Deep Sustainable Finance:
An End-to-End Text Analysis of the Financial and Environmental Narratives in Corporate Disclosures

Von der Fakultät Wirtschafts- und Sozialwissenschaften der Universität Stuttgart zur Erlangung der Würde eines Doktors der Wirtschafts- und Sozialwissenschaften (Dr. rer. pol.) genehmigte Abhandlung

Vorgelegt von
Felix Armbrust
aus Frankfurt am Main

Hauptberichter:	Prof. Dr. Henry Schäfer
Mitberichter:	PD Dr. Roman Klinger
Tag der mündlichen Prüfung:	06.05.2022

Betriebswirtschaftliches Institut der Universität Stuttgart
2022

Contents

List of Figures

List of Tables

List of Abbreviations

Adam	Adaptive Moment Estimation
AI	Artificial Intelligence
BERT	Bidirectional Encoder Representations from Transformers
BoW	Bag-of-Words
CDP	Carbon Disclosure Project
CDSB	Task Force on Climate-related Financial Disclosures
CNN	Convolutional Neural Networks
CSR	Corporate Social Responsibility
CSRD	Corporate Sustainability Reporting Directive
ELMo	Embeddings from Language Models
ESG	Environmental Social Governance
GAO	Government Accountability Office
GHG	Greenhouse Gas
GloVe	Global Vectors for Word Representation
GRI	Global Reporting Initiative
GRU	Gated Recurrent Units
IIRC	International Integrated Reporting Council
LIME	Local Interpretable Model-agnostic Explanations
MD&A	Mangement's Discussion and Analysis of Financial Conditions and Results of Operations
NFRD	Non-Financial Reporting Directive
NLP	Natural Language Processing
QQD	Quantitative and Qualitative Disclosures About Market Risk

ReLU	Rectified Linear Unit
RNN	Recurrent Neural Networks
SASB	Sustainability Accounting Standards Board
SEC	Securities and Exchange Commission
SIC	Standard Industrial Classification
SRAF	Notre Dame Software Repository for Accounting and Finance
TCFD	Task Force on Climate-related Financial Disclosures
tf-idf	Term Frequency-Inverse Document Frequency

Abstract

In the United States, companies are obliged to report on their financial position in quarterly (10-Q) and yearly (10-K) filings. These filings include the company's financial statements as well as narrative explanations to help the reader assess the financial position of the company. In 2010, the Securities and Exchange Commission (SEC) officially mandated that companies should also consider the consequences of climate change in preparing these filings. Consequently, the filings should not only include material financial information but also material climate-related discussions. However, the filings contain an unaudited section on the company's liquidity and capital resources, the results of operations, and discussions on prospective matters and known material trends. Due to the quasi-discretionary nature of this so-called "Management's Discussion and Analysis of Financial Conditions and Results of Operations" (MD&A) section, the section has been criticised for a substantial amount of boilerplate disclosures, generic language, and immaterial information. In this thesis, we address this criticism and analyse the information content of the MD&A section with respect to the future financial and environmental performance as well as the interaction of environmental performance and financial performance on a narrative basis.

Our research objective is threefold: First, we analyse whether the corporate disclosures contain meaningful information with respect to the future corporate financial performance and the future stock price performance of a company. Second, we examine whether the corporate disclosures contain meaningful environmental information to predict the company's future environmental performance. Third, we analyse whether the narrative environmental information contributes to the narrative financial information in predicting the company's financial performance.

For this reason, we introduce a holistic model that employs end-to-end neural networks based on modern Natural Language Processing (NLP) methods. These modern NLP methods represent the textual disclosures as a set of numerical feature vectors that can also take contextual information into account. In this way, we analyse the actual narrative, instead of just focussing on the quantity of information published. In total, we employ five-different textual feature extraction methods and model architectures to analyse the information content: A standard Bag-of-Words approach, a tf-idf weighting scheme, GloVe embeddings, GloVe embeddings with Convolutional Neural Networks, and BERT sentence embeddings. We first start by outlining the SEC's current regulatory framework for the MD&A section. This includes discussions on the discretionary nature of the MD&A section as well as the required climate-disclosures. We also discuss the current research on financial and environmental disclosures, including how environmental performance is measured. Then, we provide a comprehensive introduction to machine learning and deep learning. Furthermore, we guide the reader through the different applied NLP methods before introducing our holistic model. We use Local Interpretable Model-agnostic Explanations (LIME) to make the predictions of our employed neural networks explainable.

In this thesis, we find that the textual information in the MD&A section does not allow for conclusions about the future corporate financial performance. Consequently, the MD&A section appears to include mostly immaterial information on the future corporate financial performance. In addition, we find that the informational merit for investors is also limited. While investors seem to focus on the numerical information, they either do not read the MD&A section or they do not find meaningful information within the MD&A section. However, the generalisation issue of our classifiers points towards the heterogeneity of the underlying MD&A sections. We find 10-K and 10-Q reports vary greatly in size and content. Furthermore, we find that the MD&A section does not include enough meaningful environmental information. Not all companies do report environmental information in the MD&A section, with the

information published varying substantially among industries as well as among 10-K and 10-Q reports. In addition, we also find no contribution of the environmental information to predicting the financial performance from text. Consequently, our research points towards the necessity of a mandatory, coherent environmental social governance (ESG) framework to elicit more meaningful environmental disclosures. While we do not find evidence for meaningful disclosures, the LIME analysis revealed that our classifiers learn at least some relevant information. Therefore, we propose that future researchers build on our approach, as it aids researchers to discover hidden patterns in text and allows analysing the interaction between financial and environmental information based on a holistic model.

Deutsche Zusammenfassung

In den Vereiningten Staaten von Amerika sind Unternehmen dazu verpflichtet, in Quartalsberichten (10-Q) und Jahresabschlüssen (10-K) über ihre Finanzlage zu berichten. Diese Berichte enthalten neben der Bilanz und der Gewinn- und Verlustrechnung ebenfalls Erläuterungen zu der aktuellen Finanzlage des Unternehmens. Im Jahr 2010 wurden die Berichtspflichten von der Wertpapier- und Börsenaufsichtsbehörde in den USA, der sogennanten Securities and Exchange Commission (SEC), insofern erweitert, dass Unternehmen bei der Erstellung ihrer Geschäftsberichte auch die Folgen des Klimawandels berücksichtigen müssen. Folglich sollten die Berichte nicht nur wesentliche finanzielle Informationen enthalten, sondern auch materielle klimabezogene Diskussionen. Allerdings enthalten die Berichte eine dem deutschen Lagebericht ähnliche Sektion, die nicht durch Wirtschaftsprüfer geprüft wird. In dieser sogenannten "Management's Discussion and Analysis of Financial Conditions and Results of Operations" (MD&A) Sektion sollen Unternehmen über ihre Liquidität, die Kapitalressourcen des Unternehmens, die Betriebsergebnisse sowie zukünftige materielle Entwicklungen oder Pläne berichten. Auf Grund der diskretionären Natur dieses Abschnittes, wird der Informationsgehalt jedoch als unwesentlich und generisch kritisiert. In dieser Arbeit gehen wir daher auf diese Kritik ein und analysieren den Informationsgehalt der MD&A Sektion in Hinblick auf die finanzielle und ökologische Performance des Unternehmens, sowie den Beitrag der ökologischen Informationen, um die finanzielle Leistungsfähigkeit vorherzusagen.

Wir gehen in dieser Arbeit wie folgt vor: Zunächst analysieren wir, ob die MD&A Sektion aussagekräftige Informationen über die zukünftige finanzielle Leistung des Unternehmens enthält – einmal in Bezug auf die Entwicklung der Unternehmensgewinne, als auch in Bezug auf die

Aktienkursperformance des Unternehmens. Anschließend untersuchen wir, ob die MD&A Sektion aussagekräftige Umweltinformationen enthält, um die zukünftige Umweltleistung des Unternehmens vorherzusagen. Drittens untersuchen wir, ob die narrativen Umweltinformationen zu den narrativen Finanzinformationen beitragen, um die finanzielle Performance des Unternehmens vorherzusagen. Für unsere Untersuchung entwickeln wir daher ein ganzheitliches Modell, das durch end-to-end neuronale Netze und mittels moderner Methoden der natürlichen Sprachverarbeitung, dem sogenanntem Natural Language Processing (NLP), die Berichte analysiert. Diese modernen NLP-Methoden stellen die Berichte als eine Menge numerischer Merkmalsvektoren dar, wodurch auch kontextuelle Informationen berücksichtigt werden können. Auf diese Weise untersuchen wir den tatsächlichen, narrativen Informationsgehalt, anstatt uns allein auf die Quantität der Veröffentlichungen zu konzentrieren. Zur Analyse des Informationsgehalts setzen wir fünf verschiedene Merkmalsextraktions- und Modellarchitekturen ein: Einen standard Bag-of-Words-Ansatz, ein Tf-idf-Gewichtungsschema, GloVe-Embeddings, GloVe-Embeddings mit Convolutional Neural Networks und BERT-Satz-Embeddings. Am Anfang der Arbeit stellen wir zunächst den derzeitigen Regulierungsrahmen der SEC in Bezug auf die MD&A Sektion vor. Dies beinhaltet eine Diskussion des Ermessensspielraums, den die MD&A Sektion bietet, sowie der erforderlichen Klimaangaben. Zudem erörtern wir die aktuelle Forschung zu Finanz- und Umweltpublikationen, einschließlich möglicher Arten die Umweltleistung zu messen. Anschließend geben wir eine umfassende Einführung in das maschinelle Lernen und Deep Learning. Dazu führen wir den Leser durch die von uns verwendeten NLP-Methoden. Wir verwenden Local Interpretable Model-agnostic Explanations (LIME), um die Vorhersagen der von uns verwendeten neuronalen Netze erklärbar zu machen.

Unsere Forschung zeigt, dass die MD&A Sektion keine Rückschlüsse über die zukünftige finanzielle Leistungsfähigkeit des Unternehmens erlaubt. Folglich scheint die MD&A Sektion überwiegend unwesentliche Informationen über die zukünftige finanzielle Leistungsfähigkeit des Un-

ternehmens zu enthalten. Darüber hinaus stellen wir fest, dass der Informationswert für die Anleger ebenfalls begrenzt ist. Während die Anleger sich auf die numerischen Informationen zu konzentrieren scheinen, lesen die Anleger die MD&A Sektion entweder nicht oder sie finden keine aussagekräftigen Informationen in diesem Abschnitt. Zudem stellen wir fest, dass der MD&A Abschnitt nicht genügend aussagekräftige Umweltinformationen enthält. Nicht alle Unternehmen berichten Umweltinformationen in der MD&A Sektion und die Veröffentlichungen variiren erheblich zwischen den Branchen als auch zwischen den 10-K- und 10-Q-Berichten. Wir finden zumal keine Hinweise darauf, dass die Umweltinformationen zu der zukünftigen Leistungsfähigkeit der Unternehmen beitragen. Wir schließen daraus, dass die Börsenaufsichtbehörde einen verpflichtenden, einheitlichen Environmental Social Governance (ESG) Rahmen in Betracht ziehen sollte, damit mehr Unternehmen über materielle Umweltinformationen berichten und der Nutzen für die Anleger gesteigert wird. Wir möchten zudem darauf hinweisen, dass das Verallgemeinerungsproblem unserer Klassifikatoren auf die Heterogenität der MD&A Sektion hinweist. 10-K- und 10-Q-Berichte unterscheiden sich stark in Bezug auf Umfang und Inhalt. Obwohl wir keine Hinweise auf aussagekräftige Informationen finden, zeigt unsere LIME-Analyse, dass die Klassifikatoren zumindest einige relevante Informationen erlernen. Wir schlagen daher vor, dass zukünftige Forscher unseren Ansatz mit verbesserten Filtermethoden fortführen.

1. Introduction

Corporate disclosures are a means to inform the company's stakeholders, i.e., shareholders, creditors, employees, and other company-related groups (Freeman, 2010), about the financial position of a company. In this regard, it is important that corporate disclosures contain decision-useful information, as particularly investors use corporate disclosures as a tool to understand and evaluate the company's performance (IMP, 2020). However, the narrative of corporate reports has been criticised for being largely generic and immaterial (Wasim, 2019; Palmiter, 2015; Bloomfield, 2008). Furthermore, while reports should be published in a timely manner and contain comparable, verifiable, reliable, and consistent information (Leuz and Wysocki, 2008, p. 24), especially the non-financial information[1] of corporate reports, e.g., information on human capital, impacts of changes in weather patterns on the business, or other market impacts related to climate change (SEC, 2010), often lacks comparability and usefulness (Esty et al., 2020, p. 2). For this reason, this thesis addresses the narrative information content of corporate disclosures with modern Natural Language Processing (NLP) methods to analyse in a transparent, comparable, and replicable fashion whether corporate disclosures are associated with the company's financial and environmental performance. Moreover, we analyse the contribution of the environmental information to the financial performance on a narrative basis, as research indicates that non-financial performance is material for financial performance (Friede et al., 2015). This chapter provides an introduction to our study by first discussing the background and context (Section 1.1 and Section 1.2), followed by the research problem (Section 1.3), the research objectives (Section 1.4), and

[1] The terms the terms business sustainability, corporate social responsibility (CSR) or environmental, social, governance (ESG) are often used interchangeably in the financial literature (see IMP, 2020, p. 4, Rezaee, 2015, p. 2, i.a.)

finally the general outline of this thesis (Section 1.5).

1.1. A Lack of Decision-Useful Corporate Disclosures

Over the last decades, the average number of pages in annual reports increased enormously. This includes particularly the qualitative part being published, e.g., the number of footnotes or the pages in the Management's Discussion and Analysis section (Ernst & Young, 2012). While the narrative information should aid readers to understand the financial statements and the strategic goals of the company, investors face the risk of information overload (Stolowy and Paugam, 2018) and the need to process textual data faster increases – either to keep or to gain a competitive advantage (F. Z. Xing et al., 2018). Additionally, the social and political pressure companies and investors face, led to a general rise in non-financial reporting and investing. While disclosures on topics of environment, social, and governance (ESG) used to be mainly considered voluntary and published apart from legal oversight, nowadays non-financial disclosures are often required and even legally mandated (Hackett et al., 2020).

On the regulatory side, policy makers have started taking a stronger stance on sustainability-related guidance and rules. Over the past years, an increasing number of sustainability-related regulations have been enforced (Datamaran, 2020). According to the Corporate Register database (Corporate Register, 2020), more than 20,000 organizations worldwide now report on sustainability issues, either voluntarily or on a mandatory basis. In the United States, United Kingdom, and Canada regulations tightened, with the most affected topics being business ethics and climate change in the financial services industry, energy use and consumer rights in the utilities industry, and product and service safety in healthcare and pharmaceuticals (Datamaran, 2020). However, in the European Union, changes in sustainability regulations are most eminent.

In the European Union, the EU Directive 2014/95/EU, also called the Non-Financial Reporting Directive (NFRD), obliged companies to

report on ESG information for the first time in 2018 for the fiscal year 2017, affecting over 6,000 companies with an average of 500 or more employees (Kinderman, 2015). The NFRD introduced a double materiality approach, which required companies to report on: (1) How the companies' performance, position, and development are affected by sustainability issues, the so-called *outside-in perspective*, and (2) how the companies' decisions impact the stakeholders and environment, the so-called *inside-out perspective* (European Commission, 2021b, p. 1). In April 2021, the newly adopted Corporate Sustainability Reporting Directive (CSRD) amended the NFRD and set out a consistent and comparable taxonomy for reporting on sustainability information. The CSRD extends the scope to all large companies listed on regulated markets as well as requires the reported information to be audited. The CSRD introduces more detailed reporting requirements and, additionally, has the goal to make the information machine readable (European Commission, 2021b). Furthermore, the CSRD requires companies to assess the double materiality on a continuous basis (Datamaran, 2021, p. 3). While this work does not focus on the European double materiality and does not cover the European regulatory framework in detail, we do focus on material information, i.e., information that is decision-useful and relevant for users of financial statements (Eccles et al., 2012), and the current U.S.-regulatory framework.

In the United States, public companies are generally required to disclose material, financial and non-financial information within their corporate filings (SEC, 2010). However, the Securities and Exchange Commission (SEC) has remained largely silent with respect to the differences in ESG-related disclosure (Esty et al., 2020, p. 2). This is due to the fact, that the SEC applies a principle-based approach, which makes companies autonomously responsible to identify relevant risks and factors affecting the company's performance. Therefore, managers have some discretion in the information they disclose (F. Li, 2010b, p. 1053), which might cause managers to withhold unfavourable information below a critical disclosure level or to provide imprecise information (Verrecchia, 1983, 2001). Additionally, companies have an incentive to frame negative

events in rather positive terms (Loughran and McDonald, 2016, p. 1218, Loughran and McDonald, 2019, p. 6–7). This equally applies to financial as well as non-financial information (Hummel and Schlick, 2016).

While in the European Union the EU Taxonomy Climate Delegated Act, adopted in April 2021, specifies precisely what activities contribute to climate change mitigation and adoption (European Commission, 2021a), aiding companies as well as investors to make sustainable investment decisions, and the CSRD outlines detailed disclosure requirements to make sustainability information reliable and comparable, there is no such clear framework in the United States.[2] Until now, the SEC does not provide a common ESG disclosure framework and companies need to refer to voluntary non-financial disclosure frameworks, such as those provided by the Global Reporting Initiative (GRI), the Sustainability Accounting Standard Board (SASB), the Task Force on Climate-related Financial Disclosures (TCFD), and the Climate Disclosure Standards Board (CDSB), among others.

However, recent alignment efforts by standard-setting institutions reflect the increasing demand for decision-useful, climate-related information (IMP, 2020; TCFD, 2019), and even the SEC's own Investor Advisory Committee urged the SEC to instruct companies on exactly what information companies should disclose and how (Bennington, 2020). In order to fill the existing informational gaps and meet investors' needs, several ESG rating agencies have been established (Bennington, 2020; Davies et al., 2020). Prominent examples of rating providers include MSCI, Sustainalytics, Thomson Reuters, and Vigeo EIRIS (Kumar and Weiner, 2019). These ESG rating agencies evaluate the companies ESG performance and construct their ratings by surveying companies, analysing corporate releases, and collecting other unstructured data (Esty et al., 2020).

However, the landscape of ESG rating providers is crowded and confusing (Davies et al., 2020, p. 175). The lack in standardization, comparability, credibility, and missing transparency makes the ratings difficult to use

[2]The EU Taxonomy Climate Delegated Act sets out precise definitions what qualifies as environmentally sustainable (European Commission, 2020).

for investors and companies alike (Windolph, 2011). While investors do consider ESG metrics important, investors also demand more informative narratives that provide context – in particular on the managements future plans and how the company will deal with externalities (WBCSD, 2018, p. 7). As Davies et al., 2020, p. 168 notes: "much ESG information that is most useful to investors requires forward-looking projections". Under the current U.S.-regulatory framework, a section that theoretically fulfils these requirements is the "Management's Discussion and Analysis of Financial Condition and Results of Operations" (MD&A) section. The MD&A contains an analysis of material current developments and future trends from the managements' point of view (Muslu et al., 2015; Griffin, 2003). However, due to the managers discretion and a lack of a common ESG disclosure framework, it is questionable whether the financial as well as non-financial information of the MD&A section is actually informative and useful for investors (Wasim, 2019; Palmiter, 2015; F. Li, 2010b). Hence, it is necessary to evaluate the current SEC disclosure guidelines in terms of their information content and usefulness to investors.

Still, objective comparisons of corporate reports are quite difficult in reality as disclosures are often qualitative and narrative in nature (Leuz and Wysocki, 2008, p. 24). A way to address this issue is Natural Language Processing (NLP). NLP allows the user to represent textual disclosures as a set of numerical feature vectors that can be used for statistical analysis. Only recently, more advanced NLP methods make it possible to account for complex grammatical structures, "disambiguate different senses of words, [and] distinguish key points from secondary asides", Gentzkow et al., 2019, p. 537. Once a document classification mechanism is established, the decisions of a document classifier are based on a fixed set of rules. Furthermore, NLP enables (1) the fast processing of a large amount of data and (2) the discovery of hidden, subliminal patterns in texts, e.g., the unintentional use of weak modal words in corporate disclosures that might signal a troubled firm (Loughran and McDonald, 2016, p. 1191). Thus, NLP allows to efficiently and transparently analyse the narrative of corporate disclosures with respect to the financial and non-financial

information content.

1.2. Natural Language Processing and the Financial Domain

Natural Language Processing (NLP) is concerned with the computational processing of human communication and encompasses approaches to understand, interpret, and analyse textual data efficiently and objectively (Kamath et al., 2019, p. 11). NLP is considered a subfield of computer science and artificial intelligence, and can be used for tasks like text summarization, translation, correction, and sentiment analysis, among others (Fisher et al., 2016; Loughran and McDonald, 2016). With NLP methods, we can convert text into meaningful representations for computational use (Gentzkow et al., 2019, p. 537) and generate objectively reproducible results. This means that once the rules and filters on quantifying the texts are established, no additional subjectivity on the part of the researcher affects the measurement (Loughran and McDonald, 2016, p. 1209) and public availability of dictionaries (Loughran and McDonald, 2016, p. 1200) and underlying program code allows for easy, transparent replication.

In the financial domain, the application of statistical-based NLP methods is an emerging field of research (Loughran and McDonald, 2016, p. 1188). While in the financial domain lexical resource-based methods prevail, mainly due to ease of use and efficiency (Frunza, 2020; Kearney and Liu, 2014) – i.e., counting the occurrence of particular words according to a pre-designed word list – more sophisticated methods like machine learning and deep learning algorithms are rarely used (Gentzkow et al., 2019). However, NLP has made enormous progress in the past: (1) Word embeddings allow to map words into fixed length vectors preserving the syntactic and grammatical relationship between words (Chaturvedi et al., 2016; Mikolov, Sutskever et al., 2013; Bengio et al., 2003); (2) language models like ELMo, i.e., Embeddings from Language Models, (M. E. Peters

et al., 2018) and BERT, i.e., Bidirectional Encoder Representations from Transformers, (Devlin et al., 2019) can capture not only the syntax and semantics but also account for polysemy[3] (M. E. Peters et al., 2018); and, in all that, (3) NLP methods allow to analyse large amounts of textual data fast, automatically, and efficiently (F. Z. Xing et al., 2018, p. 49-50). Therefore, we can use NLP methods to classify documents according to their key features and learn about hidden structures and subtle information in texts.

