes managers to hide information by writing complex reports, we would expect that $\beta_1 > 0$; that is, this group of companies would have a higher $GFI_{i,t}$ than companies performing well. $\log MC_{i,t}$ denotes the logarithm of company's i market value, and it is expected to be

positively correlated with the $GFI_{i,t}$, since larger companies are likely to have more complex structures and consequently, more complex matters to report on. $z_{score(i,t)}$ represents Altman's z-score and is expected to be negatively correlated with $GFI_{i,t}$, as managers are likely to be clearer and more precise in highlighting a company's strong fundamentals as a reflection of their own success. $sim_{i,t}^{(t,t-1)}$ is the similarity score in company i's report between years t and $t-1$; the measure is expected to be inversely correlated with the $GFI_{i,t}$, as managers willing to hide information from investors would not be consistent in their reports over the years. $\sigma_{i,t}$ represents the annualized 52-week asset return volatility of company i and is used as a proxy for market response to news, reflecting the amount of information the market can process. Higher volatility may lead managers to obscure information in response to increased market scrutiny and uncertainty, aiming to control how the information is perceived by the market.

Tab. 2 illustrates the results of the regression model presented in Eq. (5.1) in its first column. Consistent with H_1^1,, we find evidence that managers of poorly performing companies write foggier MD&A sections than companies showing good performance. The GFI of poorly performing companies is, on average, 0.1 point lower than that of successful ones, since the coefficient β_1 is significant even at the 1% significance level. The result is robust when replacing $poor\_perf_{i,t}$ by the annualized Jensen's alpha (α_j) in the regression model (not tabulated), in which case an increase in α_j of 1% leading to a decrease of 7×10^{-4} points in the GFI. The z-score shows no significant effect on GFI and, as expected, larger companies write more complex reports than smaller ones, with a 1% change in the MC causing a decrease of 0.039 points in the GFI. Furthermore, consistency in vocabulary and similarity between years causes the GFI to decrease; a 1% increase in $sim^{(t,t-1)}$ leads to an improvement in readability of 0.02 points, and a 1% drop in the lexical diversity $(lexdiv_{i,t})$ improves readability by 0.04 points.

Next, we test hypothesis H_1^2, which suggests that managements' positive tone in their MD&A sections, when not aligned with good performance, will cause future systematic risk to increase. The regression equation reads:

$$\begin{aligned}\beta_{i,t+1} &= \gamma_i + \gamma_1 poor\_perf_{i,t} + \gamma_2 polarity_{i,t} + \gamma_3 \log MC_{i,t} + \gamma_4 lexdiv_{i,t} \\ &\quad + \gamma_5 poor\_perf_{i,t} * polarity_{i,t} + \gamma_6 z_{score(i,t)} + \eta_{i,t}. \end{aligned} \tag{5.2}$$

If our hypothesis is supported, then we should observe a positive effect in the coefficient γ_5. We control for the company's market value, as small companies are known to be considered riskier by investors (ceteris paribus), for the $z_{score(i,t)}$, which is expected to be negatively correlated with $\beta_{i,t+1}$, since the score is an indicator created as measure of financial soundness of the company, with lower z-scores suggesting lower likelihood of bankruptcy. In further control for the lexical diversity of the MD&A sections, and expect it to be positively correlated with $\beta_{i,t+1}$, since the lexical diversity was positively correlated with the GFI.

The second column of Tab. 2 reports the result of the estimation of the model in Eq. (5.2). Contrary to expectations, we fail to reject the null hypothesis that the coefficient of the interaction (γ_5) between poor performance and managements tone is different from zero; more striking, none of the variables included in the regression as control have any explanatory power on future systematic risk, and the coefficient of determination of this regression is almost zero.

6 Conclusion

In our study, we assessed the behaviors exhibited by managers in writing the MD&A sections of 10-K reports. Grounded in agency and neoclassical theories of utility maximization, we hypothesized that managers might act opportunistically, prioritizing their own interests over those of stakeholders. This behavior could manifest through the deliberate obfuscation of information in complex reports and the adoption of an overly optimistic tone to hide true company conditions, consequently increasing the future systematic risk of the company. While our analysis did reveal a positive correlation between the textual complexity level of MD&A sections, as measured by the Gunning Fog Index (GFI), and our indicator for poor performance, the observed effect size was marginal. The difference in GFI between groups representing poor-performing companies and those representing good-performing ones was only 0.10 points. Notably, our sample suffered from the so-called survival bias, since we only had access to information of companies that still active in the market. Moreover, we failed to provide evidence that a misalignment between managerial tone and company performance contributes to increased future systematic risk. Despite theoretical support for this hypothesis, the regression model with future beta as dependent variable and the interaction of

Tab. 2: Regression Results

	Dependent variable:	
	GFI	β_{t+1}
	(1)	(2)
$poor\_perf$	0.100***	0.013
	(0.030)	(0.030)
$polarity$		-0.0003
		(0.001)
$\log MC$	0.150***	-0.045
	(0.045)	(0.055)
$lexdiv$	3.981***	0.427
	(1.359)	(0.666)
$poor\_perf : polarity$		0.001
		(0.001)
z_score	-0.007	0.005
	(0.008)	(0.007)
$sim^{(t,t-1)}$	-2.003**	
	(0.984)	
σ	0.743***	
	(0.090)	
Observations	2,758	2,758
R^2	0.099	0.003
Adjusted R^2	-0.053	-0.166
F Statistic (df = 6; 2358)	43.256***	0.989

