A Computational Analysis of Financial and Environmental Narratives within Financial Reports and its Value for Investors

Felix Armbrust[1], **Henry Schäfer**[1] and **Roman Klinger**[2]
[1]Institute of Business Administration
[2]Institut für Maschinelle Sprachverarbeitung
{felix.armbrust,henry.schaefer}@bwi.uni-stuttgart.de
roman.klinger@ims.uni-stuttgart.de

Abstract

Public companies are obliged to include financial and non-financial information within their corporate filings under Regulation S-K, in the United States (SEC, 2010). However, the requirements still allow for manager's discretion. This raises the question to which extent the information is actually included and if this information is at all relevant for investors. We answer this question by training and evaluating an end-to-end deep learning approach (based on BERT and GloVe embeddings) to predict the financial and environmental performance of the company from the "Management's Discussion and Analysis of Financial Conditions and Results of Operations" (MD&A) section of 10-K (yearly) and 10-Q (quarterly) filings. We further analyse the mediating effect of the environmental performance on the relationship between the company's disclosures and financial performance. Hereby, we address the results of previous studies regarding environmental performance. We find that the textual information contained within the MD&A section does not allow for conclusions about the future (corporate) financial performance. However, there is evidence that the environmental performance can be extracted by natural language processing methods.

1 Introduction

Understanding the textual information in corporate disclosures is an important part of financial accounting research. It provides additional information about numerical financial data and aids the reader to understand the managerial information sets. In addition, the manager's communication patterns can reveal certain behavioral traits – e.g., in case the company provides truthful and meaningful information or if the company tries to obfuscate or withhold information. The textual information can, hence, guide users of corporate disclosures to better understand the firm's behavior and decisions (Li, 2010b, p. 143–144) – in particular, when information asymmetries exist (Fields et al., 2001, p. 256). However, corporate disclosures often tend to be boilerplate (Palmiter, 2015; Bloomfield, 2008; SEC, 2003). As a consequence, the Securities and Exchange Commission (SEC, 2003), concerned that a substantial part of the corporate disclosures includes generic language and immaterial detail, issued a guideline to elicit more meaningful disclosure. In 2010, the SEC further published an interpretive release on climate change related disclosures "to remind companies of their obligations under existing federal securities laws and regulations to consider climate change and its consequences as they prepare disclosure documents", (SEC, 2010, p. 6297). Therefore, the company's disclosures should not only consist of financial narratives but also contain information about the environmental aspects concerning the firm. Whether these corporate disclosures are truly informative (Li, 2010a, p. 1050) or are just boilerplate generic disclosures (Wasim, 2019; Palmiter, 2015; Li, 2010b) remains an empirical question, which we address in this paper.

While in the financial domain, the financial narratives of corporate reports have been analysed with respect to dictionary-based approaches or standard classification methods like naïve Bayes (Kearney and Liu, 2014, p. 174), little research has focused on more sophisticated methods (Gentzkow et al., 2019,

This work is licensed under a Creative Commons Attribution 4.0 International License.
License details: http://creativecommons.org/licenses/by/4.0/.

p. 569). This equally applies for studies analysing the non-financial information[1] content (Kölbel et al., 2020, p. 8). The most common methods for analysing the non-financial, narrative information content still remain a manual- or dictionary-based approach (Berkman et al., 2019; Matsumura et al., 2018; Reverte, 2016; Verbeeten et al., 2016; Clarkson et al., 2008; Cormier and Magnan, 2007, i.a.). Also, only few studies concerning non-financial information focus on the actual *narrative* information content, and rather address the *quantity* of non-financial information published (Hummel and Schlick, 2016).

In this paper, we analyse the information content of the Management's Discussion and Analysis of Financial Conditions and Results of Operations (MD&A) section of 10-K and 10-Q filings with respect to the financial performance and environmental performance using a variety of natural language processing (NLP) methods. Our contributions are that we, (1), evaluate if the MD&A section contains information that can be used to predict financial and environmental performance in a machine learning setting, (2), analyse if the environmental information complements the financial information in explaining the future financial performance.[2] As more than 2000 studies have linked non-financial performance to financial performance (Friede et al., 2015), we assume that the narrative *per se* also contributes to the financial performance. Hence, we directly train the classifier on the financial performance and the environmental performance using a sentence-wise BERT- and word-wise GloVe-embedding model, instead of regressing a certain measure on a disclosure-score previously derived from either a manually constructed word-list or a tfidf-weighting scheme, like most other studies.

The hypothesese we discuss in this paper are: (H1) Corporate disclosures, particularly the MD&A section, contain information regarding the company's financial performance. (H2) Corporate disclosures, particularly the MD&A section, contain information regarding the company's environmental performance. (H3) Narrative, non-financial information within the MD&A section, complements narrative financial information in explaining the financial performance, i.e., environmental performance serves as a mediator for financial performance.

2 Background on Financial and Non-Financial Disclosures and Related Work

2.1 Computational Linguistics in the Financial Domain

The field of computational linguistics has made enormous progress in converting text into meaningful representations for computational use (Gentzkow et al., 2019, p. 537). Content analysis – a technique that allows users to objectively and systematically identify specific characteristics from a text (Stone et al., 1966) – can lead to interesting insights to transparently analyse text and provide evidence for (economic) theories. A field of research includes the application of content analysis to efficiently examine the level of corporate disclosures (Grüning, 2011). However, in the financial domain, applications of more recently developed methods are rather in their early stages (Gentzkow et al., 2019). The most common methods still remain a dictionary-based approach or comparably, straightforward naïve Bayes approaches (Kearney and Liu, 2014, p. 174) or dictionary-based methods which lack coverage.

A complementary strain of research in computer science does use state-of-the-art methods, however, with applications in financial data which tend to primarily focus on news articles predicting stock prices in the short-term (Mishev et al., 2019; Dodevska et al., 2019; Day and Lee, 2016; Curme et al., 2015; Li et al., 2014, i.a.). Only recently, more advanced methods also find their way into the financial domain (Kölbel et al., 2020; Luccioni and Palacios, 2019, i.a.).