1.3. Motivation and Research Gap

The current regulatory framework in the United States obliges companies to report on their financial position in quarterly (10-Q) and annual (10-K) filings. These filings are accompanied by a narrative explanation to help the reader understand and assess the company's financial position. A section that is arguably one of the most read sections of these filings is the MD&A section (Clarkson et al., 1999; Tavcar, 1998; Rogers and Grant, 1997). The MD&A section provides analysts and investors with a long- and short-term analysis from the management's point of view (Muslu et al., 2015; Griffin, 2003) and also contains a forward-looking perspective (SEC, 2003). However, the section is – compared to other parts of the filings – unaudited (Cohen et al., 2008; Hüfner, 2007). Concerned that the MD&A section includes largely boilerplate discussions and immaterial detail, the SEC issued several interpretative releases to improve the disclosure quality of the MD&A section (SEC, 1987, 1989, 2003). In 2010, the SEC followed with an interpretive release on climate change "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. Consequently, company's disclosures should not only contain qualitative financial information but also material environmental information. However, whether the MD&A section provides

[3]Polysemy describes words with more than one meaning (Cambridge Dictionary, 2020).

actually useful information (F. Li, 2010b, p. 1050) or does mostly contain boilerplate discussions (Wasim, 2019; Palmiter, 2015; F. Li, 2010a; Pava and Epstein, 1993) remains an empirical question.[4]

Previous research based on the textual information content provides evidence that corporate disclosures are indeed predictive of the future company performance (Lawrence, 2013; Kothari et al., 2009) and are associated with the company's performance (Campbell et al., 2014; Merkley, 2014; Kravet and Muslu, 2013, i.a.). Furthermore, evidence suggests that corporate disclosures are associated with company's risk (Kravet and Muslu, 2013; Kothari et al., 2009; F. Li, 2006), future earnings (Moreno-Sandoval et al., 2019; Athanasakou and Hussainey, 2014; F. Li, 2010b), and, ultimately, firm value (Campbell et al., 2014; Jegadeesh and Wu, 2013; Feldman et al., 2010). However, previous work relies mostly on straightforward word-counts, word-phrase counts, or similar approaches to conclude whether disclosures are actually informative. Typically, these studies first derive some type of disclosure score based on a key word count and subsequently regress a financial measure on the score to draw inferences from the text.

Equivalently, most research addressing the non-financial information content apply lexical-based approaches (Kölbel et al., 2020, p. 8). While researchers found that voluntary, standalone non-financial disclosures are positively associated with market value and negatively with cost of equity (Verbeeten et al., 2016; De Villiers and Marques, 2016; Reverte, 2016; Plumlee et al., 2015; D. Dhaliwal et al., 2014; Clarkson et al., 2013), previous research finds that only a small percentage of companies explicitly mention climate-change in their 10-K and 10-Q filings (Government Accountability Office, 2020; Palmiter, 2015; Hirji, 2013). However, evidence suggests that mentions of climate change in 10-K filings are associated with the firm's cost of capital (Berkman et al., 2019; Matsumura et al., 2018). But even so, only few studies that deal with non-financial information address the actual *narrative* information content and focus

[4] Compare Armbrust et al., 2020.

instead on the *quantity* of non-financial information published (Hummel and Schlick, 2016). Only recently, Kölbel et al., 2020 have applied a more sophisticated approach by training a sentence-wise BERT algorithm supported by humanly annotation on TCFD sample reports to provide evidence that climate-related factors within the risk section of 10-K filings affect the spreads of credit default swaps.

While simple key word counts or word phrase counts are a reasonable first step to analyse the content of corporate disclosures, the technique's downside is that the word order is not taken into account (F. Z. Xing et al., 2018, p. 52). Since the context is ignored, pure word counts cannot capture the differences in meaning which arise from word order: For instance, the statement "AstraZeneca acquired Alexion Pharmaceuticals" is very different from the statement that "Alexion Pharmaceuticals acquired AstraZeneca". Furthermore, negations are not considered contextually, i.e., negations are not used as a function that changes the meaning of a word, and semantic similarities will not be captured by a simple word lists (F. Z. Xing et al., 2018, p. 52). Consequently, the MD&A section of 10-K and 10-Q filings is largely unexamined using state-of-the-art NLP methods. We see this methodological gap and, hence, apply more sophisticated NLP methods to examine the information content of corporate disclosures. To our knowledge, the MD&A section of 10-K and 10-Q filings remains also largely unexamined with respect to the company's environmental performance. Additionally, the effect of non-financial narratives on the relationship between financial narratives and financial performance remains largely unexamined. While more than 2000 studies provide evidence that non-financial performance is material for financial performance (Friede et al., 2015), the influence of non-financial performance on financial performance on a narrative basis remains largely unexamined. Since investors expect to either outperform markets over time by considering ESG factors or simply by avoiding companies, which are not well positioned with respect to climate-change (Bennington, 2020; Esty and Cort, 2017), we also examine the contribution of environmental information on the financial performance.

1.4. Research Objective

The broader aim of this thesis is to examine whether corporate reports contain meaningful information regarding the company's future financial performance and environmental performance, i.e., whether the corporate disclosures are associated with the future financial performance and environmental performance of a company. We, thus, formulate three connected research questions (RQ):

RQ-1: **Do corporate disclosures contain meaningful information with respect to (1) the future corporate financial performance and (2) the future stock price performance?** Since textual disclosures have been criticized for being boilerplate, we examine whether corporate disclosures actually provide meaningful information – meaningful with respect to the companies own financial performance, measured in terms of earnings, but also meaningful to the investors' performance, i.e., the company's stock return.

RQ-2: **Do corporate disclosures contain meaningful environmental information to predict the future environmental performance of the company?** In the United States, the current regulatory framework obliges companies to report on material environmental information. Thus, we expect that meaningful environmental information should provide insights on the company's future environmental performance.

RQ-3: **Does the narrative environmental information contribute to the narrative financial information in predicting the company's financial performance?** This third research question is based on the investors' expectation to outperform markets by using ESG information and the increasing demand for meaningful ESG information. Since environmental information can be an important complement of financial information, we construct a model that integrates the environmental performance into the financial

performance on a narrative basis. We aim to improve the performance of a statistical text classifier by integrating the environmental performance.

The fourth (and more general) objective is to introduce more sophisticated NLP methods to the financial domain. In the financial domain, researchers mostly still rely on straightforward word counts. We expect that methods, which mimic a human-level understanding of text, provide new insights into corporate disclosures. In directly training machine learning classifiers on the financial performance and the environmental performance, we avoid deriving a disclosure score first. Based on previous research, we expect that companies do not explicitly mention factors related to climate change and that companies frame negative events rather positively. However, we aim to extract hidden textual indicators by training machine learning classifiers on the disclosures.

Thus, our contributions are as follows:

1. We evaluate whether corporate filings, i.e., the MD&A section of 10-K and 10-Q filings, contain information that can be used to predict:
   - the future corporate financial performance, and
   - the future stock price performance.
2. We evaluate whether corporate filings, i.e., the MD&A section of 10-K and 10-Q filings, contain information that can be used to predict the environmental performance.
3. We analyse whether the narrative environmental information complements the narrative financial information in explaining the future financial performance, i.e., we analyse whether the prediction of the financial performance can be improved by the environmental information.
4. We apply state-of-the-art deep-learning models on financial disclosures and introduce them to a wider audience in the financial

domain.

The expected utility of this study is to improve our understanding of the interactions between financial performance, environmental performance, and reporting. We aim to show whether the corporate disclosures contain meaningful information with respect to the financial performance, environmental performance, and whether the environmental information contributes to the financial prediction task from text. While surveys indicate that institutional investors value climate-relevant information (Ilhan et al., 2019; Amel-Zadeh and Serafeim, 2018), a survey by the CFA Institute also showed that more than half of the participants make the lack of quantitative ESG information responsible for *not* using non-financial information in their investment process (CFA Institute, 2017). Thus, this thesis is also a first attempt to efficiently quantify textual environmental information. We expect that our classifiers are able to support the decision-making process of inventors. Additionally, our approach may help to react faster to textual information through trading automation.

While we focus on the American market, this thesis can also be seen in relation to the current developments in the European Union and the discussion on climate disclosures. However, we mainly aim to provide insights on the current disclosure guidelines of the SEC with respect to financial and environmental information. In March 2021, the SEC announced the creation of a Climate and ESG Task Force in the Division of Enforcement, which should develop initiatives to pro-actively identify ESG-related misconduct (SEC, 2021). The Task Force should focus on material gaps or misstatements of ESG information in disclosures. Consequently, our work can also be seen in the context of this Task Force.

The reader should note that this work is aimed equally at researchers from the financial domain *and* researchers from the NLP domain. Researchers from the financial domain learn more about the disclosures content and value, and are introduced to modern NLP methods that we use. Likewise, researchers from the NLP domain learn more about the current issues in the financial domain and the challenges associated with

financial disclosures. Thus, the cross-disciplinary approach of this thesis aims at closing the gap and hurdles between computational linguists and financial researchers.

1.5. Outline of the Thesis

This thesis is structured as follows: In Chapter 2, we introduce the regulatory background in the United States and name the reason for focusing on the MD&A section. We discuss related research addressing the financial and non-financial information content of the MD&A section, generate our hypotheses, and introduce the theoretical background. Then, we discuss how environmental performance is measured and how to avoid common pitfalls. In Chapter 3, we introduce the underlying machine learning techniques and relevant NLP methods. In doing so, we explain how machine learning and deep learning algorithms "learn", how to measure the performance of a machine learning and deep learning algorithm, how to extract meaningful features from texts, and introduce common NLP network elements. In Chapter 4, we introduce our underlying research model and the associated prediction tasks. We provide an overview of our corpus and how we preprocess the financial documents. We explain how we measure the financial and non-financial performance, and how we evaluate our models. In Chapter 5, we present our results with respect to the financial and environmental performance as well as to the contribution of environmental information to the financial performance. Chapter 6 contains the discussion of our empirical results and, finally, in Chapter 7, we draw our conclusion and present a future outlook.

2. Information Content and Research Framework

In this chapter, we introduce the Securities and Exchange Commission's (SEC) reporting framework (Section 2.1). We explain the reasoning for focussing on the MD&A section, briefly introduce the concept of materiality as the overarching principle of the MD&A section (Subsection 2.1.1), and take a closer look at the discretionary nature of the MD&A disclosure setting (Subsection 2.1.2). Then, we discuss the SEC's 2010 interpretative release (Subsection 2.1.3) (SEC, 2010), the voluntary non-financial disclosure frameworks that companies refer to in lack of a detailed guidance (Subsection 2.1.4), and compare the frameworks with each other (Subsection 2.1.5). Next, we discuss related research on the MD&A section (Section 2.2) with respect to the financial (Subsection 2.2.1) and the non-financial disclosures (Subsection 2.2.2). In Section 2.3, we introduce the common theories on voluntary non-financial disclosures due to the quasi-voluntary nature of the MD&A section and how the voluntary disclosures are expected to affect the value of a company. Finally, we discuss the use of ESG ratings to measure environmental performance and the common problems associated with the use of ESG ratings (Section 2.4).

2.1. The Reporting Framework in the U.S.

In the United States, most public companies are obliged to file quarterly and yearly reports under Regulation S-K with the Securities and Exchange Commission, called 10-Q Form and 10-K Form, respectively (SEC, 2011). These filings include the audited financial statements of the company, which present a numerical picture of the company's financial position, as well as narrative explanations to help users of financial statements

to assess the company's financial position (SEC, 1989). Among others, the narrative explanations must elaborate on the company's operations (Item 101, "Description of the Business"), the company's legal proceedings (Item 103, "Legal Proceedings"), the risk factors of the company (Item 503 (c), "Risk Factors"), and the company's financial condition from the management's point of view (Item 303, "Management's Discussion and Analysis of Financial Conditions and Results of Operations").

Arguably one of the most read parts of the filings include the Management's Discussion and Analysis of Financial Conditions and Results of Operations (MD&A) section (Tavcar, 1998; Rogers and Grant, 1997). The MD&A section requires the management to discuss the company's liquidity and capital resources, the results of operations, as well as any other information necessary to understand the financial condition of the company (SEC, 1989). The discussions should assist investors to understand the reasons for the change in the financial condition as well as for the changes in the results of operations. Furthermore, the section should contain a long- and short-term analysis of the company discussing prospective matters and known material trends (SEC, 2003).

The requirements of the MD&A section are comprehensive, yet the SEC intentionally decided to adopt a flexible and more general disclosure approach to foster meaningful disclosures and avoid boilerplate discussions (SEC, 1989). Thus, compared to other sections of the 10-K and 10-Q filings, the MD&A section is unaudited (Cohen et al., 2008; Hüfner, 2007). However, the SEC underscores the concept of materiality as the overarching principle of the MD&A section (SEC, 2020). Companies are expected to provide *material* financial and non-financial information within the MD&A section. Therefore, we discuss briefly the concept of materiality in more detail.

2.1.1. The Concept of Materiality

Materiality applies to financial as well as non-financial information and is a concept used by companies, auditors, and other users of financial

information. The Financial Accounting Standard Board, responsible for the United States Generally Accepted Accounting Principles (US-GAAP), defines material information as any kind of information that by omitting or misstating it could influence the decisions that users make on the basis of the financial information of a specific reporting entity (FASB, 2010, p. 17). Similarly, according to the Securities and Exchange Commission a matter is material, if there is a substantial likelihood that a reasonable person would consider it important (SEC, 1999). Both interpretations of materiality are not subject to a certain numerical threshold as to when a matter is considered material. For investors reading the MD&A section, a matter can thus be interpreted as material if the information impacts investors' decision to buy, sell, or vote a security (Davies et al., 2020).[1]

2.1.2. A Discretionary Disclosure Setting

Although disclosures in the MD&A section are mandatory and the SEC Rule 10b-5 prohibits untrue statements or the omission of material facts, the discretion of the management could lead to managers withholding inconvenient information below a critical disclosure level (Verrecchia, 2001, 1983).[2] Furthermore, managers have an incentive to frame negative events in positive terms (Loughran and McDonald, 2016, p. 1218), which also increases the likelihood that discussions become immaterial or boilerplate. However, the risk of potential litigation might serve as a disciplining tool to incentivise companies to provide material information (F. Li, 2010b; Kothari et al., 2009). Albeit, "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

[1] For an extensive overview of different materiality concepts, we refer the interested reader to Eccles et al., 2012.

[2] According to the proprietary cost theory, also called discretion disclosure theory, firms disclose information only if the costs associated with the disclosure are lower than the benefits resulting from the disclosure (Dye, 2001; Verrecchia, 2001). The discretion disclosure theory reflects company concerns that disclosures might negatively affect the firm's competitive position due to potential regulatory action, legal liabilities, reputation damage, or reduced customer demand (Dye, 1985; Verrecchia, 1983).

short-term", Kothari et al., 2009, p. 1643. Thus, the MD&A section's content can be viewed as *quasi-voluntary* or *quasi-mandatory* (Muslu et al., 2015, p. 932).

Concerned that the MD&A disclosures include largely boilerplate discussions and immaterial detail, the SEC issued several interpretative releases to improve the disclosure quality of the MD&A section (SEC, 1987, 1989, 2003). At the time of writing, the latest amendments to improve and modernize the MD&A section took effect on February 10, 2021 (SEC, 2020). The newly adopted amendments provide clarifications on the disclosure requirements, eliminate duplicative disclosures, and streamline the disclosure items (SEC, 2020). However, the amendments do not establish a common ESG disclosure guidance nor do the amendments add any *new* ESG disclosure requirements to the MD&A section (see Section 2.1.3).

2.1.3. The SEC's ESG Disclosure Requirements

In 2010, the SEC issued an interpretative release 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. Becoming effective on February 8, 2010, the guideline mandates companies to include material ESG-related information within their "Description of the Business" section (Item 101), the "Risk Factors" section (Item 503 (c)), the "Legal Proceedings" section (Item 103), as well as within the "Management's Discussion and Analysis of Financial Conditions and Results" section (Item 303). Hence, based on the concept of materiality, the MD&A section should also include any climate-related information that has affected or is expected to affect the financial position of the company. However, the SEC has provided only limited guidance on the particular ESG information that companies should disclose (Davies et al., 2020, p. 162). Thus, the guideline has been criticised as lax and ineffective (Wasim, 2019; Palmiter, 2015).

In 2016, the SEC issued a Concept Release to invite public comments

on the modernization of the Regulation S-K with respect to certain business and financial disclosure requirements (SEC, 2016). Although the Concept Release covered a wide range of topics, the majority of comments addressed sustainability issues (SASB, 2016). Investors argue that companies are not providing enough meaningful disclosures with respect to their ESG performance (Davies et al., 2020, p. 166). Following the comment letter review process, the SEC proposed new updates to the disclosure requirements in 2019. However, none of the proposals addressed the calls for enhanced ESG disclosure requirements (SEC, 2019).

In 2020, even the SEC's own Investor Advisory Committee urged the SEC to instruct companies more precisely on what type of information companies should disclose and in what format, recommending a common ESG disclosure standard (Bennington, 2020). Thus, in March 2021, the SEC announced the creation of a Climate and ESG Task Force in the Division of Enforcement, which should develop initiatives to proactively identify ESG-related misconduct (SEC, 2021). The Task Force should focus on material gaps or misstatements of ESG information in disclosures. However, the SEC's 2021 newly adopted amendments regarding Regulation S-K do not contain any improvements to the ESG disclosure requirements (SEC, 2020). Consequently, passages addressing climate-change are hard to identify (Luccioni and Palacios, 2019) and the criticism regarding the flexibility and ineffectiveness remains (Wasim, 2019; Palmiter, 2015).

2.1.4. Voluntary Non-Financial Disclosure Frameworks

In lack of a common and conclusive ESG disclosure standard, several international voluntary disclosure standards have emerged. To name here are the six leading voluntary framework and standard-setters: The Carbon Disclosure Project (CDP), the Climate Disclosure Standards Board (CDSB), the Global Reporting Initiative (GRI), the International Integrated Reporting Council (IIRC), the Sustainability Accounting Standard Board (SASB), and the Task Force on Climate-Related Financial Disclos-

ures (TCFD).[3] While the SEC's 2010 interpretative release states that much of the reporting according to these standard setters is voluntary, the SEC notes that companies should be aware that some of the information companies report according to these disclosure standards may also be required to be disclosed in the SEC filings (SEC, 2010, p. 6292). Therefore, we briefly introduce each standard setter and, subsequently, compare them with the SEC requirements (see Subsection 2.1.5).

2.1.4.1. Carbon Disclosure Project (CDP)

The Carbon Disclosure Project (CDP) is an international non-profit organization that provides stakeholders with a platform for environmental information on companies, cities, states, and regions. Founded in 2000, the CDP collects emission data from companies, ranks companies according to their environmental risks and opportunities, and makes the data transparent to different kinds of stakeholders (Depoers et al., 2016, p. 445-446). The CDP's questionnaires guide companies on how to measure and disclose climate-related information.

2.1.4.2. The Climate Disclosure Standards Board (CDSB)

The Climate Disclosure Standards Board (CDSB) is a private organization consisting of an international consortium of businesses and environmental, non-governmental organizations. Founded at the World Economic Forum in 2007, the CDSB provides detailed guidance on how to report on environmental and natural capital information (CDSB, 2019). The CDSB exclusively focuses on the environmental aspect of ESG pursuing standardized disclosures of environmental information. The CDSB addresses primarily investors as the users of environmental information to provide clear, concise, and comparable information that connects the company's environmental performance with the overall performance of the company (CDSB, 2019, p. 5-6).

[3] Compare World Economic Forum, 2020, p. 6.

2.1.4.3. Global Reporting Initiative (GRI)

The Global Reporting Initiative (GRI) is one of the oldest reporting initiatives on sustainability, founded in 1997. Hence, the GRI is also one of the world's most widely used sustainability reporting frameworks (Hoffmann, 2011, p. 68). The GRI framework guides companies on how to include material, non-financial factors within their financial reports. At its core, the GRI reporting framework consists of three universal standards that need to be applied by all companies in compliance with the GRI, i.e., GRI 101: Foundation, GRI 102: General Disclosures, and GRI 103: Management Approach. In addition, companies must apply several topic specific standards relevant to the company (GRI, 2016). The current version of the GRI framework focuses particularly on the concept of materiality and covers all dimensions of ESG (Federation of European Accountants, 2016, p. 7). The GRI addresses all stakeholder groups (GRI, 2016, p. 7).

2.1.4.4. International Integrated Reporting Council (IIRC)

Established in 2010, the main objective of the International Integrated Reporting Council (IIRC) is to foster an integrated reporting approach on sustainability issues. While some companies publish sustainability reports apart from their financial reports, the IIRC promotes the inclusion of non-financial information within financial reports. Hence, the IIRC follows a more holistic approach to address the creation of value and the shortcomings of financial reporting (Dumay et al., 2016). Companies in adherence to the IIRC should report on how their governance, organizational strategy, and future prospects lead to the creation of value (IIRC, 2013). While the IIRC provides companies with guidance on the principles and content of integrated reporting, it does not provide standards for ESG disclosures (Government Accountability Office, 2020, p. 6).

2.1.4.5. Sustainability Accounting Standard Board (SASB)

In 2011, the Sustainability Accounting Standard Board (SASB) was established to provide companies with a guideline on environmental and social issues. The SASB standards are developed with the intend to be consistent with the current financial regulation of the United States (Federation of European Accountants, 2016, p. 8). The objective is to integrate the SASB standards into the filings required by the SEC for all public traded companies (Federation of European Accountants, 2016, p. 8). In the United States, already one in four members of the S&P 500 makes reference to the SASB (The Economist, 2020). At its core, the SASB applies the concept of materiality and categorizes companies into ten sectors and 77 industries providing industry-specific reporting standards (SASB, 2017).[4] The SASB addresses all ESG dimensions and focuses on investors and creditors as its main audience.

2.1.4.6. Task Force on Climate-Related Financial Disclosures (TCFD)

The Task Force on Climate-Related Financial Disclosures (TCFD) was established in 2015 by the Financial Stability Board (FSB) to promote more effective climate-related disclosures. The TCFD applies a stakeholder point of view and provides recommendations on managing climate-related risks and opportunities. According to the TCFD, climate-related information should be consistent, comparable, clear, and reliable to allow companies and investors to manage climate-related risks and opportunities effectively (TCFD, 2020, 2019).

2.1.5. Convergence of Voluntary Disclosure Frameworks and Comparison with the SEC

The voluntary framework and standard-setters differ in their approach to non-financial information. While the TCFD, CDP, and CDSB are

[4]The SASB Conceptual Framework is currently, at the time of writing, under revision (see https://www.sasb.org/standard-setting-process/conceptual-framework/).

mainly concerned with the environmental dimension, the IIRC, SASB, and GRI address all aspects of ESG. Furthermore, both the SASB and GRI emphasize the concept of materiality – similar to the SEC requirements. However, the GRI addresses all stakeholders, while the SASB is mainly concerned with investors and creditors. These differences lead to a lack in consistency and confusion among companies, which standard to chose and apply (Davies et al., 2020, p. 171).

In September 2020, the CDP, CDSB, GRI, IIRC, and SASB announced in a common statement their intent to work together (IMP, 2020). The standard setters aim to achieve a more comprehensive corporate reporting in aligning their frameworks. In addition, the GRI and SASB have announced to demonstrate how their respective standards can be used together (IMP, 2020, p. 14). The SASB and CDSB have already articulated their complementary nature, in a publication by the TCFD (TCFD, 2019). Furthermore, the TCFD offers guidance on how to enhance the consistency and comparability of company's non-financial reporting by demonstrating overlap among standard-setters (TCFD, 2019). On November, 25 2020, the IIRC and the SASB announced their indent to merge to form the Value Reporting Foundation (SASB, 2020). The IIRC and SASB consider their reporting guidelines complementary and, hence, advocate a common sustainability disclosure standard. While the IIRC provides relevant value creation topics and the approach to integrate them into corporate reporting, the SASB supplements these with precise definitions of data to be reported for each industry.