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

The table presents the results of fixed-effects regressions for two dependent variables: (1) the GFI and (2) the future systematic risk of a company, quantified by its β_{t+1}. The results of regression (1) indicate that companies performing poorly in the fiscal year of their 10-K report produce MD&A sections tha are less readable compared to companies with better performance. However, contrary to the hypothesis, the interaction between poor performance and the polarity score in regression (2) is not significantly different from zero. This leads us to fail to reject the null hypothesis that a positive management tone alignedwith poor performance would result in higher future systematic risk for the average company.

performance and management's tone as one of the covariates yielded a coefficient of determination approximately equal to zero and none of the variables was statistically significant different than zero. Our study also explored the use of different machine learning models for sentiment analysis, including SVM, XGBoost, LM-Dictionary and finBert. We found that finBert achieved the highest performance in classifying sentiment in MD&A sections. Overall, our study contributes to the literature on financial reporting and managerial behavior. By gaining knowledge into the factors influencing readability and tone in financial reports, investors and other stakeholders can enhance their ability to interpret information disclosed by companies, facilitating more informed decision-making. As a suggestion for future research, it would be valuable to incorporate bankrupt or financially distressed companies into the sample and conduct a similar analysis to assess potential changes in results.

References

[1] Alchian, A.A. and H., Demsetz (1972), Production, Information Costs, and Economic Organization, The American Economic Review 62, p. 777 - 795.

[2] Altman, E.I. and E., Hotchkiss (2006), Corporate Financial Distress and Bankruptcy, John Wiley & Sons, Inc, New Jersey.

[3] Bloomfield, R.J. (2002), The "Incomplete Revelation Hypothesis" and Financial Reporting, American Accounting Association 16, p. 233 - 243.

[4] Bonsal IV, S.B., A.J., Leone and B.P., Miller (2017), A plain English measure of financial readability, Journal of Accounting and Economics 63, p. 329 - 357.

[5] Bonsall IV, S.B. and B.P., Miller (2017), The impact of narrative disclosure readability on bond ratings and the cost of debt, Rev Account Stud 22, p. 608 - 643.

[6] Cazier, R.A. and R.J., Pfeifer (2016), Why are 10-K Filings So Long?, Accounting Horizons 1, p. 1 - 21.

[7] Chen, T. and C., Guestrin (2016), XGBoost: A Scalable Tree Boosting System, URL: https://doi.org/10.48550/arXiv.1603.02754, query date: 15.01.24, 23:21.

[8] Gilson, R.J. and R., Kraakman (2014), Market Efficiency after the Financial Crisis: it's Still a Matter of Information Costs, Virginia Law Review 100:313, p. 313 - 376.

[9] Hardeniya, T. and D.A., Borikar (2016), Dictionary Based Approach to Sentiment Analysis - A Review, International Journal of Advanced Engineering, Management and Science (IJAEMS) 2, p. 317 - 322.

[10] Huang, A.H., H., Wang and Y., Yang (2023), finBert: A Large Language Model for Extracting Information from Financial Text, Contemporary Accounting Research 40, p. 759 - 841.

[11] James, G., D., Witten, T., Hastie, R., Tibshirani and J., Taylor (2023), An Introduction to Statistical Learning with Applications in Python, Springer, New York.

[12] Jensen, M.C. and W.H., Meckling (1976), Theory of the Firm: Managerial Behavior, Agency Costs and Ownership Structure, Journal of Financial Economics 3, p. 305 - 360.

[13] Kheiri, K. and H., Karimi (2023), Sentiment GPT: Exploiting GPT for Advanced Sentiment Analysis and its Departure from Current Machine Learning, URL:

https://arxiv.org/pdf/2307.10234.pdf, query date: 15.01.24, 09:56.

[14] Kim, C.F., K., Wang and L., Zhang (2019), Readability of 10-K Reports and Stock Price Crash Risk, Contemporary Accounting Research 36, p. 1184 - 1216.

[15] Kluger, B.D. and D., Shields (1991), Managerial Moral Hazard and Auditor Changes, Critical Perspective on Accounting 2, p. 225 - 272.

[16] Kogan, S., D., Levin, B.R., Routledge, J.S., Sagi and N.A., Smith (2009), Predicting Risk from Financial Reports with Regression, in: M., Ostendorf, M., Collins, S., Narayanan, D.W., Oard and L., Vanderwende (ed.) , Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, NA, Association for Computational Linguistics, p. 272-280.

[17] Li, F. (2008), Annual report readability, current earnings, and earnings persistence, Journal of Accounting and Economics 45, p. 221 - 247.

[18] Li, F. (2011), Textual Analysis of Corporate Disclosures: A Survey of the Literature, NA , University of Michigan, Michigan.

[19] Loughran, T. and B., McDonald (2011), When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks, The Journal of Finance LXVI, p. 35 - 65.

[20] Loughran, T. and B., McDonald (2014), Measuring Readability in Financial Disclosures, The Journal of Finance LXIX, p. 1643 - 1671.

[21] Luo, Y. and L., Zhou (2020), Textual tone in corporate financial disclosures: a survey of the literature, International Journal of Disclosure and Governance 17, p. 101 - 110.

[22] Malkiel, B.G. (2011), The Efficient-Market Hypothesis and the Financial Crisis, NA No. NA, Princeton University, Princeton.

[23] Malo, P., A., Sinha, P., Takala and P., Korhonen (2013), Good Debt or Bad Debt: Detecting Semantic Orientations in Economic Texts, URL: https://arxiv.org/pdf/1307.5336.pdf, query date: 19.12.23, 11:20.

[24] Medhat, W. and A., Hassan (2014), Sentiment Analysis algorithms and applications: A survey, Ain Shams Engineering Journal 5, p. 1093 - 1113.

[25] Siano, F. and P., Wysocki (2019), The Primacy of Numbers in Financial and Accounting Disclosures: Implications for Textual Analysis Research, Questrom School of Business, NA.