2.2 Financial Narratives

In the United States, all public companies are required to prepare audited financial statements, which have to be filed with the SEC and need to be accompanied by a narrative explanation to aid the users in

[1] While the terms business sustainability, corporate social responsibility (CSR) or environmental, social, governance information (ESG) are often used interchangeably in the financial literature (Rezaee, 2015, p. 2), we generally refer to it as non-financial information in this paper. However, in our specific case, we only consider the environmental aspect of non-financial information.

[2] Similarly suggested by Reverte (2016) and Semenova et al. (2010). Considering the direct and indirect effect of non-financial reporting, Reverte (2016) finds that the integration of non-financial information into financial investment analysis provides a richer understanding of the companies' long-term performance. Further, evidence from Semenova et al. (2010) indicates that environmental and social performance are value-relevant and complement financial information.

assessing the company's situation (Yang et al., 2018, p. 45). The SEC had recognized that a numerical presentation alone might be insufficient for investors (SEC, 1987).

One of the most read components of the corporate filings include the MD&A section (Tavcar, 1998). It provides investors with a company's long- and short-term analysis from the management's point of view including narratives on past performance as well as potential future prospects (Muslu et al., 2008; Griffin, 2003). This analysis should also include prospective matters, known material[3] trends and uncertainties relevant to the company (SEC, 2003).

The Regulation S-K and the SEC's guidelines make the MD&A section mandatory. However, the MD&A sections of the 10-K and 10-Q filings do not need to be audited (Hüfner, 2007), and the requirements for the MD&A section still allow for management discretion (Li, 2010a, p. 1053). Managers could withhold unfavourable information below a critical disclosure level (Verrecchia, 1983), or provide imprecise information. However, companies might face litigation risk when making misleading or fraudulent disclosures, which in turn serves as a disciplining tool (Li, 2010a; Kothari et al., 2009). As Kothari et al. (2009, p. 1643) notes, "the disciplining forces might be less operative when disclosures are qualitative and long-term in nature (e.g., discussion in MD&A section) rather than quantitative [...] and short-term".

Several studies have analysed the impact of corporate disclosures with respect to the financial information using the narrative information content: There is evidence for the disclosures affecting company's risk (Kravet and Muslu, 2013; Kothari et al., 2009; Li, 2006), future earnings (Moreno-Sandoval et al., 2019; Athanasakou and Hussainey, 2014; Li, 2010a), and, ultimately, firm value (Campbell et al., 2014; Jegadeesh and Wu, 2013; Feldman et al., 2008).[4] However, nearly all previous work relies purely on a word-count, word-phrase count or comparably, straightforward approaches. Thus, we hypothesise that the extraction of the information content will improve by using more advanced embedding methods or applying convolutional neural networks. While most studies tend to derive some kind of disclosure-score from the text and subsequently regress the score on financial performance or some other measure, we propose a deep-learning classifier being able to find relevant patterns in the text that provide conclusions about the financial (and environmental) performance.

2.3 Non-Financial Narratives

Various surveys have shown that institutional investors value climate-relevant information (Ilhan et al., 2019; Amel-Zadeh and Serafeim, 2018; CFA Institute, 2017) and more than 2000 studies linked non-financial information to financial performance (Friede et al., 2015). While climate-change related factors are also a risk to the company, manager's awareness and proper handling represent an opportunity for financial gain (Wasim, 2019; Jung et al., 2018).

The 2010 guideline on climate-change related disclosures highlights the importance of non-financial information within corporate reports, often referred to as environmental, social, and governance information (ESG) (Eccles and Stroehle, 2018; Eccles et al., 2012). The SEC mandates material factors to be disclosed within the MD&A section, the risk factors section, legal proceedings and the description of business (Eccles et al., 2012, p. 68). However, the principle-based approach the SEC applies makes companies autonomously responsible to identify the relevant risks and factors. Passages that address climate-change related disclosures are hard to identify (Luccioni and Palacios, 2019) and, thus, the disclosure requirements have been criticized as to be lax and ineffective (Wasim, 2019; Palmiter, 2015). Nevertheless, corporate disclosures slightly improved after the 2010 guidance (Palmiter, 2015).

The theory of discretionary-based disclosures (Verrecchia, 2001) applies to non-financial information as well (Hummel and Schlick, 2016; Clarkson et al., 2008). Previous studies found only a small percentage of companies explicitly mention climate-change in their annual report (Palmiter, 2015; Hirji, 2013). While some companies are disclosing information apart from the SEC's required disclosure regimes, these do not substitute the SEC mandated disclosures. Without a mandatory standardized framework, not all issuers will disclose, and a lack of reliability and comparability exists (Lee, 2020). However, authors

[3] Materiality refers to the relevance of an item to users of financial statements. For a more detailed discussion on materiality see Eccles et al. (2012).

[4] We refer the interested reader to other publications for more details (Kearney and Liu, 2014; Loughran and McDonald, 2016; Loughran and McDonald, 2019).

have shown that voluntary, standalone non-financial disclosures are positively associated with market value and negatively with cost of equity capital (Verbeeten et al., 2016; De Villiers and Marques, 2016; Reverte, 2016; Plumlee et al., 2015; Dhaliwal et al., 2014; Clarkson et al., 2013).[5] For 10-K forms, similar dictionary based risk measures indicate a negative association with firm value and a positive association with implied cost of capital (Berkman et al., 2019). As Matsumura et al. (2018) show, firms disclosing material factors have a lower cost of equity. Kölbel et al. (2020) provide evidence that disclosures on climate risk within the 10-K risk section affects the spreads of credit default swaps, by training a sentence-wise BERT algorithm on sample reports from the task force on climate-related financial disclosures, supported by humanly annotation. We, thus, hypothesize that the MD&A section provides similar valuable information for investors, which can be retrieved by directly training classifiers on the text without human interference.