While companies are not obliged to adhere to any of the above mentioned standards, companies can adopt the voluntary disclosure standards to improve their reporting and still be in compliance with the SEC disclosure requirements (Davies et al., 2020, p. 171). However, compared to the voluntary disclosure standards, which provide industry-specific disclosure guidance and instruct companies on how to report on non-financial subjects, the SEC's 2010 guidance basically emphasizes that material climate-related circumstances are already covered under the disclosure requirements *prior* to the 2010 release (Eccles et al., 2012,

p. 68). Thus, the SEC basically leaves it up to the management to decide what information they consider material with respect to climate-change. Consequently, the management's discretion makes it an empirical question whether meaningful climate-related information is actually disclosed.

2.2. Research Hypotheses

In this section, we focus on studies that analyse the information content of the 10-K and 10-Q filings. First, we present related research on the financial information content focusing mainly on studies addressing the MD&A section and generate our first hypothesis. Then, we present related research on the non-financial information content and generate our second and third hypothesis. For the related research on the non-financial information content, we focus on the entire 10-K and 10-Q filings, since research addressing solely the non-financial information content of the MD&A section is thin.

2.2.1. Decision-useful, Financial Information within the MD&A section

Due to the management's discretion and ongoing arguments that the MD&A section contains mostly generic language, the MD&A section has been subject to research addressing the financial information content. Research focussing on the narrative information content has, thus, mainly relied on word-count based measures or other types of disclosure scores.[5] To examine the usefulness and information content of the MD&A section, researchers associate their word-count based measures or other types of disclosure scores either with (1) future or contemporaneous stock returns, (2) future earnings, (3) analysts' forecast behaviour, or (4) stock price volatility, among others.[6]

Prior to the 2003 interpretative release (SEC, 2003), Pava and Epstein,

[5]This also includes approaches, which use tf-idf weighting (see chapter3.5.2.2).
[6]Compare Brown and Tucker, 2011.

1993 provide evidence that the MD&A section is boilerplate with respect to the future financial performance by using a qualitative approach. While Pava and Epstein, 1993 find that historical events are correctly reported, discussions of *negative* trends are either ignored or not fully disclosed, leading to positive biased MD&A disclosures. In addition, by employing readability measures, i.e., measuring the ease to read the MD&A section (see also Chapter 3.7.2), Schroeder and Gibson, 1990 provide evidence that the MD&A section is generally difficult to read and that the comprehensibility could be improved.[7] For the time period between 1996 and 2001, Kothari et al., 2009 take a more encompassing approach by employing a variety of information sources, i.e., 10-K and 10-Q filings as well as analyst reports and news articles in the business press. Kothari et al., 2009 find that pessimistic analyst reports and business-press articles increase the company's cost of capital, while using solely the corporate disclosures, i.e., filings that include the MD&A section, there is no such significant relationship. Thus, the findings indicate that the filings contain no meaningful information content. However, Cole and Jones, 2004 find that the quantifiable information in the MD&A section, i.e., sales growth, store openings or closings, and capital expenditures, are useful in predicting the company's future financial performance, measured in terms of earnings and contemporaneous stock returns.

Counting the occurrences of risk-related words in 10-K filings, i.e., including MD&A disclosures, F. Li, 2006 shows that an increase in risk sentiment is associated with lower future earnings. Similarly, Feldman et al., 2008 show that the stock return around the 10-K and 10-Q filings is significantly associated with the tone of the MD&A section, i.e., whether more positive or negative words occur according to the General Inquirer (Kelly and Stone, 1975). Furthermore, research by Muslu et al., 2015 provides evidence that the MD&A section improves the information environment indicating that MD&A disclosures are not boilerplate. Muslu et al., 2015 find that firms make more forward-looking MD&A

[7] Similarly, focusing on 10-K filings, F. Li, 2008 find that in particular firms with a poor financial performance tend to have reports with a low readability.

disclosures, when their stock prices poorly reflect the company's future earnings information. However, the disclosures seem to improve but are unable to completely mitigate the poor informational efficiency of stock prices for such firms (Muslu et al., 2015).

Other researchers have also supported their findings by employing straightforward automated text classification systems. For example, F. Li, 2010b uses a Naïve Bayes classifier to categorize the tone and content of forward looking statements within the MD&A section from 1994 to 2007. A Naïve Bayes classifier assumes that the occurrence of each word within a document is independent of other words within the document, i.e, a Naïve Bayes classifier ignores the context of a word (Jurafsky and Martin, 2020, p. 56-60). F. Li, 2010b humanly classifies sentences of forward looking statements as positive, negative, or neutral and subsequently trains the classifier on the filings to predict sentences to be negative (-1), neutral (0), or positive (1). By averaging the predictions among all sentences, F. Li, 2010b obtains an average tone of each filing. F. Li, 2010b subsequently regresses the tone on the company's earnings to find that the average tone of a filing is positively correlated with the future performance. Furthermore, F. Li, 2010b finds that no such relation exists for word-count based measures indicating that machine learning approaches might be superior to a straightforward word count. Similarly, Jegadeesh and Wu, 2013 employ a Naïve Bayes classifier to derive an overall tone measure of each filling from the stock price reaction. Jegadeesh and Wu, 2013 find a significant relation between a report's tone and the stock price reaction.

A different approach is provided by Brown and Tucker, 2011. Brown and Tucker, 2011 examine the usefulness of MD&A disclosures by measuring how much a company's filing has been modified compared to the company's previous published filing. Brown and Tucker, 2011 assume that the MD&A section must change substantially after significant economic changes, otherwise the MD&A section cannot be considered informative. Thus, Brown and Tucker, 2011 construct a score based on a filings' similarity to a previous published filing to find that firms with large economic changes, measured by a variety of financial variables, modify their MD&A section

more than those with smaller economic changes. Additionally, they find that a greater modification of the filings is positively associated with the filings' stock price response, however, analyst earnings forecast revisions are not. This suggests investors use the MD&A information but analysts do not.

We conclude that most research, especially after the SEC, 2003 release, find MD&A disclosures to be informative with respect to the future financial performance. However, research tends to derive a word-based disclosure score or tone measures first, before drawing inferences about the usefulness of corporate disclosures. Furthermore, to our knowledge, little research in the financial domain has analysed the textual disclosures with respect to more sophisticated text analysis methods, i.e., taking the word order into account. Thus, we like to fill this gap in training sophisticated machine learning classifiers directly on the MD&A section to analyse the financial information content, while also considering the textual context. Since the MD&A section has been shown to contain information content with respect to the company's earnings and stock price performance, we formally state our first hypotheses as follows:

H1a. The MD&A section of the 10-K and 10-Q filings contains information associated with the future financial performance of a company, i.e., the performance measured by earnings.

H1b. The MD&A section of the 10-K and 10-Q filings contains information associated with the future stock price performance of a company.

2.2.2. Decision-useful, Non-Financial Information within the 10-K/Q filings

Investors argue that companies do not provide enough material climate-related information (Davies et al., 2020, p. 166). Thus, researchers have analysed whether the SEC's 2010 guidance has improved companies' climate-related disclosures and whether companies provide more material climate-related information after the guidance went into effect. However,

researchers analysing the climate-related information have mainly used disclosure-scores based on hand-coded content analysis or survey rankings (Cong et al., 2020; Coburn and Cook, 2014; DiSalvio and Dorata, 2014; Eccles et al., 2012; Freedman and Park, 2014). Only few studies have applied key word counts or similar methods (Kölbel et al., 2020; Baier et al., 2020; Berkman et al., 2019; Matsumura et al., 2018; Doran and Quinn, 2009).

Prior to the SEC's 2010 guidance, evidence from Doran and Quinn, 2009 shows that only a small percentage of companies include discussions on climate-related risks and opportunities in their 10-K filings. Using a key word search, Doran and Quinn, 2009 provide evidence that during the years 1995 to 2008 the majority of firms do not even mention the phrase "*climate change*". However, in the year *preceding* the 2010 release, Feng and Gao, 2020 find that the number of words regarding environmental disclosures increased significantly, suggesting that companies may have anticipated the SEC's release. Furthermore, using a word list from the content analysis software DICTION, which measures the degree of "realism" and "certainty", Feng and Gao, 2020 find an increased degree of "realism" and "certainty" in 10-K disclosures following the 2010 release indicating more pronounced environmental disclosures.

Employing a climate change disclosure index according to Freedman and Park, 2014 for the two year period from 2010 to 2011, a study by Cong et al., 2020 finds that firms emitting more greenhouse gases (GHG) also provide more disclosures on climate change in their 10-filings. However, based on the 10-K filings of six different industries in 2011, Eccles et al., 2012, p. 68 find that the majority of companies either include boilerplate statements or no climate-related information at all. In addition, Eccles et al., 2012, p. 68 note that ESG disclosure practices vary widely among companies. Similarly, focusing on 10-K disclosures of S&P 500 companies between 2010 and 2013, Coburn and Cook, 2014 find a wide variation in environmental disclosures by companies within the same industry and a decline in average quality from 2010. Furthermore, Coburn and Cook, 2014 also find a large number of companies not disclosing climate-related

information in their 10-K reports. Similarly, using a disclosure score index, DiSalvio and Dorata, 2014 find that firms disclose more information after the 2010 release, with firms facing greater climate change exposure having a significantly larger increases in disclosure. However, DiSalvio and Dorata, 2014 note a huge variance in content and amount that companies disclose. Furthermore, DiSalvio and Dorata, 2014 show that the market reaction to the release was generally positive and mainly driven by firms more exposed to climate-risk.

A study conducted by the United States Government Accountability Office (GAO) analysed 10-K reports, proxy statements, annual reports, and voluntary sustainability reports of 32 companies published in the year 2018 (Government Accountability Office, 2020). The study showed that most companies include at least some ESG-related information but the information was not always clear or useful. For instance, the study found that some climate or resource-related information is difficult to compare as companies used different calculation methods or reported results in different units of measurement.

Other studies that analysed the climate-related information content of 10-K filings with respect to simple word count approaches include Cannon et al., 2020, Berkman et al., 2019, and Matsumura et al., 2018. Analysing the length of climate-related 10-K disclosures as well as the language used for the years 2011 to 2015, Berkman et al., 2019 find evidence that the cost of capital rises with higher exposure to climate risk. Berkman et al., 2019 also provide evidence that the tone of climate-risk disclosures within 10-K filings is significantly associated with firm value. Similarly, screening the 10-K form for climate change-related words, Matsumura et al., 2018 show that firms not disclosing material climate-related risk have a higher cost of equity. Cannon et al., 2020 use a key word search for corporate social responsibility (CSR) terms in 10-K filings from 1996 to 2015 to find that firms using more CSR-related terms have superior gross and operating margins. Recently, Kölbel et al., 2020 have applied a more sophisticated approach by employing a sentence-wise BERT algorithm supported by humanly annotation, which was trained on TCFD sample

reports. Kölbel et al., 2020 find that the climate-risk reported in 10-K reports affects the spreads of credit default swaps.

Prior studies indicate that the level of disclosure significantly increased after the SEC's 2010 release (Feng and Gao, 2020; Coburn and Cook, 2014; DiSalvio and Dorata, 2014). However, disclosure practices still vary widely among companies (Government Accountability Office, 2020; Coburn and Cook, 2014; DiSalvio and Dorata, 2014). Nevertheless, studies based on word-count measures provide evidence that climate-related 10-K disclosures can be used as an indicator for companies' risk and, ultimately, value (Kölbel et al., 2020; Cannon et al., 2020; Berkman et al., 2019; Matsumura et al., 2018). Therefore, we assume that the 10-K and 10-Q reports contain information about the companies' environmental performance.[8] Furthermore, we assume that the environmental information is value relevant and, thus, relevant for the companies earnings and stock price performance.

We contribute to the literature in addressing the usefulness of the SEC's 2010 guideline by focusing particularly on the MD&A sections' climate-related information. Furthermore, we contribute to the literature in directly training more sophisticated text analysis classifiers on the disclosures. Thus, we focus directly on the narrative information content instead of using a simple word-count or disclosure indices based on hand-coded content analyses or survey methods. Since the MD&A section should include forward-looking discussions about climate-related opportunities and risks (SEC, 2010), we assume that the MD&A section contains information about the future environmental performance of the company. We formally state our second hypothesis as follows:

H2. The MD&A section of the 10-K and 10-Q filings contains information associated with the future environmental performance of a company.

In addition, we expect that the environmental performance derived from the narrative information content contributes to the company's financial

[8] Note that evidence by Matsumura et al., 2014 shows that even non-disclosure of climate-related information affects the value of companies.

performance. If the environmental information is actually material, as the SEC's 2010 guideline suggests as well as studies based on word measures support (Kölbel et al., 2020; Cannon et al., 2020; Berkman et al., 2019; Matsumura et al., 2018), the environmental information should contribute to the companies financial performance. Our reasoning posits that the MD&A section contains environmental performance information and, thus, we can use this information to analyse whether this information is actually material. Hence, we first quantify the narrative information content and, subsequently, measure its usefulness to explain the financial performance. Consequently, we state our third hypothesis as follows:

H3. The narrative environmental performance contained within the MD&A section contributes to a company's financial performance.

2.3. Voluntary Disclosures on Environmental Performance

Only recently, governments have started turning voluntary environmental disclosure regimes into mandatory ones (Hackett et al., 2020). Thus, a large body of research addresses voluntary environmental disclosures (De Villiers and Marques, 2016; Plumlee et al., 2015; D. S. Dhaliwal et al., 2011; Plumlee et al., 2015; Clarkson et al., 2013; Clarkson et al., 2008, i.a.). While our study focuses on mandatory climate-related disclosures, the MD&A section is subject to management's discretion (see Subsection 2.1.2) and, thus, can be considered quasi-voluntary (Muslu et al., 2015). Therefore, we like to present the prevailing theories on voluntary environmental disclosures (Subsection 2.3.1) and provide a brief overview of the related research (Subsection 2.3.2).

2.3.1. Theoretical Perspective on Voluntary Environmental Disclosures

Mainly two theories have been used to explain the relationship between environmental performance and voluntary environmental disclosures: (1) Voluntary disclosure theory, and (2) legitimacy theory. While the voluntary

disclosure theory suggests that companies with superior environmental performance disclose environmental information to gain a competitive advantage (Clarkson et al., 2008), legitimacy theory suggests that companies use environmental disclosures to justify their poor environmental performance (Deegan, 2002; Patten, 2002; O'donovan, 2002). Thus, both theories indicate a competing association of environmental performance and voluntary environmental disclosures.

Voluntary disclosure theory assumes that companies with a good environmental performance disclose more on environmental information, since good environmental performance is difficult to mimic by poor performing competitors (Clarkson et al., 2008). Good environmental performers expect a competitive advantage and, subsequently, an increased market value from voluntary environmental disclosures. Therefore, voluntary disclosure theory indicates a positive relationship between environmental performance and voluntary environmental disclosures (Clarkson et al., 2008).

On the other hand, legitimacy theory, which overlaps with other sociopolitical theories (Hahn et al., 2015; Clarkson et al., 2008), suggests that companies disclose environmental information to improve the company's public perception (Deegan, 2002; Patten, 2002; O'donovan, 2002). Companies see their public legitimation at risk and, thus, use a legitimation tactic to avoid public pressure regarding the company's poor environmental performance (Hummel and Schlick, 2016). Therefore, legitimacy theory predicts low-quality and off-setting disclosures, which result in a negative relationship between environmental disclosures and the company's financial performance (Deegan, 2002).

Previous research on voluntary disclosure theory and legitimacy theory has provided mixed results: There is evidence of a positive relationship between environmental performance and voluntary environmental disclosures, i.e., good environmental performers disclosing more environmental information, (see e.g., Clarkson et al., 2008, Al-Tuwaijri et al., 2004) as well as evidence of a negative relationship (see e.g., Clarkson et al., 2011, C. H. Cho and Patten, 2007, De Villiers and Van Staden, 2006). How-

ever, Hummel and Schlick, 2016 argue that both theories are inherently connected.

Hummel and Schlick, 2016 suggest that the mixed findings result from the way the disclosure quality is measured. While according to voluntary disclosure theory good environmental performers are likely providing high-quality disclosures, i.e., easy comparable numerical data, legitimacy theory posits that poor environmental performers are likely providing low-quality disclosures, i.e., disclosures of narrative nature. Therefore, Hummel and Schlick, 2016 distinguish between high-quality and low-quality disclosures using a disclosure index and find that firms attempt to increase their market value through voluntary disclosure, while, at the same time, firms try to avoid the negative consequences of a threatened legitimacy. *However*, our study does not make any assumptions about a negative or positive relationship between environmental performance and environmental disclosures. We simply assume that a relationship exists, i.e., the narrative provides information on the environmental performance.

2.3.2. Environmental Disclosures affecting Financial Performance

In the literature, the decreasing and increasing effects of voluntary environmental disclosures on the company's value have been analysed in detail. Studies provide evidence that voluntary environmental disclosures are value-relevant offering incremental explanatory power (Verbeeten et al., 2016; Reverte, 2016; Clarkson et al., 2013) with environmental preparedness mainly increasing market value (Semenova and Hassel, 2008). Furthermore, researchers have shown that voluntary non-financial disclosures lower the costs of equity capital (D. Dhaliwal et al., 2014) and particularly high level disclosures are associated with a higher stock price (De Villiers and Marques, 2016).

However, most studies focus on the *quantity* of non-financial information published instead of the actual *narrative* information content (Hummel and Schlick, 2016).[9] For example, Reverte, 2016 uses a rating score from

[9]Compare also Verbeeten et al., 2016, p. 1363-1364.

the Observatory on Corporate Social Responsibility, which is "based on an exhaustive content analysis of sustainability reports", Reverte, 2016, p. 414. Similarly, Plumlee et al., 2015 use an environmental disclosure index to measure the level of environmental disclosure. Moreover, Clarkson et al., 2013 use an intra-industry percentile ranking to measure the degree of voluntary environmental disclosure. An exception is Verbeeten et al., 2016, who actually count the occurrence of certain environmental terms to measure the level of environmental disclosure. However, little research addresses the actual narrative information content of environmental disclosures but the quantity of environmental disclosures and the numerical metrics provided. Thus, we also hope to provide new insights to the general literature on environmental disclosures by employing a machine learning-based textual analysis.

2.4. Measuring Environmental Performance – ESG Ratings and Rating Providers

Until now, we have explained the reasons why we expect that the MD&A section contains information regarding the company's environmental performance as well as the theoretical framework on environmental disclosures. In this section, we like to discuss the use of ESG ratings to measure environmental performance and common problems associated with the use of ESG rating providers.

The increasing demand for decision-useful ESG information has led to a rise in ESG rating providers (Bennington, 2020). ESG rating agencies provide detailed information on the company's environmental, social, and governance dimensions, which help investors to account for the company's ESG performance. However, the landscape of ESG rating providers is crowded and confusing – not only for investors but also for companies (Davies et al., 2020, p. 175). Challenges are the lack in standardization, credibility, and missing transparency of the ratings (Windolph, 2011). ESG rating agencies usually collect ESG data by sending long

questionnaires to companies and subsequently adjust the data according to company releases and other secondary information sources. However, the data collected and the metrics generated from the data vary widely in quality leading to missing comparability, integrity, and utility (Esty et al., 2020, p. 2). Research shows that rating agencies consider different indicators in their rating construction and also weight these indicators differently (Escrig-Olmedo et al., 2019). In addition, the origins of rating agencies influences the underlying sustainability concepts (Eccles and Stroehle, 2018) and, hence, assigned ESG ratings disagree among rating agencies (Chatterji et al., 2016). Furthermore, the credibility is undermined by a lack in transparency, as the rating construction is hardly accessible (Windolph, 2011, p. 45).

In a recent working paper, Berg et al., 2020 point out that the disagreement in ratings results mainly from the three different sources: (1) scope divergence, which refers to a different set of attributes that rating providers use; (2) measurement divergence, i.e., different measurements of identical categories; and (3) weight divergence, i.e., the fact that rating agencies assign different weights to categories when aggregating the rating. Consequently, ESG ratings are unlikely to guide companies towards an improvement in their ESG performance, as the disagreement among rating providers diminishes any signalling effects the ratings could have (Berg et al., 2020, p. 32).

In another recently published article, Gibson et al., 2021 analyse ESG ratings from six prominent ESG rating providers for S&P 500 firms between 2013 and 2017. Gibson et al., 2021 find that the ESG rating correlation among rating providers is surprisingly low, although the highest correlation appears among the environmental dimension. Although one would expect a higher agreement in ratings if companies disclose more information, Christensen et al., 2019 find that a higher level of ESG disclosure leads to more disagreement with respect to ESG ratings. According to Berg et al., 2020, a researcher has therefore three options of dealing with the ESG rating disagreement: (1) Include several ESG ratings from different rating providers; (2) use one ESG rating provider to measure

a specific company's characteristics; or (3) construct hypotheses around attributes that are more narrowly defined than ESG performance, i.e., reliance on more verifiable and transparent measures like GHG emissions.

According to Berg et al., 2020 the first option "is reasonable when the intention is to measure 'consensus ESG performance' as it is perceived by financial markets in which several ratings are used", Berg et al., 2020, p. 31. However, this option is costly and it is unclear *how* to aggregate different ratings into a single number, e.g., with an arithmetic mean or other methods. Thus, if a rating provider considers different aspects of the environmental dimension or assigns different weights to certain topics, aggregation is likely to become arbitrary. The second option has the downside that using different rating providers can lead to different results (Berg et al., 2020). Therefore, researchers and investors have to be careful in the selection of an appropriate rating provider. While the third option seems natural, it does not measure the whole spectrum of ESG performance. For instance, the environmental performance of a company does not only consist of the company's carbon emissions but also of how the company deals with other environmental issues, e.g., water pollution, waste management, energy consumption, and so on. Hence, the sole use of GHG emissions would only consider one aspect of environmental performance.

Consequently, different measures for environmental performance are applied in the literature. While some researchers use single emission data or toxic waste (Clarkson et al., 2011; Al-Tuwaijri et al., 2004, i.a.), other researchers have used rating providers such as Kinder, Lydenberg, Domini (today MSCI) (Krüger, 2015; Hong and Kostovetsky, 2012; C. H. Cho et al., 2006, i.a.), or Thomson Reuters Asset4 ESG database (Dyck et al., 2019, i.a.). Hence, there does not seem to be a uniform solution in measuring environmental performance and researchers need to carefully evaluate how they measure environmental performance.

2.5. Summary

In this chapter, we first introduced the current reporting framework in the United States (Section 2.1). We introduced the MD&A section as one of the most read sections by analysts (Tavcar, 1998; Rogers and Grant, 1997), in which companies need to report on material events that affected the company's operating results as well as report on future trends and prospective matters that are material to the company's financial condition. We explained the concept of materiality on which the MD&A section is based (Subsection 2.1.1) and the underlying discretion, which makes it an empirical question whether managers actually include financial and non-financial information in the MD&A section (Subsection 2.1.2). We also addressed the criticism that the current framework does not elicit sufficient material ESG disclosures (Subsection 2.1.3) and compared the framework to related voluntary non-financial disclosure frameworks (Subsection 2.1.4 and 2.1.5). In Section 2.2, we found that most research provides evidence on the MD&A section to contain information regarding the financial and non-financial performance. However, research tends to use disclosure scores or simple word-count measures to analyse the information content. Thus, we developed the hypothesis that the MD&A section is associated with the future financial performance (Subsection 2.2.1) and with the future environmental performance (Subsection 2.2.2). Additionally, we expect that the narrative environmental performance is material and, therefore, helps to predict the company's future financial performance. Next, we explained the common theories on voluntary environmental disclosures as the MD&A section can be considered quasi-voluntary (Section 2.3). Finally, we discussed how environmental performance is measured and the pitfalls researchers need to avoid (Section 2.4). In the next chapter, we explain the theoretical background on machine learning to understand the text analysis methods that we apply. We explain how to evaluate the performance of a machine learning algorithm as well as current developments in Natural Language Processing.

3. Background on Machine Learning and Deep Learning

In this chapter, we introduce the background on machine learning and deep learning that is required for automatically analysing and classifying textual data. Since machine learning and more advanced Natural Language Processing (NLP) methods are not (yet) standard practice in the financial literature, the following three sections introduce the three different types of machine learning (Section 3.1), the basic structure behind neural networks (Section 3.2), and explain a simple probabilistic classifier (Section 3.3). In Section 3.4, we explain how a machine learning algorithm "learns", how to avoid overfitting of a classifier, and how to evaluate machine learning classifiers. Then, we explain how to apply this knowledge on textual data (Section 3.5) and also introduce common network elements to build efficient classifiers (Section 3.6). In the last section, we briefly talk about textual analysis in the financial domain (Section 3.7).

3.1. Overview of Machine Learning

Machine learning and deep learning are a sub-field of artificial intelligence (AI) that involves algorithms, which do not rely on hard-coded knowledge or programmed rules to produce output but instead acquire their knowledge by extracting patterns from raw data (Goodfellow et al., 2017, p. 2).

Machine learning as well as deep learning are broadly subdivided into three types of learning: (1) supervised learning, (2) unsupervised learning, and (3) reinforcement learning.

Supervised learning refers to algorithms, which are tasked to classify data according to a pre-known label, i.e., the desired output is already

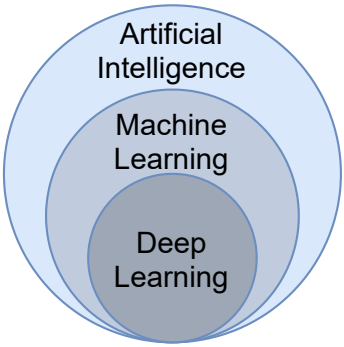


Fig. 3.1.: The Field of Artificial Intelligence.

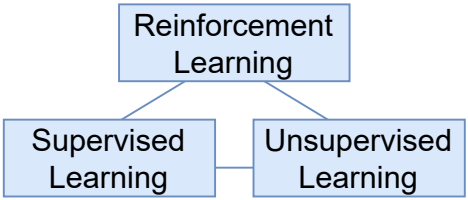


Fig. 3.2.: Supervised Learning, Unsupervised Learning, and Reinforcement Learning. Own representation drawn after Raschka, 2015, p. 2

known. In the case of Natural Language Processing (NLP), this could mean that we construct an algorithm that learns whether a given text has a positive or negative connotation. The labels – positive and negative – are in this scenario already known.

Unsupervised learning refers to a classification task of unlabelled data, i.e., the desired output is not known. An example in an NLP setting would be an algorithm that is tasked to find meaningful text characteristics to cluster the texts into two distinct categories, e.g., according to a certain genre or topic without the actual genre or topic being known (e.g., Latent Dirichlet Allocation (Blei et al., 2003) and Latent Semantic Indexing (Deerwester et al., 1990; Dumais et al., 1988)).

Reinforcement learning refers to algorithms that learn from their

environment without human guidance (Goodfellow et al., 2017, p. 25). In this case, the algorithm does not require labelled data but learns from the actions with its environment in a feedback loop. "The learner is not told which actions to take, as in many forms of machine learning, but instead must discover which actions yield the most reward by trying them out", Sutton and Barto, 1998, p. 2.[1]

3.2. Neural Networks

Machine learning and deep learning are loosely inspired by the functioning of a human brain (Goodfellow et al., 2017, p. 13). In the human brain, neurons transmit information based on the received input. If the input signal is strong enough, i.e., surpasses a certain threshold, the synapse transmits a chemical or an electrical signal to another nerve cell. Equivalently, neural networks – in the sense of machine learning – consist of interconnected nodes that process and transmit signals based on an activation function. Thus, a neural network is basically a chain of interconnected functions. If a network has just two interconnected functions $f^{(1)}$ and $f^{(2)}$, that form the chain $f(x) = f^{(2)}(f^{(1)}(x))$, we call it a **machine learning** task. If the network has more than two interconnected functions, for instance, the functions $f^{(1)}$, $f^{(2)}$, and $f^{(3)}$ that form $f(x) = f^{(3)}(f^{(2)}(f^{(1)}(x)))$, we say the network is deep, that is, **deep learning** (Goodfellow et al., 2017, p. 163). However, there is no agreement on how "deep" a neural network needs to be, to be called a deep neural network. Hence, we can think of deep learning as a part of machine learning that involves a more nested and flexible type of machine learning (Goodfellow et al., 2017, p. 8).

[1] For more information on reinforcement learning see also Mnih et al., 2013; Bertsekas and Tsitsiklis, 1996.

3.2.1. Layers

The first applied function of a neural network, $f^{(1)}$, is called the **input layer**. Likewise, the last function of a neural network is called the **output layer**. In the case of a three (or more) layer network, the intermediate layers are called **hidden layer**. Figure 3.3 depicts an exemplary neural network with an input layer, two hidden layers, and one output layer. In Fig. 3.3, the arrows show the direction of the information flow. If the input passes once through the network – from left to right (see Fig. 3.3) – we call it a **feedforward pass**. The network's prediction is denoted as $\hat{y}$.

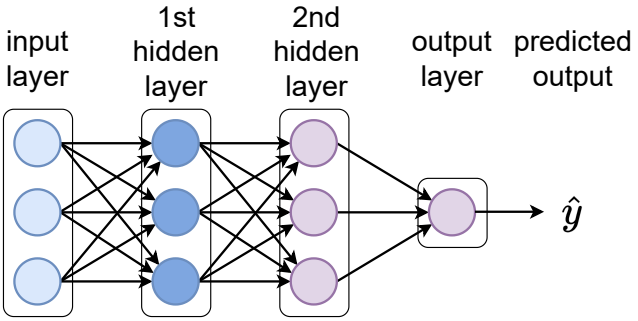


Fig. 3.3.: Depiction of a neural network with an input layer, two hidden layers, and an output layer.

3.2.2. Neurons

A layer consists of a number of neurons. Each connecting neuron is also called a node or unit. For instance, the first hidden layer depicted in Fig. 3.3 consists of three neurons, which are fully connected to the neurons in the second hidden layer. Fully connected means that the output from the previous layer serves as an input for the current layer (Goodfellow et al., 2017, p. 164). In the next section, now that we introduced some basic neural network vocabulary, we introduce a simple probabilistic classifier to explain the inner workings of a neural network.

3.3. Probabilistic Classifier

In this thesis, we focus mainly on supervised machine learning tasks, i.e., the output is already known. Generalised, our objective is to predict whether a document belongs to either a positive or negative **class**. For this reason, we employ a set of probabilistic classifiers. A probabilistic machine learning classifier returns the probability that an observation belongs to a certain category or class. Generally, a probabilistic classifier consists of the following four components (Jurafsky and Martin, 2020, p. 77): (1) A representation of the input data; (2) a classification function to compute the predicted outcome $\hat{y}$; (3) an objective function that formalizes the goal to minimize the error between the predicted outcome $\hat{y}$ and the actual outcome y; and (4) an optimization algorithm to achieve the error minimization, i.e., to optimize the objective function.

For explaining the probabilistic classifier, let us assume that a document is assigned either a positive class label ($y = 1$) or a negative class label ($y = 0$). A classification task like this – with only two distinct class labels – is called a binary classification task. In a binary classification task, the last layer of the network is a logistic function, which serves as the classification function. Hence, we also call this probabilistic classifier a **logistic classifier**.

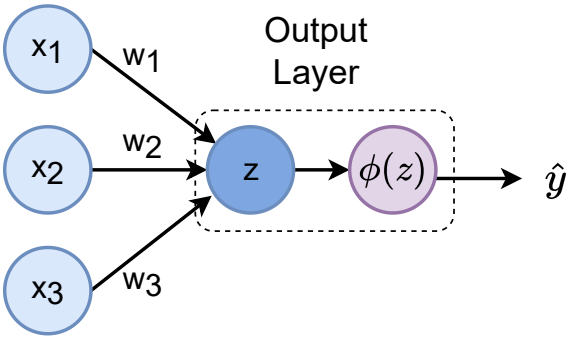


Fig. 3.4.: Exemplary Logistic Classifier with an Expanded Neural Network Structure.

Figure 3.4 depicts the neural structure of such an exemplary logistic classifier. Note that the output layer is expanded to z and $\phi(z)$ so that the inner functioning of the network becomes more clear. The neural network predicts some value $\hat{y}$ based on some given input x_1, x_2, and x_3, where each j for x_j is called a **feature**. Features are the representation of the input data. For example, distinctive features of a document could be certain word-counts, the text length, or the document's genre. For each observation i consisting of n features, we can write an observation as a vector $[x_1, x_2, ..., x_n]$ or, applying vector notation, as $\boldsymbol{x}^{(i)}$ for the i-th observation.[2]

The features are passed to a linear node z, which combines the according features with a given weight w_j, with j being the weight of feature x_j. The weights can be interpreted as the signal strength for a given feature x_j. Thus, the net input z can be written as $z = w_1x_1 + ... + w_nx_n$, or as $z = \boldsymbol{w}^T\boldsymbol{x}$ using vector notation, with $\boldsymbol{w}^T$ being the transpose of vector $\boldsymbol{w}$. Often, we also add a **bias term**, the so called intercept, to the weighted inputs. The bias term shifts the net input toward the bias term in absence of any input (Goodfellow et al., 2017, p. 107). When including the bias term, the net input becomes $z = (\sum_{j=1}^{n} w_jx_j) + b$ or, equivalently, $z = \boldsymbol{w}^T\boldsymbol{x} + b$. The resulting net input z is then passed through the classification function $\phi(z)$. In a binary classification task, the classification function is the so called **sigmoid function**, also called **logistic function**:[3]

$$\phi(z) = \frac{1}{1 + e^{-z}}\,. \tag{3.1}$$

The sigmoid function (see Figure 3.5) squashes the weighted input features to the range of 0 and 1, and has the property that $\phi(z)$ and $\phi(-z)$ sum to 1. Furthermore, the sigmoid function has the property that $\phi(-z) = 1 - \phi(z)$. Therefore, we can interpret the value of $\phi(z)$ as the probability that $\hat{y}$ takes the value of 1, and $1 - \phi(z)$ as the probability

[2]Note that we use the lower-case, bold letter to denote a vector. Furthermore, in the literature, observations are also referred to as samples or instance.

[3]Compare for example Jurafsky and Martin, 2020, p. 77 sqq.

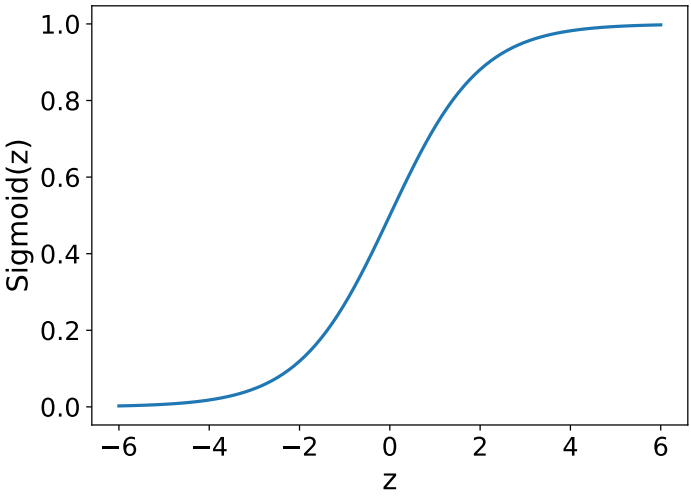


Fig. 3.5.: The sigmoid function maps real values to the range [0,1].

that $\hat{y}$ takes the value 0. Hence, if we set the **decision boundary** (or **threshold**) to 0.5, we can write the following unit step function:[4]

$$\hat{y} = \begin{cases} 1 & \text{if } \phi(z) \geq 0.5 \\ 0 & \text{otherwise,} \end{cases} \tag{3.2}$$

where $\hat{y}$ takes the value 1 if $\phi(z) \geq 0.5$ and 0 otherwise. Consequently, our network predicts whether a certain set of features belong to class $y = 1$ or $y = 0$. In other words, the network maps a set of input values to distinct output values (Goodfellow et al., 2017, p. 5).

However, the calculated output of the network, $\hat{y}$, depends on the weight parameters. This means that we need to choose the weight parameters in such a way that our network produces the desired output y – and not an output $\hat{y}$ that does not correspond to the true class label. Therefore, in

[4]Compare Jurafsky and Martin, 2020, p. 78-79.

the next section, we define the objective function that is used in a binary classification task and explain how to find the optimal weight parameters so that the network produces a $\hat{y}$ close to y for all our observations. This process of optimizing the network's weights is referred to as **training**.

3.4. Training and Testing

In this section, we explain how a neural network "learns", that is, how the weights are adjusted. Furthermore, we explain how to avoid an overfit of the neural network, and explain how to test and evaluate a neural network based on its output.

3.4.1. Training a Machine Learning Classifier

In a supervised learning task, we know the true class label $y^{(i)}$ for each observation $\boldsymbol{x}^{(i)}$. However, when training a classifier, we do not use all observations. Instead, we split the observations into a **training set** and a **test set**. Heuristics imply that a train-test split between 70/30, i.e., 70% training data and 30% test data, and 90/10 works best depending on the total number of observations (Raschka, 2015, p. 109). The training set is then used to train the classifier and the test set is used to evaluate the classifier.

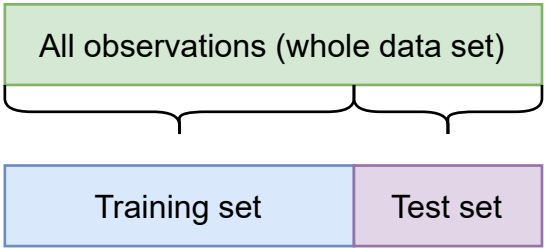


Fig. 3.6.: Training set and test set.

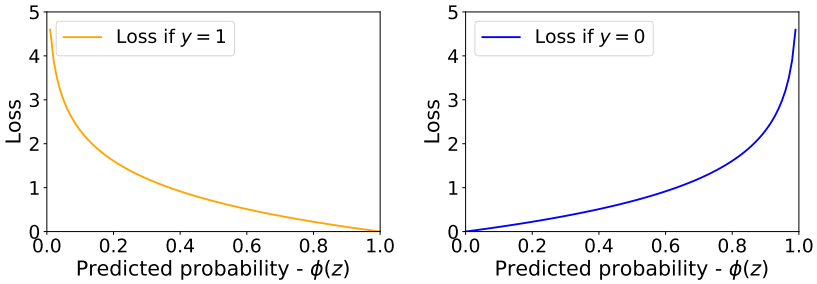
Let us assume that our network is still the network according to Fig. 3.4

and that we are faced with a binary classification task, i.e., the classification function is a logistic function. Therefore, our goal is to minimize the prediction error between the network's output $\hat{y}$ and the true class label y. To achieve this error minimization, we introduce the **loss function**.

The loss function calculates the distance, that is, the prediction error, between the network's prediction $\hat{y}$ and the true class label y. For a single observation i, the loss function for a binary logistic regression can be defined as (Raschka, 2015, p. 60)

$$L(\hat{y}, y) = -y \log(\phi(z)) - (1 - y) \log(1 - \phi(z)) \,, \tag{3.3}$$

with $\log(\cdot)$ being the natural logarithm. Since Equation 3.3 is also the binary cross-entropy that measures the performance of a classifier whose output is a probability value between 0 and 1, we sometimes refer to the loss as the cross-entropy loss. The cross-entropy loss increases if the predicted probability diverges from the actual class label y and decreases if the predicted probability gets closer to the true class label.



(a) Loss function if $y = 1$.

(b) Loss function if $y = 0$.

Fig. 3.7.: Loss function for class labels $y = 1$ and $y = 0$.

Figure 3.7 shows the loss for different values of $\phi(z)$. If the true class label is equal to 1 and the predicted probability is $\phi(z) = 1$, the loss is equal to 0 (see Fig. 3.7a). Equivalently, the loss tends to infinity if the

true class label is equal to 1 and $\phi(z) = 0$, i.e., the prediction is far away from the true class label. On the other hand, the opposite is true if the true class label y is equal to 0 (see Fig. 3.7b): In the case of $\phi(z) = 1$, the loss goes towards infinity, and if $\phi(z) = 0$ the loss is equal to 0. Hence, the loss function penalizes for high deviations from the true class more than for small deviations. In the following, we show how we can use this property to iteratively optimize our weights.

3.4.1.1. Gradient Descent

Our objective is to minimize the distance between $\hat{y}$ and y, i.e., to minimize the prediction error of the network. Therefore, we have to minimize the loss function to improve our predictions. We do that in calculating the **gradient** of the loss function. The gradient of the loss function is basically a vector, which contains the partial derivatives of the loss function with respect to all weights. The gradient shows the direction to the global minimum of the loss function, i.e., the set of weights that minimize the loss function. As the loss function is parameterized by the weights, we write the gradient as (Jurafsky and Martin, 2020, p. 83)

$$\nabla_{\boldsymbol{w}} L(\phi(x; \boldsymbol{w}), y) = \begin{bmatrix} \frac{\partial}{\partial w_1} L(\phi(x; \boldsymbol{w}), y) \\ \frac{\partial}{\partial w_2} L(\phi(x; \boldsymbol{w}), y) \\ \vdots \\ \frac{\partial}{\partial w_n} L(\phi(x; \boldsymbol{w}), y) \end{bmatrix} . \tag{3.4}$$

Note that we denote $\hat{y}$ in $L(\hat{y}, y)$ as $\phi(x; \boldsymbol{w})$, with the parameters $\boldsymbol{w}$ being separately shown, to symbolize the dependence of the loss on the weight parameters.

However, since the loss function is a function formed by other function, we need to apply the chain rule of calculus to calculate the partial derivatives of the loss function with respect to the weights. The chain

rule of calculus is defined as (Goodfellow et al., 2017, p. 199)

$$\frac{\partial f(g(x))}{\partial x} = \frac{\partial f(g(x))}{\partial g(x)} \frac{\partial g(x)}{\partial x} \,. \tag{3.5}$$

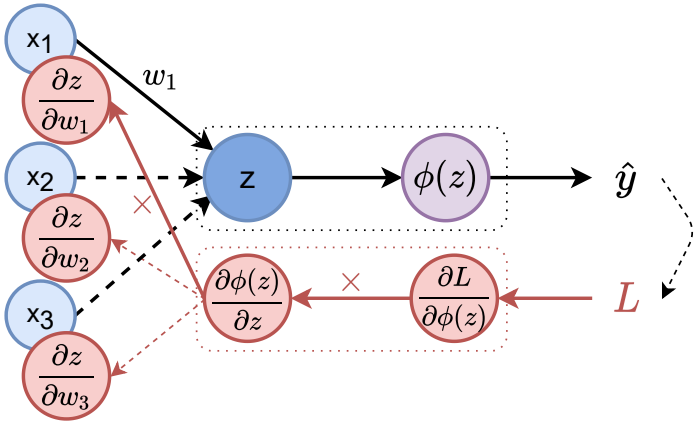


Fig. 3.8.: Schematic depiction of the chain rule of calculus.

For instance, let us assume we like to compute the first element of the gradient vector $\nabla L(\phi(x; \boldsymbol{w}), y)$, i.e., the partial derivative of $L(\phi(x; \boldsymbol{w}), y)$ with respect to w_1 (see Eq. 3.4). Further, let us assume that we just calculated the output of the network $\hat{y}$, and used $\hat{y}$ to calculate the loss according to Eq. 3.3. As the loss function depends on the sigmoid function $\phi(z)$ with z being also a function depending on w_1, we apply the chain rule of calculus as follows (see Fig. 3.8): Starting from the last node in our network, we take the partial derivative of the loss function with respect to the previous input $\phi(z)$, i.e., $\frac{\partial L(\phi(z))}{\partial \phi(z)}$. Since $\phi(z)$ is a function depending on z, we also take the partial derivative of $\phi(z)$ with respect to z, i.e., $\frac{\partial \phi(z)}{\partial z}$. Then, we apply the chain rule of calculus to compute $\frac{\partial L(\phi(z))}{\partial z}$ by multiplying both partial derivatives, i.e., $\frac{\partial L(\phi(z))}{\partial z} = \frac{\partial L(\phi(z))}{\partial \phi(z)} \frac{\partial \phi(z)}{\partial z}$. Next, we move one node further to the initial weights. It is now possible to compute $\frac{\partial L(\phi(z))}{\partial w_1} = \frac{\partial L(\phi(z))}{\partial z} \frac{\partial z}{\partial w_1}$, since we just calculated $\frac{\partial L(\phi(z))}{\partial z}$.

Thus, the total term becomes $\frac{\partial L(\phi(z))}{\partial w_1} = \frac{\partial L(\phi(z))}{\partial \phi(z)} \frac{\partial \phi(z)}{\partial z} \frac{\partial z}{\partial w_1}$. To update all of our network's weights, we apply this process for all weights in our network. Hence, we moved from the last node in the network to the first node in the network to compute the gradient, i.e., from right to left (see Fig. 3.8). This specific order of applying the chain rule of calculus is called **backpropagation**, since we calculate backwards through our network (Goodfellow et al., 2017, p. 199), and was first introduced by Rumelhart et al., 1986.