3 Study Design and Research Methodology

Our aim is to examine whether NLP methods can capture the meaning of the MD&A section of corporate disclosures with respect to financial and non-financial narratives. If we find evidence, this is an indicator that such information is contained, if not, this is an indicator that this information might not be available, which, however, would motivate future work to investigate this further. Therefore, we directly train classifiers on the text in the MD&A section to predict financial *and* environmental performance.

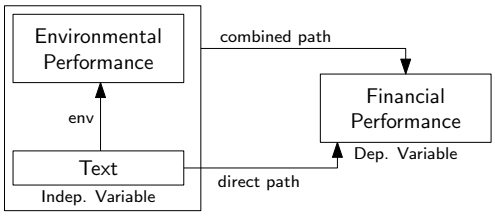


Figure 1: Our model predicts financial performance under consideration of environmental performance.

Figure 1 elicits our approach: First, we predict the financial performance from the MD&A section, i.e., *Direct Path*. Second, we predict the environmental performance from the text, represented by path *env*. Third, we build an extended model, which predicts the financial performance from the text while also considering the (predicted) environmental performance, i.e., *Combined Path*. In case all paths show a meaningful relation to their labels and the *Combined Path* shows an improved F_1-score compared to the *Direct Path*, we conclude that the environmental narrative complements the financial narrative in explaining the financial performance.

While the company might provide information in a timely fashion through other communication channels, e.g., ad-hoc messages, press releases and twitter posts, the MD&A section gives companies a more comprehensive way to discuss their future prospects and operating strategy. On top, the dissemination in the MD&A section is for free. Hence, we assume companies also provide previously released information within their MD&A (in more detail).[6] Furthermore, we assume that any material information from sustainability reports will also be published in the MD&A section.

Nevertheless, the impact of the financial information content of the 10-K and 10-Q filings might be quite limited since companies typically make an earnings announcement that includes information on their earnings, cash flows, and other key performance indicators approx. three weeks before disclosing the filings (Muslu et al., 2008; Lansford, 2006; Schrand and Walther, 2000). Hence, the majority of quantitative information is not expected to provide much information content to investors. However, the *narrative* information content probably will.

3.1 Data and Preprocessing

Our sample consists of S&P500 firms for the five-year period from the beginning of 2014 to the end of 2018 for which we are able to calculate all class labels (see Section3.2).[7] To minimise survivorship bias,[8]

[5]For an overview see, for example, Verbeeten et al. (2016, p. 1363-1365).

[6]For a confirmation of these assumptions see Muslu et al. (2008).

[7]The time horizon is limited by the ESG data available on Bloomberg since for most companies the sustainability data was not available before the beginning of 2014.

[8]By including *only* companies that were part of the S&P500 *throughout* the years 2014 till 2018, one would neglect poor performing companies that did not fulfil the S&P500 criteria at any point in time during the time horizon – namely having a too

we also include all companies that were constituents at any time during the observation period. This adds up to 615 companies under consideration, for which the 10-K and 10-Q reports were retrieved from the Notre Dame Software Repository for Accounting and Finance (SRAF) website (McDonald, 2019; Loughran and McDonald, 2016).[9] The files taken from SRAF are the entire reports containing all sections. We manually filter to the MD&A sections and only include the "Quantitative and Qualitative Disclosures About Market Risk" (QQD) section in addition, due to the fact that most companies include the QQD within the MD&A section. We opted against an automatic parsing procedure to avoid downstream errors, which would have been possible due to some companies reporting sections titled item 7 (MD&A) and 7A (QQD), but still include the *actual content* in a later part of the filing, e.g., in exhibit 13 or exhibit 99.1 – an additional attachment to the filings (Center, 2014). We lose a total of 2815 observations due to missing data. The final sample consists of 8772 observations (554 unique firms). The corpus includes 2.914.623 sentences and 56.063.378 tokens in total (45.872 unique tokens).[10]

3.2 Class Labels

As financial performance can be measured in several ways, we consider two sources for labels for the *Direct Path*. The first one uses the change in earnings per share as the according label[11] and the second one uses buy-and-hold returns as the class label (Barber and Lyon, 1997, p. 344)[12]. While earnings do not directly allow conclusions about the relevance of the filings to investors, it measures the informativeness of the MD&A section, i.e., how much information is actually conveyed that affects the future financial performance of the company. We denote this first label as $\Delta\mathrm{EPS}_{i,q}$, which is calculated as the difference in quarterly earnings previously reported and earnings published in the subsequent quarter. Here, q is the quarterly change and i denotes to a particular firm. For the buy-and-hold returns, we use the compounded SP500 index return, which we subtract from the contemporaneous, compounded return of the company under consideration.[13] Compared to Loughran and McDonald (2011) and Jegadeesh and Wu (2013), who use primarily a 3-day event window for the buy-and-hold-returns, we extend the period to 30 days to match the change in environmental performance ratings. This is important for the *Combined Path*, when considering the environmental narrative as well as the financial narrative. We denote the buy-and-hold-returns as $\mathrm{BHAR}_{i,t}$, which indicates how the information is perceived by investors (i again denotes the company, t denotes the 30-day period).

To measure the information conveyed by environmental narratives, we employ ESG data from Sustainalytics accessed via Bloomberg.[14] The ESG data is adjusted on a monthly basis. To match the earnings class label, we use a three-month time horizon for the environmental narrative. Hence, the first environmental class label is the change in the ESG percentile score denoted as $\Delta\mathrm{Env}_{i,q}$ and calculated as the change in the environmental percentile score over the subsequent quarter – measured by the time it takes the company to file the next report with the SEC. Index i is again for the company i and q denotes the quarterly change. The second class label for the environmental narrative uses a one-month change to match the second financial class label. The calculation is done in a similar fashion as for the quarterly environmental performance label and denoted as $\Delta\mathrm{Env}_{i,t}$ with i for the company and t for the day period. We calculated the label as the difference in the previously available percentile score on the filing date and the subsequent score available.