The gradient returns the slope of the loss function with respect to each weight. Thus, if we move in the opposite direction of the slope, we get closer to the global minimum, i.e., the slope gets flatter. Therefore, we update the weights as[5]

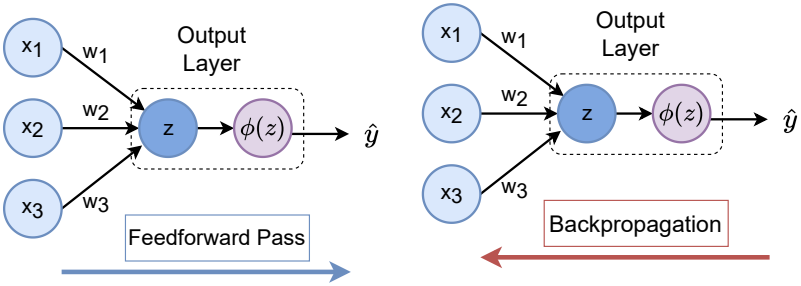
$$\boldsymbol{w} \coloneqq \boldsymbol{w} - \alpha \nabla_{\boldsymbol{w}} L(\phi(x; \boldsymbol{w}), y)\,. \tag{3.6}$$

Equation 3.6 is called the general equation of **gradient descent**. $\nabla_{\boldsymbol{w}} L(\phi(x; \boldsymbol{w}), y)$ denotes the gradient of the loss function and $\boldsymbol{w}$ is the corresponding weight vector. The parameter α is the so called **learning rate**. The learning rate is a positive scalar, which determines the size of the weight adjustment, i.e., the step size in the direction of the minimum.

The learning rate is a parameter that is *not* learned by the network. We call such parameters **hyperparameters**. Hyperparameters can be used to control the behaviour of a neural network. They need experience and fine-tuning to be chosen optimally (see Subsection 3.4.6).[6] If the learning rate is set too high, we likely overshoot the global minimum. If the learning rate is set too low, the learning process takes a considerable amount of time. This is due to the fact that the gradient declines as we approach the global minimum.

We see from Equation 3.6 that the closer the gradient is to 0, the smaller the weight adjustment. On the other hand, the greater the gradient, the larger the weight adjustment. For example, consider again

[5] Compare for example Jurafsky and Martin, 2020, p. 85.
[6] Compare for example Goodfellow et al., 2017, p. 117.



(a) Feedforward pass to compute $\hat{y}$.

(b) Backpropagation to compute gradient.

Fig. 3.9.: Feedforward pass and backpropagation.

the loss function for the different values of y (see Fig. 3.7b): If $y = 0$, the gradient, i.e., the slope, is more positive the steeper the incline of the loss curve. This means if $y = 0$ and the prediction is far away from the true class label, the more positive the gradient and, consequently, the greater the decrease in weights. On the other hand, if $y = 1$ and the prediction is far away from the true class label, the more negative the gradient and, consequently, the greater the increase in weights (see Fig. 3.7a). We refer to this process of using the gradient to adjust the weights as gradient descent (Goodfellow et al., 2017, p. 198).

Accordingly, we can summarize the training process of the network as follows: Initially, the weights are randomly set, often according to a uniform distribution. Then, we pass the input through the network and calculate the output, i.e., forward propagate the input (see Fig. 3.9a). The output is used to compute the loss function. Next, we back-propagate the error through the network (see Fig. 3.9b) and update the weights according to Eq. 3.6. If we repeat this process several times, after each iteration, the loss will decrease and our prediction will be closer to the true class value.[7] We apply this process until the network produces a desired result, i.e., the network predicts $\hat{y}$ close to y.

[7]Note: If the loss does not decrease, this is an indication that we overshoot the global minimum and the learning rate must be decreased.

3.4.1.2. Batch Training

Each pass over the entire training set, i.e., the observations used for training the classifier, is called an **epoch**. The total number of epochs are the number of complete passes over the entire training dataset. However, it can be computationally efficient to split an epoch into several **batches** (Raschka, 2015, p. 43). The batch size is the number of observations that are processed before the model is updated, i.e., the weights are adjusted. For instance, if we have 1.000 observations, a batch size equal to 5, and we do 10 epochs over the training set, this means that we create 200 batches containing five observations each. Therefore, the model adjusts the weights 200 times in one epoch and a total of 10 passes over the entire training set are done. Both the number of epochs and the batch size are hyperparameters (Brownlee, 2018).

If we use all training observations to create one batch, the learning algorithm is called **batch gradient descent**. This means that the errors are accumulated across an epoch before updating the weights. When the batch contains just one observation, the learning algorithm is called **stochastic gradient descent**. If the batch size is between 1 and the size of the training set, we call it **mini-batch gradient descent** (Brownlee, 2018). Consequently, we need to adjust the loss function for considering more than one observation.

In equation 3.3, we assumed that we use only one observation to compute the loss. To pass several observations at once to the network, we simply average each individual loss. This average loss function is referred to as the cost function:[8]

$$C(\boldsymbol{w}) = \frac{1}{m} \sum_{i=1}^{m} [-y^{(i)} \log(\phi(z^{(i)}) - (1 - y^{(i)}) \log(1 - \phi(z^{(i)}))] \,, \qquad (3.7)$$

where m is the number of observations in the batch that we use for training, $y^{(i)}$ is the true class label for observation i, and $z^{(i)}$ is the net input of observation i that is passed through the sigmoid function.

[8] Compare for example Jurafsky and Martin, 2020, p. 88.

Hence, the cost function is basically the average loss among m number of observations (Kahn, 2019b). However, the training process according to gradient descent remains the same (see Subsection 3.4.1.1).

3.4.2. Training a Deep Learning Classifier

Optimizing the weights for a deep learning classifier is quite similar to training a machine learning classifier. First, we propagate the input through the network and calculate the output, i.e., doing a feedforward pass. Then, we use the output to calculate the cost function, that is, the error between the prediction and the true class label. Finally, we apply the back-propagation algorithm to find the partial derivatives with respect to each weight so that we are able to update the weights. This process is repeated several times till our network predicts the output precisely and also performs well on unseen observations that were not used for training. However, in a deep neural network we have several layers and, consequently, more weights to consider. Thus, to explain the training process of a deep neural network, we demonstrate the training process on a straightforward deep neural network.

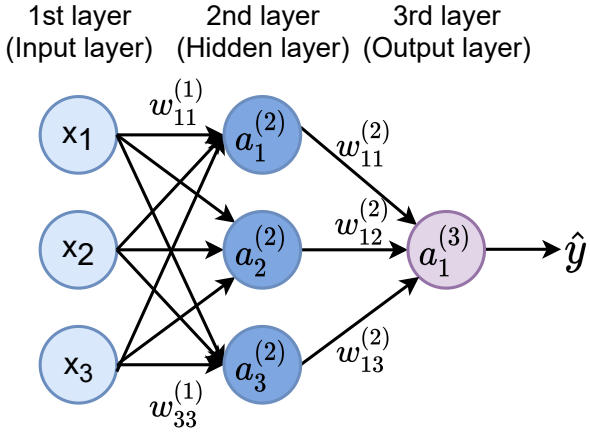


Fig. 3.10.: Three layer neural network.

Figure 3.10 shows an exemplary deep neural network, which consists of three layers: An input layer, a hidden layer, and an output layer. Each layer l is connected to all nodes in layer $l+1$ through the weights $w_{ij}^{(l)}$, where i denotes the node number of the connection layer $l+1$ and j refers to the node number in the previous layer l. Similarly, each activation node is denoted as $a_j^{(l)}$, with j referring to the number of the node and l being the layer number. Mathematically, we can write the network as the equation system

$$\begin{aligned} a_1^{(2)} &= f(w_{11}^{(1)} x_1 + w_{12}^{(1)} x_2 + w_{13}^{(1)} x_3)\,, \\ a_2^{(2)} &= f(w_{21}^{(1)} x_1 + w_{22}^{(1)} x_2 + w_{23}^{(1)} x_3)\,, \\ a_3^{(2)} &= f(w_{31}^{(1)} x_1 + w_{32}^{(1)} x_2 + w_{33}^{(1)} x_3)\,, \\ a_1^{(3)} &= f(w_{11}^{(2)} a_1^{(2)} + w_{12}^{(2)} a_2^{(2)} + w_{13}^{(2)} a_3^{(2)})\,. \end{aligned} \tag{3.8}$$

Each consecutive node takes the previous nodes' output as its input. For instance, $a_1^{(3)}$ takes $a_1^{(2)}$, $a_2^{(2)}$, and $a_2^{(2)}$ as its input. The input is weighted with the corresponding weights and the resulting net input is passed through the **activation function** $f(\cdot)$, with the activation function being some non-linear function, e.g., the sigmoid function.

From Eq. 3.8, we also see that the weights between the input layer and the hidden layer can be denoted as the matrix

$$W^{(1)} = \begin{bmatrix} w_{11}^{(1)} & w_{12}^{(1)} & w_{13}^{(1)} \\ w_{21}^{(1)} & w_{22}^{(1)} & w_{23}^{(1)} \\ w_{31}^{(1)} & w_{32}^{(1)} & w_{33}^{(1)} \end{bmatrix}\,, \tag{3.9}$$

with $W^{(1)} \in \mathbb{R}^{h \times n}$, where h is the number of units in the second layer and n is the number of input units. Furthermore, we see that the weights between the hidden layer and the output layer can be denoted as the vector – or, more generally speaking, as the array – $W^{(2)} = [w_{11}^{(2)}, w_{12}^{(2)}, w_{13}^{(2)}]$, with $W^{(2)} \in \mathbb{R}^{t \times h}$, where t is the number of output units and h the number

of units in the second layer.[9] Thus, by applying vector notation, we can also write the network for a single observation i as[10]

$$\begin{aligned} \boldsymbol{z}^{(2)} &= W^{(1)}\boldsymbol{x}^{(i)}\,, \\ \boldsymbol{a}^{(2)} &= f(\boldsymbol{z}^{(2)})\,, \\ \boldsymbol{z}^{(3)} &= W^{(2)}\boldsymbol{a}^{(2)}\,, \\ \boldsymbol{a}^{(3)} &= f(\boldsymbol{z}^{(3)})\,. \end{aligned} \tag{3.10}$$

Here, $\boldsymbol{z}^{(2)}$ is the net input of the second layer, i.e., $\boldsymbol{z}^{(2)}$ serves as the input for activation node $\boldsymbol{a}^{(2)}$. Accordingly, $\boldsymbol{z}^{(3)}$ is the net input of layer 3. Note that $\boldsymbol{a}^{(2)} = [a_1^{(2)}, a_2^{(2)}, a_3^{(2)}]$, and $f(\cdot)$ is the activation function applied to the units of each layer. In general, a multi-layer network can thus be expressed as

$$\begin{aligned} \boldsymbol{z}^{(l+1)} &= W^{(l)}\boldsymbol{a}^{(l)}\,, \\ \boldsymbol{a}^{(l+1)} &= f(\boldsymbol{z}^{(l+1)})\,. \end{aligned} \tag{3.11}$$

Now, if the activation function $f(\cdot)$ of the last activation node $\boldsymbol{a}^{(3)}$ in Eq. 3.10 is a logistic function, the neural network predicts whether a certain set of features belong to class $y = 1$ or $y = 0$, i.e., the network is a logistic classifier. Again, for this binary classification task, we minimize the error between the predicted outcome $\hat{y}$ and the actual outcome y. For this reason, we employ the binary cross-entropy that we introduced in the previous subsection (see Eq. 3.7):[11]

$$C(\boldsymbol{w}) = \frac{1}{m}\sum_{i=1}^{m} L(\hat{y}^{(i)}, y^{(i)})\,. \tag{3.12}$$

To find the minimum of this cost function, we compute the gradient.

[9] Compare Raschka, 2015, p. 350.

[10] Compare for example Raschka, 2015, p. 345-450.

[11] Note that practical implementations in Python use a more stable loss function, e.g., TensorFlow: $C(Y, Z) = \frac{1}{m}\sum max(Z, 0) - ZY + \ln(1 + e^{|-Z|})$ with the maximum function being the element-wise maximum (Kahn, 2019a).

Therefore, we take the partial derivative of C with respect to each single weight by applying the chain rule of calculus. Generally, we can denote this for each weight in our network as

$$\frac{\partial C(w)}{\partial w_{ij}^{(l)}} = \frac{\partial C}{\partial a_i^{(l+1)}} \frac{\partial a_i^{(l+1)}}{\partial z_i^{(l)}} \frac{\partial z_i^{(l)}}{\partial w_{ij}^{(l)}} \,. \tag{3.13}$$

However, in a deep network we also need to back-propagate through each activation node to adjust also the weights of the previous layers. We do this by calculating[12]

$$\frac{\partial C(w)}{\partial a_j^{(l)}} = \sum_{i=0}^{s_{l+1}-1} \frac{\partial C}{\partial a_i^{(l+1)}} \frac{\partial a_i^{(l+1)}}{\partial z_i^{(l)}} \frac{\partial z_i^{(l)}}{\partial a_j^{(l)}} \,, \tag{3.14}$$

where s_{l+1} is the number of nodes in layer $l+1$.

To illustrate the backpropagation algorithm, Figure 3.11 depicts the general gradient flow in a deep neural network. After the input is passed through the network from left to right, we use the predicted outcome $\hat{y}$ to calculate the cost function. Then, at each node, we compute the partial derivative of the cost function with respect to the node itself to derive the upstream gradient $\frac{\partial C}{\partial z}$. Next, we calculate the partial derivatives of that node with respect to the inflowing information, i.e., the local gradients.[13] We multiply the upstream gradient with each local gradient to compute the out-flowing downstream gradients. The downstream gradients are then passed to the next node in line. At the subsequent node, the downstream gradient becomes our next upstream gradient and we repeat this process until we are able to update the weights between the input layer and the hidden layer. However, the gradients do not propagate back to the input nodes as this would alter the input data.[14]

Now, let us demonstrate the gradient flow for our exemplary deep neural

[12] Compare Sanderson and Pullen, 2021, i.a.

[13] Note: Since both x and y are incoming information, we have to calculate two local gradients. However, sometimes there is only one local gradient to compute.

[14] Compare Kahn, 2019b.

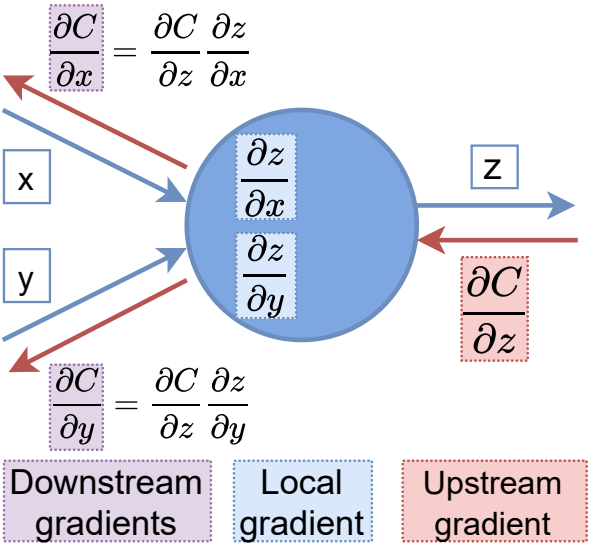


Fig. 3.11.: Illustration of the Gradient Flow for a single node z, with x and y being the inflowing information from the feedforward pass (blue arrows) and the backpropagation is signalled by the red arrows; drawn after F.-F. Li, 2019, p. 58.

network (see Fig. 3.10). First, we generalize the network, to include all m observations in a batch. Hence, let X_{batch} be a $[n \times m]$ dimensional matrix with n being the number of features and m denoting the number of observations used in the batch. We can then rewrite Eq. 3.10 for all m samples as

$$\begin{aligned} Z^{(2)} &= W^{(1)} X_{\text{batch}}\,, \\ A^{(2)} &= f(Z^{(2)})\,, \\ Z^{(3)} &= W^{(2)} A^{(2)}\,, \\ A^{(3)} &= f(Z^{(3)})\,. \end{aligned} \tag{3.15}$$

$A^{(2)}$ is the activation matrix, with $A^{(2)} \in \mathbb{R}^{h \times m}$. Note that we apply the activation functions $f(\cdot)$ to each value in the net input matrix $Z^{(2)}$ and $Z^{(3)}$, with $Z^{(2)} \in \mathbb{R}^{h \times m}$ and $Z^{(3)} \in \mathbb{R}^{t \times m}$. $A^{(3)}$ contains the estimated probabilities for each observation, which are used to calculate the cost.[15]

The backpropagation algorithm for the network is as follows: We start by computing the first upstream gradient $\frac{\partial C}{\partial f(Z^{(3)})}$. The local gradient is $\frac{\partial f(Z^{(3)})}{\partial Z^{(3)}}$, so we can compute the downstream gradient as $\frac{\partial C}{\partial Z^{(3)}} = \frac{\partial C}{\partial f(Z^{(3)})} \frac{\partial f(Z^{(3)})}{\partial Z^{(3)}}$. At the next node, the downstream gradient $\frac{\partial C}{\partial Z^{(3)}}$ becomes our new upstream gradient. Now, we have two local gradients $\frac{\partial Z^{(3)}}{\partial W^{(2)}}$ and $\frac{\partial Z^{(3)}}{\partial A^{(2)}}$. Thus, we compute $\frac{\partial C}{\partial W^{(2)}} = \frac{\partial C}{\partial Z^{(3)}} \frac{\partial Z^{(3)}}{\partial W^{(2)}}$ and $\frac{\partial C}{\partial A^{(2)}} = \frac{\partial C}{\partial Z^{(3)}} \frac{\partial Z^{(3)}}{\partial A^{(2)}}$. Then, we repeat this process for the next node. Since we do not propagate $W^{(2)}$ any further, we consider only $\frac{\partial C}{\partial A^{(2)}}$, which becomes our new upstream gradient. At the next node, the local gradient is $\frac{\partial A^{(2)}}{\partial Z^{(2)}}$. The downstream gradient can thus be computed as $\frac{\partial C}{\partial Z^{(2)}} = \frac{\partial C}{\partial A^{(2)}} \frac{\partial A^{(2)}}{\partial Z^{(2)}}$, which consequently becomes our next upstream gradient. In the last step, the local gradient is $\frac{\partial Z^{(2)}}{\partial W^{(1)}}$, so we compute the downstream gradient as $\frac{\partial C}{\partial W^{(1)}} = \frac{\partial C}{\partial Z^{(2)}} \frac{\partial Z^{(2)}}{\partial W^{(1)}}$.

Having derived the gradients of the cost function with respect to all weight matrices in the network, we update the weights as

$$\begin{aligned} W^{(2)} &:= W^{(2)} - \alpha \frac{\partial C}{\partial W^{(2)}} \,, \\ W^{(1)} &:= W^{(1)} - \alpha \frac{\partial C}{\partial W^{(1)}} \,. \end{aligned} \tag{3.16}$$

We can generalize Eq. 3.16 to[16]

$$W^{(l)} := W^{(l)} - \alpha \frac{\partial C(w)}{\partial W^{(l)}} \,. \tag{3.17}$$

Thus, we demonstrated how to update a deep neural network using gradient descent. However, if we repeat the learning process for too

[15] Compare Raschka, 2015, p. 350.
[16] Compare for example Kahn, 2019b.

many epochs, the classifier will indeed memorize the observations used for training but fail to work well on unseen observations, i.e., on observations that were not used for training. Since this is not a desired result, we use regularization techniques.

3.4.3. Regularization

After training the network for several epochs, our network will almost perfectly predict the observations used for training. However, this most likely leads to a classifier, which works well on the training data but fails to predict observations that were not used for training. This problem is known as **overfitting**. As we prefer a classifier that also performs well on unseen data, i.e., a classifier that is generalizable (Goodfellow et al., 2017, p. 115 sqq.), two common techniques are used to avoid overfitting: (1) L1/L2 regularization; and (2) dropout.

3.4.3.1. L1 and L2 Regularization

To prevent overfitting, L1 and L2 regularization add a regularization term to the cost function. The L1 and L2 regularization terms are defined as

$$\begin{aligned} L1 &= \lambda \|\boldsymbol{w}\|_1^1 = \lambda \sum_{j=1}^{m} |w_j| \,, \text{ and} \\ L2 &= \lambda \|\boldsymbol{w}\|_2^2 = \lambda \sum_{j=1}^{m} w_j^2 \,. \end{aligned} \tag{3.18}$$

Both regularization terms penalize extreme weight parameters when performing gradient descent. However, the L1 term eliminates features that are not considered important for the prediction task. This means that certain weights are set equal to zero if we apply the gradient descent process with L1 regularization. On the other hand, the L2 term calculates the sum of the squared weights. This shrinks the weights towards zero but does not fully eliminate the weights. Both L1 and L2 regularization reduce

the risk of an overfit. In general, we employ either L1 regularization, L2 regularization, or a combination of both.[17]

3.4.3.2. Dropout

Another regularization technique is dropout regularization (Hinton et al., 2012; Dahl et al., 2013; Srivastava et al., 2014, i.a.). With dropout regularization, a number of nodes within a layer are randomly omitted when training the model. The omitted nodes are temporarily removed from the network, along with the incoming and outgoing connections (Srivastava et al., 2014, p. 1930). This means a single unit cannot rely on the input of a particular previous node as it might be omitted during the learning process. Hence, the network spreads out the weights more than it would do without dropout. This ultimately leads to a regularization of the weights. The dropout *rate* states how many of the nodes are omitted. The higher the rate, the more nodes are randomly omitted. For example, a dropout rate of 50% states that for each iteration 50% of the nodes, to which the dropout regularization is applied, are randomly omitted during training.

3.4.4. Evaluation – Precision, Recall, F-score

In this subsection, we introduce some common evaluation metrics to evaluate the performance of a machine learning or deep learning classifier.[18] The evaluation metrics tell us how well the classifier performed in predicting the observations' original class labels y. We refer to the original pre-assigned class labels also as the **gold labels**. A confusion matrix illustrates the classifier's performance with respect to the gold labels (see Fig. 3.12). The confusion matrix has two dimensions: (1) The classifiers predicted output $\hat{y}$; and (2) the gold labels y. Generally, we prefer a classifier that predicts $y = 1$ as $\hat{y} = 1$ (true positives) and

[17] For more details on L1 and L2 regularization see Jurafsky and Martin, 2020, p. 88 sqq., Goodfellow et al., 2017, p. 221 sqq., or Raschka, 2015, p. 365 sqq.

[18] For the calculation of the evaluation metrics, compare for example Jurafsky and Martin, 2020, p. 65 sqq.

$y = 0$ as $\hat{y} = 0$ (true negatives). Thus, the first metric we introduce is **Accuracy**.

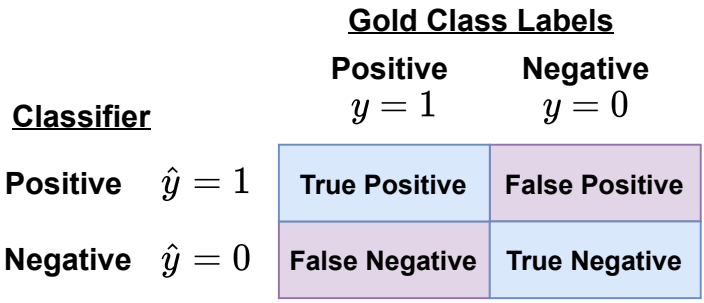


Fig. 3.12.: Confusion matrix for a binary classification task.

Accuracy is the percentage of all observations that the classifier labelled correctly. More precisely, accuracy shows the number of correctly predicted class 1 observations (true positives) and correctly predicted class 0 observations (true negatives) in proportion to all predictions. The accuracy is calculated as

$$\text{accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FN} + \text{FP}}\,, \tag{3.19}$$

with TP being the true positives, TN are the true negatives, FP are the false positives, and FN are the false negatives.

However, in a text classification task, accuracy is generally not used since accuracy is not suited for unbalanced classes (Jurafsky and Martin, 2020, p. 65). For instance, consider a case with 10 actual positive observations and 1000 actual negative observations, i.e., the classes are heavily unbalanced. If a classifier simply decides to predict that all observations belong to the negative class, the accuracy will be 99%. However, the classifier would be quite useless, since the classifier is unable to predict the positive classes – which might be the purpose for which the classifier was designed. Therefore, in text classification tasks, we use measures like precision, recall, and F_1-score.

Precision is the ratio of positive predicted classes that are indeed positive. Consequently, precision tells us the proportion of positive identified classes that are actually correct. Precision ranges from 0 to 1 and is calculated as

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \,. \tag{3.20}$$

From Equation 3.20, we see that we obtain a precision of 1, if the classifier predicts all actual positive classes correctly and avoids classifying negative classes as positive. For instance, let us assume we have two observations, one observation that belongs to class 1 ($y = 1$) and the other observation belongs to class 0 ($y = 0$). If the classifier predicts both observations to be class 1 ($\hat{y} = 1$), precision will be 0.5, i.e., 50% is classified correctly. However, if the classifier predicts only the observation that truly belongs to class 1 as class 1 (avoiding false positives), precision will be 1.

On the other hand, **recall** states the proportion of actual positives that were identified correctly. Therefore, recall can be interpreted as the true positive rate, i.e., the true positives out of all positives (Raschka, 2015, p. 192). Recall ranges from 0 to 1 and is calculated as

$$\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \,. \tag{3.21}$$

So considering again our example with one observation belonging to class 1 ($y = 1$) and the other observation to class 0 ($y = 0$): A classifier that predicts both observations to be class 1 has a recall of 1, simply by avoiding false negatives. However, we know the classifier is not perfect, as it predicted one out of two observations incorrectly, i.e., generating a false positive. Optimally, we want a classifier to avoid both false positives (type I errors) and false negatives (type II errors). However, increasing recall often decreases precision and vice versa. Thus, the last measure that we introduce is the F_1-score.

F_1-score sets precision and recall in relation to each other. F_1-score weights the importance of precision and recall equally, and is calculated

as (van Rijsbergen, 1975)

$$F_1\text{-}score = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \,. \tag{3.22}$$

Like recall and precision, F_1-score ranges from 0 to 1. The higher the F_1-score, the better the classifier.

3.4.5. Comparing Two Classifiers

Often we like to compare whether a classifier A is significantly better than another classifier B. If we have trained both classifiers on the same training data X_{Train}, we can calculate the absolute difference in performance on the test set X_{Test} as (Berg-Kirkpatrick et al., 2012)

$$\delta(X_{\text{Test}}) = |F_1\text{-}score(A, X_{\text{Test}}) - F_1\text{-}score(B, X_{\text{Test}})| \,. \tag{3.23}$$

$\delta(X_{\text{Test}})$ is referred to as the test statistic.[19] We perform a bootstrap test, to analyse whether the performance difference is statistically significant (Efron and Tibshirani, 1993). Under the assumption that our test set X_{Test} correctly represents the population, we can pretend that we draw from the original population by creating a bootstrap distribution. We do that by drawing n observations from our original test set X_{Test} with replacement. A newly created sample $\hat{X}_{\text{Test}}$ has the same number of observations n as the original sample X_{Test}. For each $\hat{X}_{\text{Test}}$, we calculate the test statistic according to Eq. 3.23. We repeat this process b times, also called the number of bootstrap re-samples.[20] Then, we can calculate the p-value as (Riezler and Maxwell, 2005)

$$p\text{-}value = \frac{1 + \sum_{i=1}^{b} \mathbb{1}(\delta(\hat{X}_{\text{Test}}) - \tau_b \geq \delta(X_{\text{Test}}))}{1 + b} \,, \tag{3.24}$$

[19] Note that Berg-Kirkpatrick et al., 2012 denote the test set as x. However, to be consistent with our previous notation, we write X_{Test}.
[20] Note that b has to be quite large, e.g. 1,000 or 10,000.

with τ_b being the sample mean over the bootstrap samples $\frac{1}{b}\sum_{i=1}^{b}\delta(\hat{X}_{\text{Test}})$. This p-value can be used for a standard hypothesis test. We reject the null hypothesis $H_0 : \delta(X_{\text{Test}}) \leq 0$, that is, classifier A is not better than B, if the *p-value* is less than or equal to the specified significance level. In this case, we find support for $H_1 : \delta(X_{\text{Test}}) > 0$, i.e., A is better than B.[21]

3.4.6. Optimization of Hyperparameters

We have introduced hyperparameters as the parameters that are not learned when training the network and that can be used to control the behaviour of the network. In practice, it requires experience and tinkering to find the optimal hyperparameters for a model, that is, the hyperparameters that optimize our classifier. Hyperparameters are, for example, the learning rate, the drop-out rate, the nodes in a layer, and also the number of layers. In this subsection, we introduce some methods to find the optimal parameters.

3.4.6.1. Train, Validation, and Test Sets

We have introduced the training set as the set, which is used to train the classifier, and the test set as the set of observations to evaluate the performance of the classifier. However, if we use the test set to select the optimal hyperparameters and the selection process is repeated several times, the test set would ultimately become part of our training set. To avoid this, we use part of the training set as a **validation set** to select the optimal parameters (see Fig. 3.13). One way to approach this is to apply a 90/10 training-validation split on our training set. However, a more common technique is applying **k-fold cross-validation**.

K-fold cross-validation is a re-sampling technique without replacement. The training set is split into k-folds, where one fold is used as the validation set. After each training iteration, a new fold is selected for evaluating the model parameters. For instance, if we apply 10-fold cross-validation, the

[21] Also see Dror et al., 2020.

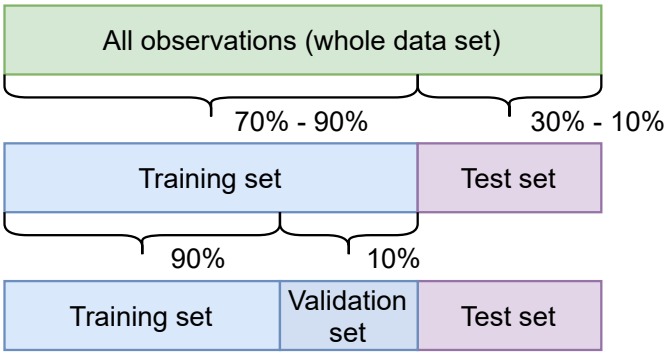


Fig. 3.13.: Training data, validation data, and test data.

training set is split into 10-folds of which one fold is used for evaluating the model after each epoch. To ensure that each fold contains a similar amount of classes, i.e., class 1 ($y = 1$) and class 0 ($y = 0$), we apply stratified k-fold cross-validation (Kohavi et al., 1995). Stratified k-fold cross-validation ensures that each fold has approximately a proportional amount of classes as the whole training set.

3.4.6.2. Grid Search

In using a validation set, we ensure that the test set does not become part of our training set. Hence, we can now select a set of different values for the hyperparameters and "search" an optimal combination of hyperparameters when training the network. We do that by applying a **grid search**. Grid search is simply a brute-force, exhaustive search for the optimal parameters (Raschka, 2015, p.185-187). In this way, we can fine-tune the hyperparameters.

3.4.6.3. Learning rate – ADAM

The learning rate α controls how fast a neural network learns. If the learning rate is set too low, each iterative step of adjusting the weights will only have a minor effect on the new weights. Thus, the time to approach

the global minimum will be very long. On the other hand, if the learning rate is set too high, we might overshoot the global minimum and never approach the minimum of the objective function. Therefore, we introduce an algorithm to adjust the learning rate while training the network. This algorithm is called Adaptive Moment Estimation, i.e., **Adam**, and was first introduced by Kingma and Ba, 2015.

Adam is an optimization algorithm that adapts the learning rate at each timestep by employing first and second moments of the gradient (Kingma and Ba, 2015). This leads to an iteratively adjusted learning rate that approaches the minimum of the loss function much faster than other optimizers (Ruder, 2016). With Adam, all elements of the initial first and second moment vectors are set to zero. We denote the initial first and second moments as $\boldsymbol{m}_0$ and $\boldsymbol{v}_0$, respectively. At each update step t, we compute the gradient $\nabla_{\boldsymbol{w}} C_t(\boldsymbol{w_{t-1}}) \coloneqq \boldsymbol{g}_t$. Then, we update the moment vectors as

$$\begin{aligned} \boldsymbol{m}_t &= \beta_1 \boldsymbol{m}_{t-1} + (1-\beta_1)\boldsymbol{g}_t \,, \\ \boldsymbol{v}_t &= \beta_2 \boldsymbol{v}_{t-1} + (1-\beta_2)\boldsymbol{g}_t^2 \,, \end{aligned} \tag{3.25}$$

with $\beta_1, \beta_2 \in [0, 1]$ being the exponential decay rates. Here, $\boldsymbol{g}_t^2$ indicates the elementwise square, i.e., $\boldsymbol{g}_t \odot \boldsymbol{g}_t$ with $\odot$ denoting the hadamard product. Subsequently, we correct the first and second moment estimates by computing[22]

$$\begin{aligned} \hat{\boldsymbol{m}}_t &= \frac{\boldsymbol{m}_t}{(1-\beta_1^t)} \,, \\ \hat{\boldsymbol{v}}_t &= \frac{\boldsymbol{v}_t}{(1-\beta_2^t)} \,. \end{aligned} \tag{3.26}$$

In the last step, we can use the biased-corrected moments $\hat{\boldsymbol{m}}_t$ and $\hat{\boldsymbol{v}}_t$ to update the weight parameters as

$$\boldsymbol{w}_t = \boldsymbol{w}_{t-1} - \alpha \frac{\hat{\boldsymbol{m}}_t}{(\hat{\boldsymbol{v}}_t^{0.5} + \epsilon)} \,, \tag{3.27}$$

[22] Note that β_1^t and β_2^t are to the power of t.

with $\epsilon = 10^{-8}$ to ensure that we do not divide by zero. Kingma and Ba, 2015 assumed $\alpha = 0.001$ $\beta_1 = 0.9$ and $\beta_2 = 0.999$.[23]

Now, that we explained how to efficiently train and evaluate classifiers, we discuss a variety of approaches to extract meaningful features from texts. Hence, the next section addresses how to preprocess texts and obtain meaningful feature representations from texts.

3.5. Text Classification

In order to process textual data, texts must be in a computer readable format so that we can feed the texts to a neural network (McEnery, 2005, p. 449). Hence, the first step is to construct a **corpus**. A corpus is a collection of texts that are used as a "statistically valid base for computational processing", Hanks, 2005, p. 53. Theoretically, a corpus can consist of a collection of single sentences, multiple sentences, or entire documents. However, a collection of texts is usually unstructured. Therefore, the texts need to be standardized – a step called **preprocessing**.

3.5.1. Textual Preprocessing

Before we can feed the texts to a neural network, we need to preprocess the texts. For this reason, let us consider the case where we have a collection of documents. Each document consists of several sentences, which themselves consist of several words. Thus, we have to split the documents into smaller linguistic units, i.e., **tokens**. A token can be a word, a number, or other non-alphanumeric characters and strings (Mikheev, 2005). The process of splitting documents into tokens is referred to as **tokenization**. For a human, the tokenization is a straightforward task, however, it is a complex task for a computer. Consider a simple whitespace split, i.e., words are split into tokens at the whitespace: A word like "San Francisco" would be split into two independent tokens

[23] For more details on Adaptive Moment Estimation please refer to Kingma and Ba, 2015. For a comparison of different adaptive learning rate algorithms see Ruder, 2016, i.a.

"San" and "Francisco". However, we might want to consider "San Francisco" as one token. Furthermore, to find the end of sentences, it is not sufficient to look for words followed by a period. For example, words such as "Dr." and "Ms." complicate the localisation of the end of a sentence. Additionally, hyphenated words, apostrophes, or numeric sequences – like dates or email addresses – complicate the tokenization process (Mikheev, 2005).[24] Therefore, we apply specialized **tokenizers**.

Specialized tokenizers can recognize patters of email addresses, dates, or, URLs. Python libraries like spaCy (Honnibal et al., 2020) or NLTK (Bird et al., 2009) provide specialized tokenizers, which deal with most of the complex word and sentence splitting. These tokenizers return a "clean" list of tokens (Mikheev, 2005, p. 208-209).

Another preprocessing step is to remove **stopwords**. Stopwords are words that have functional purpose and do not transmit comparable content such as nouns, verbs, and adjectives. Stopwords are for example repositions, pronouns, and determiners (Tzoukermann et al., 2005, p. 531). Therefore, stopwords are often removed, when preprocessing texts.

Figure 3.14 depicts the preprocessing of a corpus. For each sentence in the corpus, the words are split into tokens. The stopwords "this", "is", and "an" are removed. In the preprocessing step, capitalized letters are also made lower case. Otherwise, a computer does not recognize that a capitalized word like "Money" and a lower case word like "money" are the same word.[25] Now let us move on to the extraction of features.

[24] Consider the following sentence: "The Coca-Cola Company conducted a stock split on the 05.01.1996.". Splitting sentences at the period and words at the whitespace character would split the name "The Coca-Cola Company" into three different tokens and end the sentence at "05.".

[25] Note that another possible step of preprocessing is **stemming** (Porter, 1980), i.e., obtaining the word root from words, e.g. "walk" or "goes" is stemmed to "go" (Tzoukermann et al., 2005, p. 531). However, we will not apply stemming for our purpose and refer the interest reader to Porter, 1980.

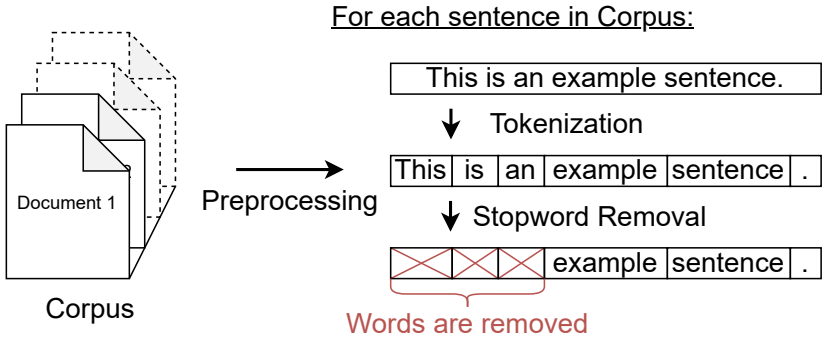


Fig. 3.14.: Depiction of a corpus of documents, where each sentence is tokenized into several words and stopwords are removed.

3.5.2. How to Extract Features From Texts – Feature Representation

In this section, we present several methods to extract meaningful features from a text. First, we introduce the standard Bag-of-Words approach and the term frequency-inverse document frequency (tf-idf) weightings. Next, we explain more complex methods like word embeddings and Bidirectional Encoder Representations from Transformers (BERT).

3.5.2.1. Bag-of-Words

Tokens need to be transformed into feature vectors to be machine readable. One way to do that is by counting the occurrence of a particular word within a text. This method is known as **Bag-of-Words** (BoW). Bag-of-Words treats the document as a unstructured list of words ignoring their previous position and keeping only track of their respective frequency (Jurafsky and Martin, 2020, pp. 56–57).

Figure 3.15 illustrates a Bag-of-Word approach. Suppose we have three documents, each document consisting of only one sentence. The vocabulary of this corpus is only 7 words, with the occurrence of "stock" being 4, "A" being 2, and so on. Now, we count the occurrence of each of those words within a document. Thus, we receive a feature vector for

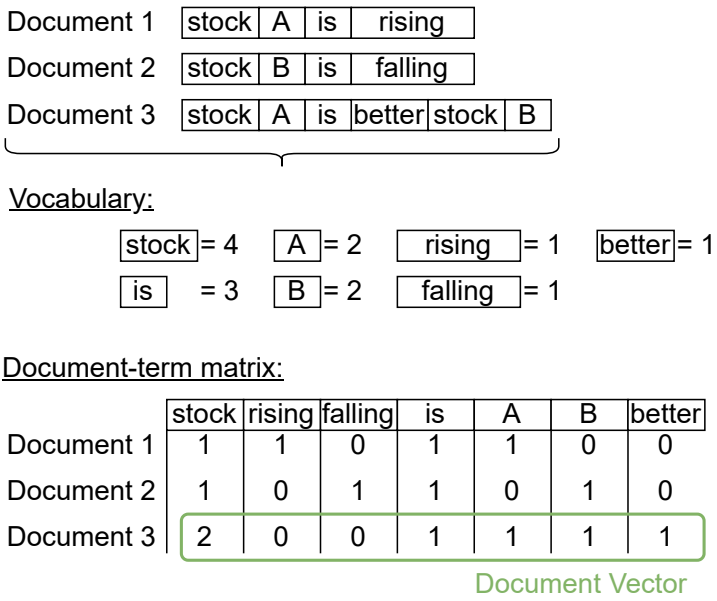


Fig. 3.15.: Example of the Bag-of-Words approach.

each document, depicted in the document-term matrix, i.e., the terms per document. However, we see that a document vector contains entries for the whole training set vocabulary. This makes the vectors quite sparse, i.e., the vectors contain a lot of zero entries.[26]

3.5.2.2. tf-idf weighting

The raw count of words is a first step towards meaningful vector representations. However, some words will occur more often in longer documents – and these words are often less meaningful words. Thus, many different weighting techniques can be applied (Salton and McGill, 1983). We briefly introduce the term frequency-inverse document frequency (tf-idf) technique (Salton and McGill, 1971; Salton and Lesk, 1968). Let $\mathrm{tf}_{t,d}$ be

[26] Note that in a real example, we use only the training set vocabulary of the corpus to create the document vectors.

the term frequency (Luhn, 1957) for term t in document d and n_d being the total number of documents. We compute the according weight for each term in a document as[27]

$$\textit{tf-idf}(t, d) = \mathrm{tf}(t, d) \times (\mathrm{idf}(t, d)) \,. \tag{3.28}$$

The inverse document frequency $\mathrm{idf}(t, d)$ is calculated as (Jones, 1972)

$$\mathrm{idf}(t, d) = \log\left(\frac{n_d}{\mathrm{df}_{d,t}}\right) + 1 \,, \tag{3.29}$$

with $\mathrm{df}_{d,t}$ being the number of documents that contain term t. The addition of 1 leads to terms occurring in all documents being not completely ignored. All words are then represented with a tf-idf weighting instead of the pure word count.

3.5.2.3. Word Embeddings

Compared to BoW or tf-idf feature vectors, word embeddings are short dense vector representations of words (Mikolov, Sutskever et al., 2013). Each embedding vector is of size $1 \times d$ with d denoting the dimension of the vector. Usually, the dimension is set to 100, 200, or 300. By representing words in a more compact form, the results of NLP classifiers can be significantly improved (Jurafsky and Martin, 2020, p. 122). Another advantage of word embeddings is that they preserve the relationship between words (Bengio et al., 2003). Studies have shown that words similar to one another are placed closer together in the vector space (Mikolov, Sutskever et al., 2013; Mikolov, Yih et al., 2013, i.a.) and even algebraic operations on the word vectors come close to human reasoning (Mikolov, Yih et al., 2013). For example, taking the vector(King) - vector(Man) + vector(Woman) results in a vector close to the vector(Queen) (Mikolov, Yih et al., 2013). However, word embeddings can be constructed in a variety of ways. Therefore, we briefly introduce the **skip-gram method** (Mikolov,

[27] Note that there are slight variations of how to implement tf-idf. We stick to the implementation from Pedregosa et al., 2011 as this is the one we will use.

Sutskever et al., 2013) and the **GloVe method** (Pennington et al., 2014) to construct word embeddings. Both methods are unsupervised learning models.

Skip-gram method: Suppose our vocabulary contains V unique words. Initially, all words are **one-hot-encoded**, meaning that a feature vector $\boldsymbol{x}$ representing word i, with $\boldsymbol{x} = [x_1, \cdots, x_i, \cdots, x_V]$ of size $V \times 1$, contains only zeros except for the position i, which is set to 1. Now, we feed each word in our training set to a network consisting of an input layer, a hidden layer, and an output layer. Given the context window C of the words $x_{1:C}$, the network predicts the probability that the target word i belongs to the context words $x_{1:C}$, that is, the network predicts a word's context $x_{1:C}$ given a word i.

For instance, consider the training sentence "The company turns a profit" (see Fig. 3.16). If the target word is "turn" and the context window is 2, the network tries to predict whether the predicted context $\hat{y}$ matches the actual context y. Thus, we update our network with gradient descent so that the network predicts the context for each word almost perfectly.

Under the assumption that the context words are independent from each other, the network is computed as[28]

$$\begin{aligned} \boldsymbol{h} &= W^{(1)}\boldsymbol{x}\,, \\ \boldsymbol{u} &= W^{(2)}\boldsymbol{h}\,, \end{aligned} \tag{3.30}$$

with $W^{(1)} \in \mathbb{R}^{N\times V}$ being the weight matrix between the input layer and the hidden layer $\boldsymbol{h}$, where N is the number of units in the hidden layer, and $W^{(2)} \in \mathbb{R}^{V\times N}$ being the weight matrix between the hidden layer $\boldsymbol{h}$ and the output layer $\boldsymbol{u}$. Since we predict the context words given an input word, we are faced with a multi-classification problem. Thus, we

[28]We follow the notation of Minnaar, 2015.

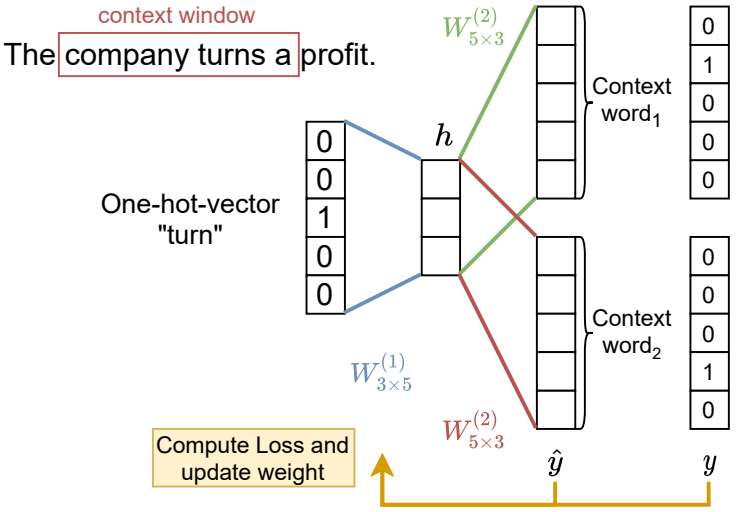


Fig. 3.16.: Depiction of a simplified skip-gram model with 5 words in the vocabulary and a window size of 2.

apply the softmax function

$$P(j|i) = y_j = \frac{e^{u_j}}{\sum_{j'=1}^{V} e^{u_{j'}}}\,, \tag{3.31}$$

where $P(j|i)$ denotes the probability that word j is a context word given word i as the input word. u_j denotes unit j of the output layer $\boldsymbol{u}$. Since the context words are assumed to be independent, we can multiply the probabilities as $\prod_{c=1}^{C} \frac{e^{u_{j_c^*}}}{\sum_{j'=1}^{V} e^{u_{j'}}}$, with j_c^* being the vocabulary indices of the context words $c \in [1, C]$. Note that we want to maximize this probability. However, we can turn the calculation into a minimization problem by minimizing the cross-entropy

$$L = -\sum_{c=1}^{C} u_{j_c^*} + C \log \left(\sum_{j'=1}^{V} e^{u_{j'}} \right)\,. \tag{3.32}$$

After training the network on a large body of text, the transpose of matrix $W^{(1)}$ contains in row i the word embeddings for word i. This reduces the size of our word vectors significantly, compared to the previous Bag-of-Words method. Additionally, the vectors contain the local context information of a word.