Again, the first class labels on financial and environmental narratives consider the quarterly change and the second labels the monthly impact of the MD&A section. Hence, we construct a long-term and a short-term model. While the long-term model considers the general informativeness and significance

small market capitalization in any of the years between 2014 till 2018. This would lead to logical errors and one would create biased statistics (see e.g., Carpenter et al. (1999), Brown et al. (1992)).

[9] Files and pre-processing description at `https://sraf.nd.edu/data/stage-one-10-x-parse-data/` (accessed 15.03.2020); unprocessed files are at `https://www.sec.gov/Archives/edgar/Feed/`

[10] The token number for BERT varies, as BERT uses its own tokenizer, see Devlin et al. (2019).

[11] Similar to Li (2010a).

[12] Similar to Loughran and McDonald (2011) and Jegadeesh and Wu (2013).

[13] The t-test $\mathrm{t}_{\mathrm{BHAR}} = \frac{\overline{\mathrm{BHAR}_{i,t}}}{\sigma(\mathrm{BHAR}_{i,t})/\sqrt{n}}$ showed that BHAR are significant to the 1%-significance level.

[14] To proxy environmental performance, other studies have also used data from KLD (today MSCI), Asset4, Viegeo Eiris, or single emission data.

of corporate disclosures, the second model considers the informativeness of corporate disclosures with respect to investors. We binary encode all labels, i.e., converting positive $\mathrm{BHAR}_{i,t}$, $\Delta\mathrm{EPS}_{i,q}$, $\Delta\mathrm{Env}_{i,t}$ and $\Delta\mathrm{Env}_{i,q}$ to 1 and negative values to 0, respectively. The idea for the *Combined Path* is to predict the environmental class label first and then using the environmental label to predict the financial label from text.

3.3 Model Architecture and Feature Extraction

We use five different feature extraction methods and model architectures: (1) **BOW**: Bag-of-words in a maximum entropy classifier. (2) **TF-IDF**: Bag-of-words weighted with TF-IDF in a maximum entropy classifier. (3) **GloVe**: Embeddings (Pennington et al., 2014) in a maximum entropy classifier. (4) **CNN**: GloVe embeddings with a convolutional neural network, following the architecture by Zhang and Wallace (2017) (three convolutional layers, each followed by max-pooling, each with drop-out, followed by dense layer with drop-out and a sigmoid output layer; filter sizes of 2, 3, 4; 2 filters for each length; dropout rate 0.5). (5) **BERT**: BERT sentence-embeddings followed by a Bidirectional-GRU, a relu layer, an attention layer with drop-out and a fully-connected sigmoid layer. For the BERT embeddings, we split each document into sentences using NLTK (Bird et al., 2009), encode each sentence with the required [CLS]-token using the BERT tokenizer and do a feed-forward pass to the BERT base network (Devlin et al., 2019).[15] The resulting [CLS]-tokens can be considered a vector representation of the sentences (Alammar, 2019, i.a.). Due to BERT base model input length restrictions of 512-tokens (Devlin et al., 2019), we limit each sentence to the first 200 tokens.[16] For models (1), (2), (3) and (4), we remove stop words using NLTK (Bird et al., 2009), non-alphabet characters and words that are shorter than two characters. Since BERT can make use of sentences that are grammatically correct, we do not remove any tokens for BERT. For all models, we use the adam optimizer (Kingma and Ba, 2015) and binary cross-entropy as a loss function. All models are fit using *50* epochs and a batch size of 32, except for the BERT model, which is trained for only *10* epochs. For hyperparameter optimization, we evaluate the following settings: We use dense layers with sizes of $\{50, 100\}$ for each model with an additional dense layer and set the drop-out probabilities to a probability of $\{40\%, 50\%, 60\%\}$. For the CNN, we also applied both l1 and l2-regularization with regularization weights of $\{1, 0.1, 0.01\}$. The optimal set of parameters are dense layers of size 50 and drop-out probabilities of 50%. The reported CNN models apply 0.01-weighted l2-regularization. The size of each max-pooling window was set to 32 and GloVe word-embeddings are of size 100. For BERT, the GRU are of size 50 and the attention layer size is 100. The data is split into 10% test, 10% validation, and 80% training data.[17] We chose the epoch of the best performing model with the validation data. Since larger companies tend to obtain a higher ESG-rating (Doyle, 2018, p. 9), we control path *env* models for size. Size is measured by the logarithm of the market capitalization. The geographical bias is comparably small, since we only consider stocks listed within the United States. To account for any potential industry bias, we also use the Standard Industrial Classification (SIC) to find the according division responsible,[18] which we include in path *env* models as one-hot-encoded input before the sigmoid output layer. Hence, path *env* models include a concatenate layer before the sigmoid layer to control for size and industry.

3.4 Evaluation

We assume that if the classifier is able to capture the meaning of the disclosures, i.e., has a high F_1-score, then the evidence that the MD&A section is associated with the class labels is consistent with the hypotheses that managers are truthfully disclosing information in the MD&A section and that corporate filings have information content.[19] However, if the MD&A section, based on the classifier, is not associated

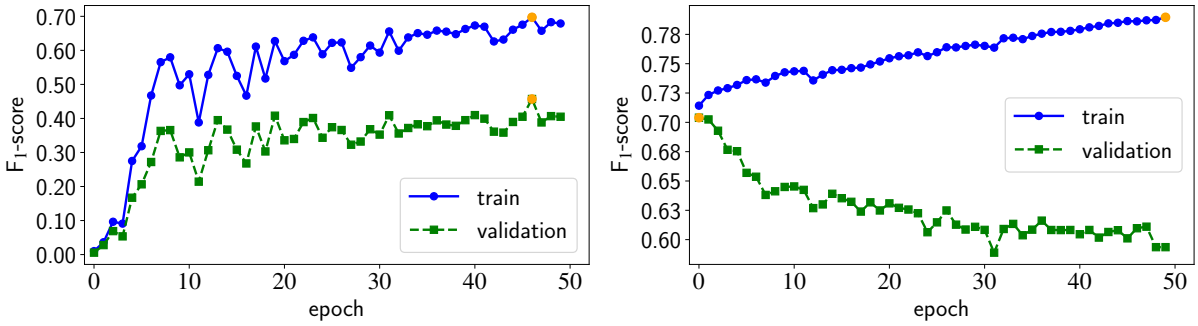
[15] For BERT, we consider a maximum of 1,000 sentences per document, as the median and arithmetic mean of sentences per document is 270 and 332.26, respectively. We then pad each document to the same length before doing the feed-forward pass.