GloVe method: While the skip-gram algorithm relies only on the local information, i.e., the surrounding words, **GloVe embeddings** (Pennington et al., 2014) take also the global word statistic into account.[29] GloVe embeddings (global vector word representations) consider the global word statistic by employing a co-occurrence word matrix.

Given a corpus, the co-occurrence matrix X denotes how often a particular word i occurs together with another word k. First, we compute a probability $P_{ik} = P(k|i)$ of how often word i and k appear together. This is done by dividing the number of times i and k appear together by the total number of times i appeared in the corpus. The same procedure is applied for calculating the probability $P_{jk} = P(k|j)$ of how often word j and word k appear together. Then, the ratio

$$F(\boldsymbol{x}_i, \boldsymbol{x}_j, \tilde{\boldsymbol{x}}_k) = \frac{P_{ik}}{P_{jk}} \tag{3.33}$$

is calculated, where $\tilde{\boldsymbol{x}} \in \mathbb{R}^d$ denotes a context word vector, $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ denote the word vectors for word i and j, and $F(\cdot)$ denotes the natural exponential function $e^{(\cdot)}$. Pennington et al., 2014 show that this can be simplified to

$$\boldsymbol{x}_i^T \tilde{\boldsymbol{x}}_k + b_i + \tilde{b}_k = \log(X_{ik})\,, \tag{3.34}$$

where b_i and b_k are bias terms associated with word vectors $\boldsymbol{x}_i^T$ and $\tilde{\boldsymbol{x}}_k$, respectively, and X_{ik} is the number of times $\boldsymbol{x}_i$ occurs together with word

[29] For example, skip-gram does not capture whether a stopword is actually a context word belonging to the target word or just a pronoun or determiner.

$\boldsymbol{x}_k$. Therefore, the loss can be written as

$$L = \sum_{i,j=1}^{V} f(X_{ij})(\boldsymbol{x}_i^T \tilde{\boldsymbol{x}}_j + b_i + \tilde{b}_j - \log X_{ij})^2 \,, \tag{3.35}$$

where X_{ij} is the number of times word j occurs with word i, b_i and $\tilde{b}_j$ are the bias terms, V is the vocabulary size and $f(X_{ij})$ is a weighting function.[30] The benefit of using GloVe embeddings is that the embeddings use global *and* local word vector statistics. Moreover, we can readily employ pre-trained GloVe embeddings, which were trained on a huge corpus of text. However, GloVe and skip-gram embeddings are static word embeddings, i.e., each word has a fixed embedding in the vocabulary. In the next subsection, we introduce contextualized embeddings, which assign the word vector depending on the context in which the word is used (Jurafsky and Martin, 2020, p. 112). In this way, the word "bank", in the sense of a money lending institution, and the word "bank", in the sense of a river bank, obtain two separate word embedding vectors.

3.5.2.4. BERT – Deep Bidirectional Transformers

Bidirectional Encoder Representations from Transformers (BERT) is a powerful tool developed by Devlin et al., 2019 at Google, which employs masked language models[31] and next sentence prediction tasks[32] to efficiently create deep bidirectional representations from unlabeled text using left and right context in all layers (Devlin et al., 2019). BERT is build on ideas such as semi-supervised sequence learning (Dai and Le, 2015), deep contextualized word representations (M. Peters et al., 2018),

[30] Pennington et al., 2014 assume a weighting function of the form $f(z) = \begin{cases} (z/z_{max}^{\alpha}) & \text{if } z < z_{max}, \\ 1 & \text{otherwise.} \end{cases}$

[31] Masked language models mask some part of the tokens, i.e., "hide" the token temporarily, and try to predict the masked token from the context (Devlin et al., 2019).

[32] Next sentence prediction predicts the next sentence by using pairs of sentences (Devlin et al., 2019)

transfer learning (Howard and Ruder, 2018), and decoder transformer models (Radford et al., 2018).

BERT is not only powerful but also a flexible tool that can be applied for question answering tasks, sentence pair classification, sentence tagging, and single sentence classification tasks (Devlin et al., 2019). Since we use BERT only for creating sentence-level embeddings, we look at BERT from a high-level perspective and refer the interested reader to the paper from Devlin et al., 2019 for more details.

In its "base" form ($\text{BERT}_{\text{BASE}}$), the model consists of 12 encoder blocks that are stacked on top of each other. Encoder blocks are used in transformer language models[33] to map an input sequence $(x_1, \cdots, x_n)$ to a sequence representation $z = (z_1, \cdots, z_n)$ by using self-attention (see Subsection 3.6.5 for more information on self-attention and encoder blocks). In this context, a sequence refers to a sequence of consecutive tokens. In the $\text{BERT}_{\text{BASE}}$ model, the encoder blocks take a token-sequence of up to 512-tokens and apply bidirectional self-attention in each layer. BERT's encoder structure is similar to Vaswani et al., 2017, however, BERT applies *bidirectional* self-attention. To avoid that a word indirectly sees itself during training, BERT masks 15% of the input sequence during training. This helps BERT to understand the context and the language itself, to create an efficient language model. BERT implementations are usually already pre-trained on a large corpus of textual data and, thus, can be readily applied to get contextualized embeddings of words and sentences. Furthermore, it can be fine-tuned for specific tasks.

3.5.3. LIME – Local Interpretable Model-agnostic Explanations

Text classification based on deep neural networks has the advantage of a high prediction performance (Nguyen, 2018). However, the increasingly complex model architectures of deep neural networks complicate the interpretability from a human perspective (Silva et al., 2018). Machine

[33] Transformer models were introduced by Vaswani et al., 2017 and traditionally apply multi-headed self-attention to encode a specific word. Transformers understand which words in a sequence of words are most important for the current word.

learning classifiers and, especially, deep learning classifiers have thus been criticised as black boxes that are hard to grasp (Shrikumar et al., 2017). Recently, several methods have been proposed to make the reasoning behind complex neural networks easier to understand and interpret (Lundberg and Lee, 2017; Shrikumar et al., 2017; Ribeiro et al., 2016; Datta et al., 2016, i.a.). However, it is not yet clear how the different methods relate and when a method should be preferred to another (Lundberg and Lee, 2017).

In this work, we use LIME, i.e., Local Interpretable Model-agnostic Explanations, to explain our model predictions (Ribeiro et al., 2016). LIME separates the model explanations from the underlying machine learning model. Thus, LIME treats the original model as a black box, i.e., LIME is model-agnostic. This increases flexibility as any complex machine learning model can be used and, at the same time, interpreted on a human level.

The intuition behind LIME is straightforward: For a single observation, LIME creates perturbed observations and generates predictions for these newly created observations using the original classifier. Then the newly created observations are used to train an interpretable model, like a regression or a random forest, which is weighted by a proximity measure that measures the distance to the original observation. Consequently, the model can then be interpreted from a human perspective. For a single text, this means that LIME creates a set of perturbed texts by removing words from the original text. The set of newly created texts are then used to create predictions using the original model as well as by training an exponential kernel on the cosine distance. The cosine distance simply measures the proximity of a newly created text to the original text in the Euclidean space. Mathematically, LIME can be expressed as (Ribeiro et al., 2016)

$$\xi(x) = \underset{g \in G}{\operatorname{argmin}} \, L(f, g, \pi_x) + \Omega(g) \,. \tag{3.36}$$

f is the original classifier that we try to explain using a single observation x. For this reason, LIME employs an interpretable explanation model

g that minimizes the loss L. For example, if we use a linear regression model as the explanation model, we would use the mean squared error as the loss function. G is the set of potentially interpretable explanation models. π is the proximity measure between the perturbed observation z and the original instance x. Thus, $L(f, g, \pi_x)$ measures the error – or "unfaithfulness" – of g in approximating f. The complexity of the explanation model g is expressed by $\Omega(g)$, e.g., the number of non-zero weights in a linear model. From this, LIME selects the m features that best describe the original classifiers output from the perturbed data.

While a learned explanation model approximates a model's predictions locally, this is not necessarily true globally. For this reason, Ribeiro et al., 2016 propose an algorithm called *submodular pick*, which selects observations that are representative for the entire dataset. These non-redundant explanations should help explain how the model behaves globally. The procedure for submodular pick is as follows: Submodular pick first calculates the lime values for a set of observations n, so that m features are selected for each observation. From this, we obtain an explanation matrix W that contains the explanations for each observation. Thus, the explanation matrix is of size $W \in \mathbb{R}^{n \times d}$ with n being the number of observations and d being the number of unique features found by LIME. Submodular pick then calculates the feature importance score as $I_j = \sqrt{\sum_{i=1}^{n} |W_{ij}|}$ and uses a greedy optimization algorithm to find the observations that best represent the underlying dataset (Ribeiro et al., 2016).

3.6. Network Architectures

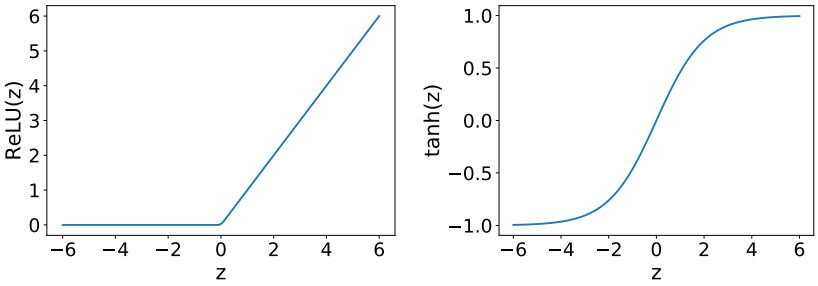
In this section, we introduce some common network elements. These network elements are used to build efficient machine learning or deep learning classifiers.

3.6.1. Activation Function

In Section 3.3, we introduced the logistic function as a common activation function. However, sometimes we do not want our activation to contain negative values. Therefore, another activation function used is the so called ReLU function, i.e., rectified linear unit. The ReLU function is defined as[34]

$$R(z) = \max(0, z)\,. \tag{3.37}$$

The ReLU function ranges from zero to infinity (see Fig 3.17a.).



(a) ReLU activation function. **(b)** tanh activation function.

Fig. 3.17.: Common activation functions.

Furthermore, a common activation function is the hyperbolic tangent function, i.e., tanh(·) function, which is defined as[35]

$$\tanh(z) = 2\phi(2z) - 1\,. \tag{3.38}$$

The tanh(·) function rescales the logistic function $\phi(\cdot)$ so that the output ranges from -1 to 1 (see Fig. 3.17b).

[34] Compare Goodfellow et al., 2017, p. 187.
[35] Compare Goodfellow et al., 2017, p. 189.

3.6.2. CNN – Convolutional Neural Networks

Convolutional Neural Networks (CNN) are typically applied in computer vision. However, recently they have also found applications in NLP (Kim, 2014; Kalchbrenner et al., 2014; Zhang and Wallace, 2017, i.a). While computer vision typically applies 2-dimensional convolutional layers, NLP focuses on one-dimensional convolutional layers. A one-dimensional CNNs commonly consist of a convolutional layer followed by a maxpooling layer – or a similar layer, which flattens a 3-dimensional output to a 2-dimensional output.

So to explain the functioning of a CNN, let us consider a set of documents. Each document D is represented as a $D \in \mathbb{R}^{n \times k}$ document matrix. n is the number of words in the longest document and k is the dimension of the word vectors. Consequently, each word in a document is represented by a k-dimensional vector $\boldsymbol{x}_i$ with i being the i-th word in a document. The words in each document are simply stacked together, i.e., concatenated, as (Kim, 2014)

$$\boldsymbol{x}_{1:n} = \boldsymbol{x}_1 \oplus \boldsymbol{x}_2 \oplus ... \oplus \boldsymbol{x}_n \,, \tag{3.39}$$

with $\oplus$ being the concatenation operator. However, in a set of documents, some documents contain more words than others. Thus, before applying CNNs, all documents need to be padded to the same length, that is, all documents need to be made equally long by filling them with zeros (Goodfellow et al., 2017, p. 338). In this case, the zeros are just placeholders, since the network needs an input of equal length to perform its operations. Therefore, all documents are padded to a length of n words, with the longest document consisting of a maximum of n words.

Now, the CNN applies over each subset of words $\{\boldsymbol{x}_{1:h}, \boldsymbol{x}_{2:h+1}, \cdots, \boldsymbol{x}_{n-h+1:n}\}$ a sliding window, also called filter, of size $k \times h$. This filter $W \in \mathbb{R}^{h \times k}$ produces a new feature $\boldsymbol{c}_i$ as (Kim, 2014)

$$\boldsymbol{c}_i = f(W\boldsymbol{x}_{i:i+h-1} + \boldsymbol{b}) \,, \tag{3.40}$$

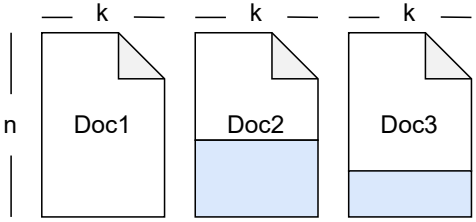


Fig. 3.18.: Documents of unequal length are padded to the same length, i.e., filled with vectors containing zeros.

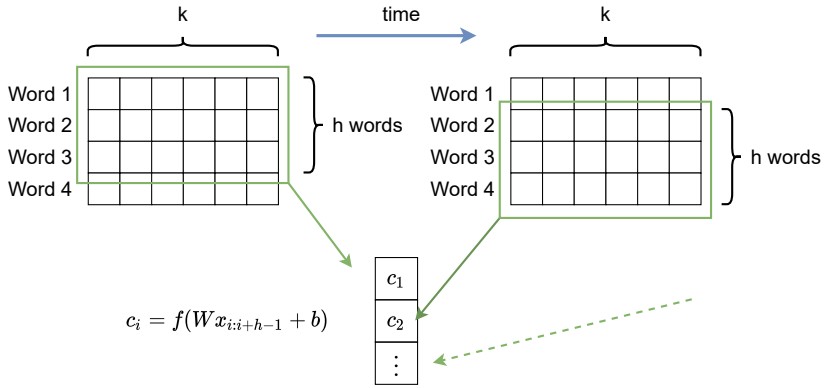


Fig. 3.19.: Word structure used for Convolutional Neural Networks.

with $f(\cdot)$ being a non-linear function, e.g., the ReLU function (Nair and Hinton, 2010; Glorot et al., 2011). By applying the filter over each possible subset of words, we produce a new *feature map* $\boldsymbol{c} = [c_1, c_2, \cdots, c_{n-h+1}]$. In the last step, a maxpooling layer (Goodfellow et al., 2017, p. 330) is applied to the newly produced feature map to get the maximum value $\hat{c} = \max(\boldsymbol{c})$ of each produced feature map.[36] Usually, multiple filters of different sizes are applied to a text to learn the relation between words

[36]Note that a maxpooling layer does exactly that: The maxpooling layer calculates the maximum value of the feature map.

that are further apart or closer together.[37]

3.6.3. GRU – Gated Recurrent Unit

Until now, we discussed only feedforward networks. Feedforward networks process the input data in a feedforward pass without considering the sequential structure of the input.[38] However, a sentence is made up of words, where the meaning of a sentence only becomes apparent, if we read the sentence from left to right. Each word by itself has a certain meaning to it but without the context, we can only loosely interpret the words. Similarly, a sentence's meaning might depend on a previous sentence's meaning so that the general interpretation of a document is only possible, if we consider the sequence of sentences. Consequently, text can be interpreted as sequential data.

Gated Recurrent Units (GRU) (K. Cho et al., 2014) belong to the category of recurrent neural networks (RNN). RNNs introduce a measure for time to keep track of the sequential structure of the underlying data by feeding their produced output back into themselves with a delay. In this sense, GRUs use gates to control for the information flow. GRUs employ an update gate and a reset gate, which decide what information should be passed to the output. In the first step, for each unit j, the update gate $\boldsymbol{z}_t$ computes (K. Cho et al., 2014)

$$\boldsymbol{z}_t = \phi(W^{(z)}\boldsymbol{x}_t + U^{(z)}\boldsymbol{h}_{t-1})\,, \tag{3.41}$$

where $W^{(z)}$ is the weight matrix for belonging to z and $U^{(z)}$ is the weight matrix for the lagged hidden layer output $\boldsymbol{h}_{t-1}$. A sigmoid function ϕ is applied to the result, to squash the result in the range of $[0, 1]$. The update gate is responsible for how much of the information from the past $\boldsymbol{h}_{t-1}$ is passed on. At the same time, the reset gate computes in a similar

[37] For more details see also Zhang and Wallace, 2017.
[38] Note: CNNs are also a type of feedforward networks, as they contain no memory.

fashion (K. Cho et al., 2014)

$$\boldsymbol{r}_t = \phi(W^{(r)}\boldsymbol{x}_t + U^{(r)}\boldsymbol{h}_{t-1})\,. \tag{3.42}$$

We see that equation 3.42 is equivalent to the update gate, where we take the corresponding weights, multiply them with the input $\boldsymbol{x}_t$ and the past information $\boldsymbol{h}_{t-1}$, and apply the sigmoid function to the sum. The reset gate is responsible for how much of the past information should be forgotten (e.g., if the reset gate is set to zero, the previous state will be forgotten (Yang et al., 2016)).

Then, we can compute the current memory as (Yang et al., 2016)

$$\tilde{\boldsymbol{h}}_t = \tanh(W^{(h)}\boldsymbol{x}_t + \boldsymbol{r}_t \odot (U^{(h)}\boldsymbol{h}_{t-1}))\,, \tag{3.43}$$

where $\odot$ is the hadamard product, i.e., the element-wise multiplication, $W^{(h)}$ and $U^{(h)}$ are again the corresponding weight matrices, and $\tanh(\cdot)$ is the hyperbolic tangent function.

The update gate $\boldsymbol{z}_t$ and the current memory $\tilde{\boldsymbol{h}}_t$ is then used to compute the current information $\boldsymbol{h}_t$ that is passed to the next unit (Yang et al., 2016):

$$\boldsymbol{h}_t = \boldsymbol{z}_t \odot \boldsymbol{h}_{t-1} + (1 - \boldsymbol{z}_t) \odot \tilde{\boldsymbol{h}}_t\,. \tag{3.44}$$

Thus, $\boldsymbol{h}_t$ is basically a linear interpolation between the current new state $\tilde{\boldsymbol{h}}_t$ and the previous state $\tilde{\boldsymbol{h}}_{t-1}$.

A GRU can work in both ways of the sequence. The direction is denoted by an arrow above $\boldsymbol{h}_t$. A bidirectional GRU works also from the left side of the sequence to the right $\overrightarrow{\boldsymbol{h}}_t$ (forward hidden state) as well as from right to left $\overleftarrow{\boldsymbol{h}}_t$ (backward hidden state) (Bahdanau et al., 2015).[39]

3.6.4. Attention Layer

Although GRUs can keep track of sequences by memorizing part of the content, some words in a text will be more meaningful than others to

[39] See also Chung et al., 2014 for more details on GRUs.

understand a document and, therefore, should receive more *attention*. To pay special attention to these informative words, we use **attention layers** (Bahdanau et al., 2015).

Following Yang et al., 2016, a document D_i with i denoting the i-th document consists of t words with $t \in [1, T]$. A previous bidirectional GRU on the sequence of words produces $\overrightarrow{\boldsymbol{h}}_t$ and $\overleftarrow{\boldsymbol{h}}_t$, which can be concatenated to $\boldsymbol{h}_t = [\overrightarrow{\boldsymbol{h}}_t, \overleftarrow{\boldsymbol{h}}_t]$. $\boldsymbol{h}_t$ contains the context of each word and is referred to as the *annotations* (Bahdanau et al., 2015). The annotations are passed through a one-layer network:

$$\boldsymbol{u}_t = \tanh(W_w \boldsymbol{h}_t + \boldsymbol{b}_w)\,. \tag{3.45}$$

The tanh function sets the input in the range of $[-1, 1]$. Now, to measure the importance of a word, each word is accompanied by a trainable context vector $\boldsymbol{u}_w$, which is randomly initialized. The importance of a word is then measured as the similarity of $\boldsymbol{u}_t$ with a word level context vector $\boldsymbol{u}_w$. The importance is normalized through a softmax function:

$$a_t = \frac{e^{\boldsymbol{u}_t^T \boldsymbol{u}_w}}{\sum_t e^{\boldsymbol{u}_t^T \boldsymbol{u}_w}}\,. \tag{3.46}$$

These normalized importance weights a_t are then summed and concatenated with the previously calculated context annotations:

$$\boldsymbol{s} = \sum_t a_t \boldsymbol{h}_t\,. \tag{3.47}$$

The context vector $\boldsymbol{s}$ is the weighted sum of the annotations $\boldsymbol{h}_t$, with the weights a_t. To classify documents, we can apply this attention mechanism on a word level basis, a sentence level basis, or both (Yang et al., 2016).

3.6.5. Self-Attention Layer

The previously explained attention layer needs a GRU layer to compute a context vector (see Subsection 3.6.4). However, Vaswani et al., 2017

showed that we can also use a process called self-attention to extract context information. Following Vaswani et al., 2017, let a word be represented by a k-dimensional vector $\boldsymbol{x}_i$ with i being the i-th word in a document. To consider the word order, we add a positional encoding vector $t \in \mathbb{R}^k$ to the word embeddings. Consequently, each document D is represented as a $D \in \mathbb{R}^{n \times k}$ document matrix. Now, for each word, we create a query vector, a key vector, and a value vector, which are inputs to the network but created during the training of the network. We do this as follows: First, we multiply the document matrix D with the weight matrices $W^Q, W^K, W^V \in \mathbb{R}^{k \times d}$, which are randomly initialized, to calculate the query matrix Q, the key matrix K, and the value matrix V. Then, we take the dot product of Q and K^T and divide it by the square root of the dimension of the key vectors d. Subsequently, we pass it through a softmax function and multiply the result with the value matrix:

$$\text{Attention}(Q, K, V) = Z = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V\,. \tag{3.48}$$

This process can be repeated for multiple query, key, and value matrices to compute multiple *attention heads* Z_i.[40] To obtain a single matrix for the multi-head self-attention, which can be passed to the next layer, we concatenate all these attention heads and multiply them by an additional weight matrix $W^O \in \mathbb{R}^{hd \times k}$. Running multiple attention heads allows the model to consider different representations. Usually, the single matrix for the multi-head self-attention is then normalized by adding the position encoded document matrix D to it. If the normalized matrix is then passed to a simple feedforward network and again normalized, we call this whole procedure a encoder block (see Figure 3.20). Encoder blocks are the underlying structure of the BERT algorithm (see Subsection 3.5.2.4).