[16] To build our models, we use the Python library Keras (Chollet and others, 2015) using the tensorflow backend (Abadi et al., 2015). To leverage the classifiers, we use the Scikit-learn library (Pedregosa et al., 2011). To implement BERT, we use the Hugging Face transformer library (Wolf et al., 2019).

[17] Our python code is available at: `https://github.com/ForgeFin/Fin-Env-Narrative`

[18] See `https://www.osha.gov/pls/imis/sic_manual.html` (accessed 05.06.2020) for divisions.

[19] This is equivalent to Li (2010a).



(a) F_1-score, $\Delta \text{Env}_{i,q}$

(b) F_1-score, $\Delta \text{EPS}_{i,q}$

Figure 2: Exemplary F_1-score curve for the tfidf-classifier on $\Delta \text{Env}_{i,q}$ and $\Delta \text{EPS}_{i,q}$

with the class labels, then we cannot reject the hypothesis that the MD&A section has no information content because the result could be due to the low power of the classifier. Nevertheless, this might be an indicator that this is eventually the case.

We compare our model's performances to a majority baseline (i.e., always predict class 1, namely "improving"). If a model has a substantially higher F_1-score than the baseline, we are able to predict financial and environmental improvements more precisely compared to always assuming an improvement. In a next step, we compare the F_1-score of the *Direct Path* with the F_1-score of the *Combined Path* using a Bootstrap approach. If the *Combined Path* – considering both environmental performance and text – is significantly better than the *Direct Path*, it indicates that the environmental performance contributes to the relationship between text and financial performance.

4 Empirical Results

Table 1 shows the results of *path env*, the *Direct Path* and the *Combined Path* on the test data as well as the baseline for each classifier. All classifiers do not show a substantial improvement compared to the baseline, meaning always guessing an improvement in environmental or financial performance performs at least as good as a trained classifier. Hence, we plotted the F_1-score for all classifiers to analyse the results further – exemplary shown for **TF-IDF** in Figure 2. While the curve increases for $\Delta \text{Env}_{i,q}$ on the validation data (see Fig. 2a), the model does not show any meaningful increase for $\Delta \text{EPS}_{i,q}$ on the validation data (Fig. 2b). Both graphs show that the classifiers learn from the training data. However, generalisation remains low as the predictive power for the validation data remains low and the F_1-score remains below the baseline. Additionally, we observe that the loss curve (not shown) declines in all instances for the training data but the validation loss decreases only slightly for the first few epochs. Shortly thereafter, loss increases for the validation data indicating an overfit for each model. The results for other classifiers (not shown here) indicate a similar pattern. While there is no classifier for the earnings or the buy-and-hold return label that shows an increasing F_1-score with a decreasing loss function on the validation data, there is a slight improvement for the long and short-term environmental performance. For the environmental classifiers on the validation data, recall mostly improves during training while precision remains relatively constant. This applies for both short-term and long-term environmental performance. Even so, the classifiers remain below the baseline. For the financial performance, the classifiers do not learn from the data. Although, the training F_1-score increases for both financial labels, the results cannot be generalized (similar to Figure 2b). This might be due to low information content, generic language or due to our small sample. Although, the performance of all classifiers remained below expectations, we present the *Combined Path* with the original environmental performance (see Table 1 Task Combined). As could be expected from the previous results, no model proves to be better than the baseline. Further, some models showed that the best model is indeed always guessing class 1.

		$\Delta\text{EPS}_{i,q}$			$\text{BHAR}_{i,t}$			$\Delta\text{Env}_{i,q}$			$\Delta\text{Env}_{i,t}$		
Model	Task	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
Baseline		.53	1	.70	.49	1	.65	.44	1	.61	.37	1	.54
BOW	Single	.56	.89	.69	.51	.75	.61	.46	.64	.54	.40	.56	.46
BOW	Combined	.58	.81	.68	.51	.76	.61						
TF-IDF	Single	.54	1	.70	.49	.63	.55	.47	.45	.46	.50	.22	.31
TF-IDF	Combined	.54	1	.70	.48	.59	.53						
GloVE	Single	.54	.96	.69	.50	.89	.64	.44	.92	.59	.39	.57	.46
GloVE	Combined	.54	.94	.69	.49	1	.65						
CNN	Single	.53	1	.70	.49	1	.65	.46	.51	.48	.39	.16	.22
CNN	Combined	.53	1	.70	.49	1	.65						
BERT	Single	.54	1	.70	.48	1	.65	.44	.52	.48	.39	.10	.16
BERT	Combined	.54	1	.70	.52	.39	.44						

Table 1: Model results with precision, recall, and F_1, respectively. Task *Single* shows *Direct Path* and path *env* for each model. Task *Combined* presents *Combined Path* for each model.

5 Analysis and Discussion

Analysing the reports revealed that the MD&A section varies largely in content and size. While 10-Q reports are rather short and often refer to the 10-K report for more detail, the 10-K reports tend to be rather extensive, with some reports ranging up to more than 100 pages. However, the heterogeneity of 10-K reports make inter-comparisons quite challenging. For our sample, we could not show that the MD&A sections have information content with respect to financial or environmental performance. Given our sample, we cannot reject the null hypothesis for H1, H2, and H3 and, hence, conclude that the MD&A section is not informative with respect to financial and environmental information. However, this can be due to the low performance of the classifiers. As performance of trained classifiers could not be generalised, it is likely that there is not sufficient, material environmental and financial information within the MD&A section, i.e., reports tend to be boilerplate.

Analysing the top 200 weights of the tfidf-classifier, we observe that particularly for the environmental labels the majority of words can be related to the Pharmaceutical or Chemical Industry. The top 200 weights show that for the short-term environmental performance 25 words can be related to the Pharmaceutical or Chemical Industry (12.5%) and 41 words for the long-term environmental performance (20.5%). Filtering for reports which include at least 5 of these words reveals that these reports are related to the Manufacturing division (particularly SIC showing pharmaceutical preparations; surgical&medical instruments; biological products industry etc.). This indicates that the classifier rather learns industries instead of environmental performance, even as we are controlling for industry and market capitalization. However, it is important to note that including at least five of these words in a report does not necessarily result in a performance improvement. While for $\Delta\text{EPS}_{i,q}$ 21 words can be categorized as financial words, the top 200 words appear to be quite random for $\text{BHAR}_{i,t}$.

6 Conclusion

In this paper, we analysed the information content of the MD&A section with respect to financial and environmental narratives employing various NLP methods. We proxy environmental performance with the change in ESG rating percentile scores and the financial performance with earnings and buy-and-hold-returns. Although, we sought to implement a model that uses the predicted environmental performance and text to predict financial performance, results proved to be not as promising. Therefore, we implemented the model with the original environmental performance data instead of the predicted data. We find that our classifiers are not able to surpass a baseline model of always guessing an improvement in financial and environmental performance. As a consequence, we were not able provide evidence that the MD&A

section has information content with respect to financial and environmental narratives, and the section appears to be rather boilerplate. This asks for more structured reports with respect to volume and content as well as the inclusion of more material information. However, it is clearly a limitation of this study to focus solely on one ESG rating provider. As rating construction among providers varies, this raises general data quality concerns (Eccles and Stroehle, 2018) and limits the generalisability of our study. Although other information sources were neglected, corporate disclosures should be by itself useful, informative and reliable. While one could argue that the information is already priced into stock prices shortly after dissemination, the general information content with respect to earnings remains remarkably low. Surprisingly, the usefulness of the MD&A section appears to be quite limited with respect to future corporate performance. An explanation are the reports ambiguity and low polarization. As other researchers have shown, specific word lists are more likely to capture the specific context in this setting. Alternatively, a larger sample could overcome the distinctions of the reports for classification purpose. Future research might use different rating providers, support the classifiers by humanly annotation, or apply BERT models specific for the financial domain, like the recently published FinBert (Yang et al., 2020).

Acknowledgements

This research paper was made possible through the help and support from Christoph Klein at ESG Portfolio Management and our former colleague Sven Raith. Further, we would like to thank Laura Oberländer for her fruitful discussions and suggestions.

References

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. 2015. TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org.

Jay Alammar. 2019. A visual guide to using BERT for the first time. `http://jalammar.github.io/a-visual-guide-to-using-bert-for-the-first-time/`. (accessed: 2020-01-09).

Amir Amel-Zadeh and George Serafeim. 2018. Why and how investors use ESG information: Evidence from a global survey. *Financial Analysts Journal*, 74(3):87–103.

Vasiliki Athanasakou and Khaled Hussainey. 2014. The perceived credibility of forward-looking performance disclosures. *Accounting and business research*, 44(3):227–259.

Brad M. Barber and John D. Lyon. 1997. Detecting long-run abnormal stock returns: The empirical power and specification of test statistics. *Journal of financial economics*, 43(3):341–372.

Henk Berkman, Jonathan Jona, and Naomi S. Soderstrom. 2019. Firm-specific climate risk and market valuation. Available at SSRN 2775552. `https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2775552`.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. O'Reilly Media, Inc.

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

Stephen J. Brown, William Goetzmann, Roger G. Ibbotson, and Stephen A. Ross. 1992. Survivorship bias in performance studies. *The Review of Financial Studies*, 5(4):553–580.

John L. Campbell, Hsinchun Chen, Dan S. Dhaliwal, Hsin-min Lu, and Logan B. Steele. 2014. The information content of mandatory risk factor disclosures in corporate filings. *Review of Accounting Studies*, 19(1):396–455.

Jennifer N. Carpenter and Anthony W. Lynch. 1999. Survivorship bias and attrition effects in measures of performance persistence. *Journal of financial economics*, 54(3):337–374.

Reynolds Center. 2014. 10-K filings guide: Traps and mistakes. https://businessjournalism.org/2014/02/10-k-filings-guide-traps-and-mistakes/. (accessed: 2020-04-09).

CFA Institute. 2017. ESG survey: Global perceptions of environmental, social, and governance issues in investing. https://www.cfainstitute.org/en/research/survey-reports/esg-survey-2017.

François Chollet et al. 2015. Keras. https://keras.io.

Peter M. Clarkson, Yue Li, Gordon D. Richardson, and Florin P. Vasvari. 2008. Revisiting the relation between environmental performance and environmental disclosure: An empirical analysis. *Accounting, organizations and society*, 33(4-5):303–327.

Peter M. Clarkson, Xiaohua Fang, Yue Li, and Gordon Richardson. 2013. The relevance of environmental disclosures: Are such disclosures incrementally informative? *Journal of Accounting and Public Policy*, 32(5):410–431.

Denis Cormier and Michel Magnan. 2007. The revisited contribution of environmental reporting to investors' valuation of a firm's earnings: An international perspective. *Ecological economics*, 62(3-4):613–626.

Chester Curme, H. Eugene Stanley, and Irena Vodenska. 2015. Coupled network approach to predictability of financial market returns and news sentiments. *International Journal of Theoretical and Applied Finance*, 18(07):1–26.

Min-Yuh Day and Chia-Chou Lee. 2016. Deep learning for financial sentiment analysis on finance news providers. In *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 1127–1134. IEEE.

Charl De Villiers and Ana Marques. 2016. Corporate social responsibility, country-level predispositions, and the consequences of choosing a level of disclosure. *Accounting and Business Research*, 46(2):167–195.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Dan Dhaliwal, Oliver Zhen Li, Albert Tsang, and Yong George Yang. 2014. Corporate social responsibility disclosure and the cost of equity capital: The roles of stakeholder orientation and financial transparency. *Journal of Accounting and Public Policy*, 33(4):328–355.

Lodi Dodevska, Viktor Petreski, Kostadin Mishev, Ana Gjorgjevikj, Irena Vodenska, Ljubomir Chitkushev, and Dimitar Trajanov. 2019. Predicting companies stock price direction by using sentiment analysis of news articles. *Computer Science and Education in Computer Science*, pages 37–42.

Timothy M. Doyle. 2018. Ratings that don't rate: The subjective world of ESG ratings agencies. Report, American Council for Capital Formation. https://accfcorpgov.org/wp-content/uploads/2018/07/ACCF_RatingsESGReport.pdf.

Robert G. Eccles and Judith C. Stroehle. 2018. Exploring social origins in the construction of ESG measures. Available at SSRN 3212685. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3318225.

Robert G. Eccles, Michael P. Krzus, Jean Rogers, and George Serafeim. 2012. The need for sector-specific materiality and sustainability reporting standards. *Journal of Applied Corporate Finance*, 24(2):65–71.

Ronen Feldman, Suresh Govindaraj, Joshua Livnat, and Benjamin Segal. 2008. The incremental information content of tone change in management discussion and analysis. Available at SSRN 1126962. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1126962.

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

Gunnar Friede, Timo Busch, and Alexander Bassen. 2015. ESG and financial performance: aggregated evidence from more than 2000 empirical studies. *Journal of Sustainable Finance & Investment*, 5(4):210–233.

Matthew Gentzkow, Bryan Kelly, and Matt Taddy. 2019. Text as data. *Journal of Economic Literature*, 57(3):535–74.

Paul A. Griffin. 2003. Got information? investor response to form 10-K and form 10-Q EDGAR filings. *Review of Accounting Studies*, 8(4):433–460.

Michael Grüning. 2011. Artificial intelligence measurement of disclosure (aimd). *European Accounting Review*, 20(3):485–519.

Zahra Hirji. 2013. Most U.S. companies ignoring SEC rule to disclose climate risks. InsideClimate News. https://insideclimatenews.org/news/20130919/most-us-companies-ignoring-sec-rule-disclose-climate-risks.

Bernd Hüfner. 2007. The SEC's MD&A: does it meet the informational demands of investors? *Schmalenbach Business Review*, 59(1):58–84.

Katrin Hummel and Christian Schlick. 2016. The relationship between sustainability performance and sustainability disclosure–reconciling voluntary disclosure theory and legitimacy theory. *Journal of Accounting and Public Policy*, 35(5):455–476.

Emirhan Ilhan, Philipp Krueger, Zacharias Sautner, and Laura T. Starks. 2019. Institutional investors' views and preferences on climate risk disclosure. Swiss Finance Institute Research Paper. No. 19-66.

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

Juhyun Jung, Kathleen Herbohn, and Peter Clarkson. 2018. Carbon risk, carbon risk awareness and the cost of debt financing. *Journal of Business Ethics*, 150(4):1151–1171.

Colm Kearney and Sha Liu. 2014. Textual sentiment in finance: A survey of methods and models. *International Review of Financial Analysis*, 33:171–185.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. Proceedings of the 3rd International Conference on Learning Representations (ICLR). Contribution to International Conference on Learning Representations, May 7-9, 2015, San Diego.

Julian F. Kölbel, Markus Leippold, Jordy Rillaerts, and Qian Wang. 2020. Does the CDS market reflect regulatory climate risk disclosures? Available at SSRN 3616324. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3616324.

Sabino P. Kothari, Xu Li, and James E. Short. 2009. The effect of disclosures by management, analysts, and business press on cost of capital, return volatility, and analyst forecasts: A study using content analysis. *The Accounting Review*, 84(5):1639–1670.

Todd Kravet and Volkan Muslu. 2013. Textual risk disclosures and investors' risk perceptions. *Review of Accounting Studies*, 18(4):1088–1122.

Benjamin Lansford. 2006. Strategic coordination of good and bad news disclosures: The case of voluntary patent disclosures and negative earnings surprises. Available at SSRN 830705. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=830705.

Allison Herren Lee. 2020. "modernizing" regulation S-K: Ignoring the elephant in the room. https://www.sec.gov/news/public-statement/lee-mda-2020-01-30.

Xiaodong Li, Haoran Xie, Li Chen, Jianping Wang, and Xiaotie Deng. 2014. News impact on stock price return via sentiment analysis. *Knowledge-Based Systems*, 69:14–23.

Feng Li. 2006. Do stock market investors understand the risk sentiment of corporate annual reports? Available at SSRN 898181. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=898181.

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

Feng Li. 2010b. Textual analysis of corporate disclosures: Survey of the literature. *Journal of accounting literature*, 29:143–165.

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

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

Tim Loughran and Bill McDonald. 2019. Textual analysis in finance. Available at SSRN 3470272. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3470272.

Alexandra Luccioni and Hector Palacios. 2019. Using natural language processing to analyze financial climate disclosures. In *Proceedings of the 36th International Conference on Machine Learning, Long Beach, California*.

Ella Mae Matsumura, Rachna Prakash, and Sandra C Vera-Muñoz. 2018. Capital market expectations of risk materiality and the credibility of managers' risk disclosure decisions. Available at SSRN 2983977. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2983977.

Bill McDonald. 2019. Software repository for accounting and finance. https://sraf.nd.edu/. (accessed: 2020-04-09).

Kostadin Mishev, Ana Gjorgjevikj, Irena Vodenska, Ljubomir Chitkushev, Wataru Souma, and Dimitar Trajanov. 2019. Forecasting corporate revenue by using deep-learning methodologies. In *2019 International Conference on Control, Artificial Intelligence, Robotics & Optimization (ICCAIRO)*, pages 115–120. IEEE.

Antonio Moreno-Sandoval, Pablo Alfonso Haya Ana Gisbert, Marta Guerrero, and Helena Montoro. 2019. Tone analysis in spanish financial reporting narratives. In *Proceedings of the Second Financial Narrative Processing Workshop (FNP 2019)*, pages 42–50.

Volkan Muslu, Suresh Radhakrishnan, KR Subramanyam, and Dongkuk Lim. 2008. Causes and consequences of forward looking disclosures in the management discussion and analysis (MD&A). Technical report, Working Paper.

Alan R. Palmiter. 2015. Climate change disclosure: A failed SEC mandate. Available at SSRN 2639181. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2639181.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Marlene Plumlee, Darrell Brown, Rachel M. Hayes, and R. Scott Marshall. 2015. Voluntary environmental disclosure quality and firm value: Further evidence. *Journal of accounting and public policy*, 34(4):336–361.

Carmelo Reverte. 2016. Corporate social responsibility disclosure and market valuation: evidence from spanish listed firms. *Review of Managerial Science*, 10(2):411–435.

Zabihollah Rezaee. 2015. *Business sustainability: Performance, compliance, accountability and integrated reporting*. Greenleaf Publishing.

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

SEC. 1987. Securities act release no. 6711. april 24. Washington, DC. https://www.sec.gov/rules/interp/33-6835.htm.

SEC. 2003. Commission guidance regarding management's discussion and analysis of financial condition and results of operations. https://www.sec.gov/rules/interp/33-8350.htm.

SEC. 2010. Commission guidance regarding disclosure related to climate change. http://www.sec.gov/rules/interp/2010/33-9106.pdf.

Natalia Semenova, Lars Hassel, and Henrik Nilsson. 2010. The value relevance of environmental and social performance: Evidence from swedish six 300 companies. *The Finnish Journal of Business Economics*, 3:265–292, 01.

Philip J. Stone, Dexter C. Dunphy, and Marshall S. Smith. 1966. The general inquirer: A computer approach to content analysis. MIT press.

Lawrence R. Tavcar. 1998. Make the MD&A more readable. *The CPA Journal*, 68(1):10.

Frank HM Verbeeten, Ramin Gamerschlag, and Klaus Möller. 2016. Are CSR disclosures relevant for investors? empirical evidence from germany. *Management Decision*.

Robert E. Verrecchia. 1983. Discretionary disclosure. *Journal of Accounting and Economics*, 5:179 – 194.

Robert E. Verrecchia. 2001. Essays on disclosure. *Journal of Accounting and Economics*, 32(1-3):97–180.

Roshaan Wasim. 2019. Corporate (non) disclosure of climate change information. *Columbia Law Review*, 119(5):1311–1354.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. arXiv: 1910.03771. `https://arxiv.org/abs/1910.03771`.

Fang Yang, Burak Dolar, and Lun Mo. 2018. Textual analysis of corporate annual disclosures: a comparison between bankrupt and non-bankrupt companies. *Journal of Emerging Technologies in Accounting*, 15(1):45–55.

Yi Yang, Mark Christopher Siy UY, and Allen Huang. 2020. Finbert: A pretrained language model for financial communications.

Ye Zhang and Byron C. Wallace. 2017. A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 253–263.

Appendix

A Data preprocessing details

For matching the environmental ratings with the filing date, we apply the following process: Reports are usually filed sometime during the month, e.g., 11.02.2016. However, the ESG data available on Bloomberg is adjusted on a monthly basis at the end of each month, e.g., 28.02.2016. To measure a change in the rating data, we need to match each filing date with the previous months rating. This means a file published on the 11.02.2016 is assigned the ESG percentile score of the 31.01.2016 – provided that a rating exists. Subsequently, the according change in the percentile score is calculated. For reports that were published in February 2014 or March 2014, it is likely that no prior data on the environmental rating is available, hence, the report is not considered. This also applies to some companies that are not rated at all.

B Label calculation

The first financial class label is the change in earnings per share. It serves as a measure of the informativeness of the MD&A section and is defined as $\Delta\mathrm{EPS}_{i,q} = \mathrm{EPS}_{i,q+1} - \mathrm{EPS}_{i,q}$, where the change in earnings per share $\Delta EPS_{i,q}$ is measured as the difference in the earnings reported in the next period $q+1$ and the previously reported earnings for company i.

For the second financial class label, we use buy-and-hold returns and classify the 10-K and 10-Q filings based on the sign of the buy-and-hold returns, $\mathrm{BHAR}_{i,t} = \prod_{t=0}^{t}[1 + R_{it}] - \prod_{t=0}^{t}[1 + R_{mt}]$, where R_{it} and R_{mt} are the returns on stock i and on the SP500 index on date t. Barber and Lyon (1997) state that BHAR are particularly suited for longer periods compared to other measures. BHAR indicates how the information is perceived by investors.

For the first label on the environmental narrative, we use the quarterly change in the environmental percentile score to match the according $\Delta\mathrm{EPS}_{i,q}$ label. The change in the percentile score can be formally expressed as $\Delta\mathrm{Env}_{i,q} = \mathrm{Env}_{i,q+1} - \mathrm{Env}_{i,q}$, with q indicating the according quarter for company i.

The second class label on the environmental narrative uses a one months time horizon to match the $\mathrm{BHAR}_{i,t}$ label. $\Delta\mathrm{Env}_{i,t}$ is the change in the environmental percentile score over the subsequent month, formally expressed as$\Delta\mathrm{Env}_{i,t} = \mathrm{Env}_{i,t+1} - \mathrm{Env}_{i,t}$, with the previously available percentile score being $\mathrm{Env}_{i,t}$ for company i on the filing date and $\mathrm{Env}_{i,t+1}$ being the next environmental score available.