[40] Since the weight matrices are randomly initialized the results will vary.

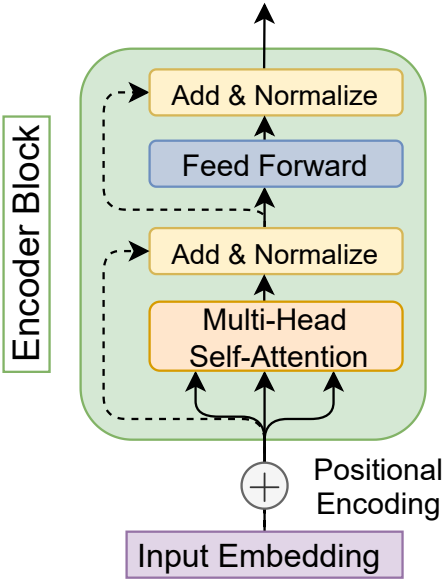


Fig. 3.20.: Encoder Block, drawn after Vaswani et al., 2017, p. 3.

3.7. Text Analysis in the Financial Domain

In the financial domain, particularly in financial accounting research, text analysis is an emerging field of research (Loughran and McDonald, 2016, p. 1188). Finance-related researchers apply mostly straightforward approaches and little research focuses on more sophisticated methods like machine learning and deep learning algorithms (Gentzkow et al., 2019). While NLP researchers working on financial data do apply various state-of-the-art methods, the NLP researchers mainly tend to focus on the short-term effect of narratives, to improve on trading (Armbrust et al., 2020, p. 182).

3.7.1. Sources for Text Analysis

In the financial domain, common sources for text analysis are corporate disclosures, earning calls, analyst reports, and news articles (Kearney and Liu, 2014, p. 172–173). While news articles predominantly cover events that have already happened, earning calls, analysts reports, and corporate disclosures offer a more forward-looking perspective. Thus, these forward-looking narratives have the potential to predict the future performance of the company (Kearney and Liu, 2014, p. 174). However, due to the low corporate disclosure release frequency, they are not perfectly suited for time-series modelling but more appropriate for cross-sectional analysis and event studies that examine the effect of the disclosure's sentiment on stock returns (Kearney and Liu, 2014, p. 174).

3.7.2. Readability

One common method to analyse textual data in the financial domain is readability, i.e., referring to how easy a text is to read and understand. Readability is often used to derive some ease-to-read score from the text and, subsequently, draw inferences from the score about the current or future financial performance (Baxamusa et al., 2018; Butler and Kešelj, 2009; Henry, 2008), earnings quality and persistence (Lo et al., 2017; Y.-J. Lee, 2012; F. Li, 2008), manager's risk taking behaviour (Chakrabarty et al., 2018), or fraud potential (Hoberg and Lewis, 2017; Othman et al., 2012; Humpherys et al., 2011).[41] Employing common readability measures like the length of texts, the Flesh Reading Ease Score (Flesch, 1948), or the Gunning-Fog Index, several studies linked poor readability to poor performance (Lehavy et al., 2011; Y.-J. Lee, 2012; Loughran and McDonald, 2009, 2014, i.a.). However, only few studies addressed the readability of non-financial disclosures. Nevertheless, there is evidence that non-financial disclosures are quite difficult to read compared to purely financial disclosures (Smeuninx et al., 2020; Farewell et al., 2014) and, similar to financial disclosures, the readability deteriorates as the financial

[41] Compare e.g., Fisher et al., 2016, p. 169

performance does (Abu Bakar and Ameer, 2011). However, readability measures do not address the actual content but the presentation of a text. While readability answers interesting questions with respect to the reader's understanding of a text and the accessibility of textual information, readability does not answer content related questions (Loughran and McDonald, 2016, p. 1188).

3.7.3. Financial Sentiment Analysis

The objective of sentiment analysis is to classify entire documents according to an overall positive or negative polarity (Pang et al., 2002). In the financial context, a positive polarity is associated with an increasing (bullish) market and a negative polarity with a decreasing (bearish) market (F. Xing et al., 2020, p. 978). However, sentiment analysis is a challenging task as there is often a lack in sufficient training data and it is difficult to obtain appropriate labels for texts, as it involves expert knowledge (F. Xing et al., 2020, p. 978). The sentiment, i.e., the views and opinions of market participants and commentators (Kearney and Liu, 2014, p. 172–173), is an important factor that drives markets as well as information flows (F. Z. Xing et al., 2018). Prominent examples of the sentiment analysis of 10-K and 10-Q filings include F. Li, 2010b, 2006; Feldman et al., 2008; Jegadeesh and Wu, 2013; Loughran and McDonald, 2011, i.a.

3.7.3.1. Word Lists and Sentiment Scores

To measure the polarity of financial documents, researchers apply word lists that usually contain a set of positively connotated words and a set of negatively connotated words. Researchers count the occurrence of positive and negative connotated words within a particular document to derive a sentiment score from the text. This score is then used to explain some financial measure like risk (e.g., measured by volatility) or return (e.g., measured by earnings or stock return). The financial variable used is then regressed on the score to draw some inferences about the statistical

significance. More recent methods also build word lists from a set of seed word embeddings, which are then used to propagate the sentiment for other word embeddings (Hamilton et al., 2016; Tai and Kao, 2013).

Common word lists are the General Inquirer (Kelly and Stone, 1975), the Henry Word List (Henry, 2008), and the Loughran & McDonald Word List (Loughran and McDonald, 2011) to extract the positive or negative sentiment. However, Loughran and McDonald, 2011 showed that terms in the financial context are typically not captured well by word lists developed for other disciplines. As noted by F. Li, 2010b, measuring the tone of the management and discussion section in 10-K reports using word lists like the General Inquirer or similar word counts does not relate positively to the future performance of a company. Thus, the results from F. Li, 2010b indicate that word lists do not work well on financial statements. This is due to the fact that words like "depreciation" or "liability" are neither positively or negatively connotated as it might be the case in other disciplines. Hence, Loughran and McDonald, 2011 developed a word list particularly for the financial context.

Also most studies addressing the non-financial information content apply lexical-based approaches (Kölbel et al., 2020, p. 8). For example, Loughran et al., 2009 consider ethical terms or terms associated with corporate responsibility to determine if these words appear more often in disclosures of stocks that produce alcohol, tobacco, or that are involved in gaming (sin stocks). Loughran et al., 2009 find that firms whose managers are discussing these terms are more likely to have low corporate governance and be sued the subsequent year following the filing. Similar research that used non-financial word lists are for example Feng and Gao, 2020, Verbeeten et al., 2016, and Matsumura et al., 2018.

3.7.3.2. What Word Lists Cannot Measure

Even so, most lexical-based methods remove stopwords from the underlying text (e.g., "and", "the", "not", etc.) and the drawback of this technique is that word order is not taken into account (F. Z. Xing et al.,

2018, p. 52). Since the context is ignored, word lists can turn a sentence's meaning. For example, "AstraZeneca acquired Alexion Pharmaceuticals" is very different from "Alexion Pharmaceuticals acquired AstraZeneca", with the stock responding likely very differently. Also semantic similarities will not be captured with simple word lists (F. Z. Xing et al., 2018, p. 52).

As in the financial domain dictionary-based approaches or similar straightforward approaches prevail (Frunza, 2020), for our research question, we will apply a classifier directly learned on the texts to figure out whether the texts are associated with positive or negative financial or environmental performance. In this way, we avoid deriving a sentiment score first. Thus, this thesis is not directly a sentiment analysis in the sense of deriving a sentiment score. Instead, we efficiently analyse whether a text is associated with the according label based on state-of-the-art NLP methods. Accordingly, we expect to find relevant features or patterns in the data that may allow conclusions about the financial and environmental performance and the interaction between them.

4. Methodology

In this chapter, we introduce the study design. We first review our hypotheses and introduce our general model to answer our research questions. Second, we explain how we preprocess the underlying financial disclosures and describe the underlying corpus. Then, we explain how we measure the financial and non-financial performance by showing the calculation of the according class labels. Next, we explain the methods that we use in order to obtain meaningful feature representations and the architecture of each classifier. Finally, we explain the subsequent evaluation of each classifier.

4.1. Review of the Hypotheses

The research objective of this dissertation is to determine whether the financial disclosures of a company, i.e., the MD&A section of 10-K and 10-Q reports, are associated with the company's future financial performance and environmental performance. Furthermore, as several studies indicate that the ESG performance contributes to the financial performance (Friede et al., 2015), we also examine whether the environmental performance contributes to the financial performance prediction task from text.

We focus solely on the environmental dimension of the ESG dimensions, since the 2010 SEC disclosure guideline (SEC, 2010) explicitly states that companies have to report on material issues related to greenhouse gas emissions and climate change. According to the guideline, environmental disclosures should be made within the "Description of business" (Item 101 of Regulation S-K), the "Legal proceedings" section (Item 103 of Regulation S-K), the "Risk factors" section (Item 503(c) of Regulation S-K), and the "Management's Discussion and Analysis of Financial Conditions and

Results of Operations" (MD&A) section (Item 303 of Regulation S-K). In this thesis, we focus on the MD&A section, since the MD&A section is arguably one of the most read sections by analysts (Clarkson et al., 1999; Tavcar, 1998; Rogers and Grant, 1997), the section provides a long- and short-term analysis from the management's point of view (Muslu et al., 2015; Griffin, 2003), and the section contains a forward-looking perspective (SEC, 2003). Thus, the MD&A section most likely provides insights on the future financial *and* environmental performance of a company. Additionally, the section might provide insights on the value that managers place on environmental performance, since the company has to report on environmental issues only if they are considered material (see Section 2.1.1). Consequently, we propose the following theories:[1]

H1a. The MD&A section of the 10-K and 10-Q filings contains information associated with the future financial performance of a company, i.e., the performance measured by earnings.

H1b. The MD&A section of the 10-K and 10-Q filings contains information associated with the future stock price performance of a company.

H2. The MD&A section of the 10-K and 10-Q filings contains information associated with the future environmental performance of a company.

H3. The narrative environmental performance contained within the MD&A section contributes to a company's financial performance.

4.2. Study Design

In order to test our hypotheses, we propose the following model (see Fig. 4.1): First, we predict the financial performance from the MD&A section, represented by the *Direct* arrow. Second, we predict the environmental performance from the text, i.e., arrow *Env*. Third, we use the environmental performance and the MD&A section to predict the financial performance, i.e., arrow *Combined*.

[1] Compare Chapter 2.2.

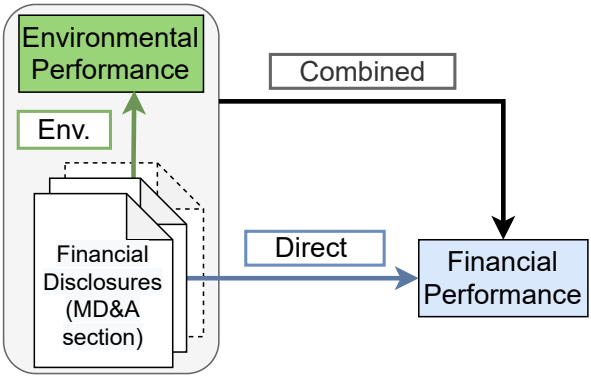


Fig. 4.1.: Illustration of the general idea to predict the financial performance from the MD&A section (Direct arrow), the environmental performance from the MD&A section (Env. arrow), and the financial performance using both the environmental performance and the MD&A section (Combined arrow).

Since we hypothesize that the MD&A section contains information to predict the company's performance measured by earnings (H1a.) and the performance measured by the stock price performance (H1b.), we build two separate models: (1) A *long-term* model that uses earnings as the financial performance metric; and (2) a *short-term* model that uses the stock price performance as the financial performance metric. We refer to model (1) as being long-term, since the company's earnings numbers are only released quarterly. With this model, we measure the general informativeness of the corporate filings, as the earnings metric shows how much information affecting the company's future financial performance is actually conveyed by the corporate filings. On the other hand, the company's stock price performance measures the informativeness of the filings to investors. As investors react to the filing's information by buying or selling the company's stock, the subsequent market price reaction approximates how relevant the information is to investors.[2] However,

[2]Compare Armbrust et al., 2020, p. 185.

for the short-term model (2), we consider two time horizons of different lengths. The first time horizon matches previous research (Loughran and McDonald, 2011; Jegadeesh and Wu, 2013, i.a.) and is equal to four days – we number this version of the short-term model (2a) –, and the second time horizon is equal to one month – we number this version of the model (2b). Hence, we create two separate *short-term* models, (2a) and (2b). For each model – (1), (2a), and (2b) –, we select an environmental performance metric that approximately corresponds to the time horizon of the financial performance metric. The environmental performance metric measures the informativeness of the filings with respect to the environmental performance. Thus, we consider in total three separate models, i.e., model (1), (2a), and (2b), each using different financial and environmental performance measures.

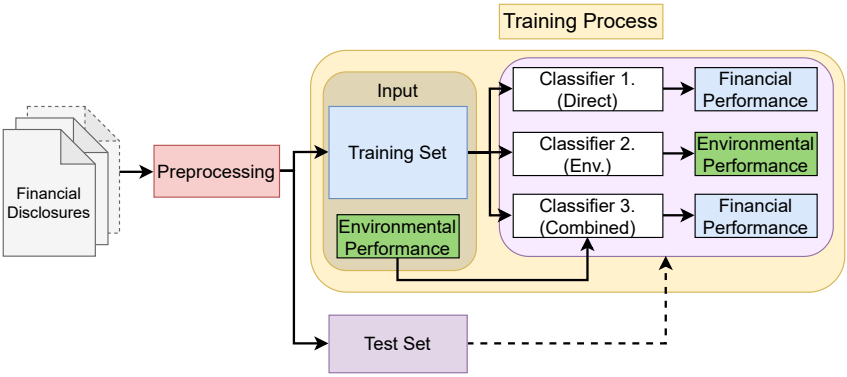


Fig. 4.2.: The financial disclosures are preprocessed and passed to our model, which predicts the environmental performance from the MD&A section (Classifier 1.), the financial performance from the MD&A section (Classifier 2.), and the financial performance using both the environmental performance and the MD&A section (Classifier 3.).

Now to perform the prediction tasks and to test our hypotheses, we employ a set of machine learning and deep learning classifiers. Figure 4.2 presents our model in more detail. First, the financial disclosures are preprocessed, so that we can pass them to the classifiers. Then, the whole

data set is split into a training set and a test set. We use the training set to train three different classifiers: (1) A classifier that predicts the future financial performance from the MD&A section, called *direct*-classifier; (2) a classifier that predicts the future environmental performance from the text, i.e., *env.*-classifier; and (3) a classifier that predicts the future financial performance using the text and the environmental performance, i.e., *combined*-classifier. Subsequently, we pass the test set to all classifiers to obtain the performance of each classifier, i.e., how well the classifiers perform on unseen observations.

In case the *direct*-classifier and the *env.*-classifier show a meaningful relation to their labels, we conclude that the MD&A section is informative with respect to the financial and environmental performance, respectively – that is, the MD&A section is associated with the financial and environmental performance, i.e., (H1a), (H1b), and (H2). If the *combined*-classifier shows a meaningful relation to the financial class label and is also superior to the *direct*-classifier, we conclude that the environmental performance complements the MD&A section in explaining the financial performance, i.e., environmental performance contributes to the financial performance prediction task from text, i.e., (H3).[3]

However, the information content of the MD&A section may be limited for mainly two reasons: (1) The company might publish information through other information channels in a more timely fashion (e.g., through press releases, ad-hoc messages, twitter posts, or sustainability reports); (2) companies make an earnings announcement approximately three weeks before the disclosure that includes information on their earnings, cash flows, and other key performance indicators (Muslu et al., 2015; Lansford, 2006; Schrand and Walther, 2000).

Although these limitations apply, the MD&A section enables companies to provide a more comprehensive reporting with respect to the

[3]Note that for (H3) we measure the narrative environmental information with the actual environmental performance. While we would have used the predicted environmental performance, the predicted environmental performance proves to be unreliable.

future prospects, the operating strategy, and any other matter that could materially impact the companies financial performance. On top, the dissemination is for free. Therefore, we assume that companies provide previously released information within their MD&A section in more detail including any material information from their sustainability reports. Addressing the second limitation: While companies typically make an earnings announcement a few days before the disclosure of the filing, the MD&A section contains more *qualitative* information by the management than preliminary earnings announcements usually do. This is due to the SEC's requirements that demand extensive qualitative information in the MD&A sections (Feldman et al., 2008, p. 14). Therefore, we expect that the majority of quantitative information does not contain much new information for investors but the qualitative, textual information content probably does. Hence, we focus entirely on the textual information content of the corporate filings and do not make use of other external sources, such as social media.[4]

4.3. Data and Text Preprocessing

Our corpus contains the 10-K and 10-Q filings of the S&P500 constituents for the five-year period from the beginning of 2014 to the end of 2018. We also include all companies that were constituents of the S&P500 at any time during our sample period to minimize survivorship bias. However, the corpus includes only those reports, for which we are able to calculate all class labels (see Section 4.4). We retrieve the 10-K and 10-Q filings from the Notre Dame Software Repository for Accounting and Finance (SRAF) website (McDonald, 2019; Loughran and McDonald, 2016). The filings on the SRAF are already pre-processed in such a way that they exclude any markup tags (i.e., $<, \cdots, >$), ASCII-encoded graphics, and tables.[5] Hence, the 10-K and 10-Q filings contain only the plain text.[6]

[4] For a confirmation of these assumptions see Armbrust et al., 2020.

[5] For a detailed pre-processing description see: https://sraf.nd.edu/data/stage-one-10-x-parse-data/ (accessed 25.03.2021)

[6] The unprocessed files are at https://www.sec.gov/Archives/edgar/Feed/

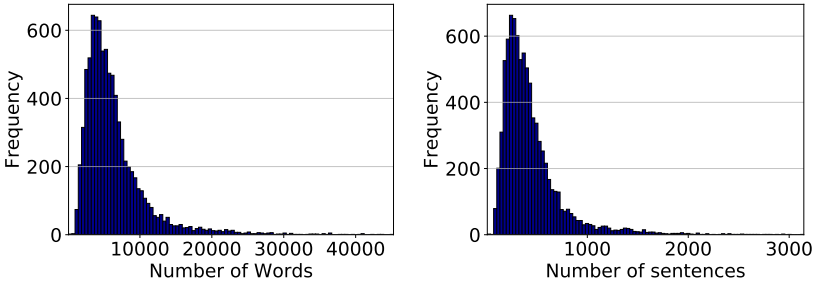
The files taken from the SRAF contain all sections of the filing. Therefore, we need to filter for the MD&A section. We decided to include the "Quantitative and Qualitative Disclosures About Market Risk" (QQD) section in addition, since most companies include the QQD section within the MD&A section. Although an automatic parsing process would have been beneficial due to speed and reproducibility, we opted for a manual parsing process to avoid downstream errors. Downstream errors could occur since some firms report sections titled item 7 (MD&A) and item 7A (QQD), but still include the *actual content* of these sections in a later part of the filing, e.g., in exhibit 13 or exhibit 99.1 – an additional attachment to the filings (Center, 2014).

The filings also contain the Standard Industrial Classification (SIC) code of each company. The first two digits of the SIC system classifies companies into 10 different industries: Agriculture, Forestry, Fishing (SIC 01-09), Mining (SIC 10-14), Construction (SIC 15-17), Manufacturing (SIC 20-39), Transportation & Public Utilities (SIC 40-49), Wholesale Trade (SIC 50-51), Retail Trade (SIC 52-59), Finance, Insurance, Real Estate (SIC 60-67), Services (SIC 70-89), Public Administration (SIC 91-99). We use these industry classifications to control the *env.*-classifiers for industry, as environmental performance has been shown to be influenced by the company's industry (Doyle, 2018, p. 6).

We use the SpaCy tokenizer (Honnibal et al., 2020) to split each filing into tokens. The final corpus includes 54.872.675 tokens in total (45.388 unique tokens) after removing stopwords using the NLTK stopword list (Bird et al., 2009) and keeping only alphabetic characters.[7] Figure 4.3 illustrates the distribution of words (see Fig. 4.3a) and sentences (see Fig. 4.3b) per filing. We lose a total of 2925 observations due to missing data. The final sample consists of 8592 observations (554 unique firms).[8]

[7] For the BERT classifiers this number varies, as BERT uses its own tokenizer (Devlin et al., 2019). However, for BERT we do not remove stopwords.

[8] Note that the data and text preprocessing is congruent to Armbrust et al., 2020. However, we use SpaCy (Honnibal et al., 2020) to sentenize the filings instead of NLTK (Bird et al., 2009).



(a) Distribution of the number of words per filing after removing stopwords and keeping only words with alphabetic characters.

(b) Distribution of sentences per filing.

Fig. 4.3.: Distribution of the number of words and number of sentences per filing in the final corpus.

4.4. Class Labels

For the *long-term* model (1) and the two *short-term* models (2a) and (2b), we use different financial and environmental performance class labels. In this section, we explain how we measure the financial and environmental performance for each of these models and show the calculation of each class labels.

4.4.1. Earnings

For the *long-term* model, we use earnings to measure the financial performance. We define the earnings class label as the change in earnings per share. We calculate the change in earnings as

$$\Delta \text{EPS}_i = \text{EPS}_{i,q+1} - \text{EPS}_{i,q}\,, \tag{4.1}$$

where $\text{EPS}_{i,q}$ are the earnings for company i in quarter q and $\text{EPS}_{i,q+1}$ are the earnings in the subsequent quarter $q+1$. The earning per share data is taken from Bloomberg (Field: IS_COMP_EPS_GAAP).[9] To account for dilutive effects, we employ the diluted earnings per share that are published with each filing and reported according to the Generally Accepted Accounting Principles. We binary encode the earnings class labels, i.e., converting positive ΔEPS_i to 1 and negative values to 0. This turns the classification task into a binary classification problem and increases the likelihood that the classifiers are able to predict the according performance.

4.4.2. Stock Price Performance

The company's stock price performance is measured by using an event study design. For each filing i, we measure the effect on the stock by calculating the returns in excess to the market. Therefore, we employ buy-and-hold returns to calculate the excess returns (Barber and Lyon, 1997, p. 344). The buy-and-hold returns (BHAR) are calculated as

$$\text{BHAR}_i = \prod_{t=T_1}^{T_2} [1 + R_{i,t}] - \prod_{t=T_1}^{T_2} [1 + R_{m,t}] \,, \tag{4.2}$$

where $R_{i,t}$ is the company's stock return on date t and $R_{m,t}$ is the contemporaneous return on the S&P500 index. T_1 is the start of the holding period and T_2 is the end of the holding period. The individual daily returns are calculated as (P_{t+1}/P_t), with P_{t+1} being the stock price on date $t+1$ and P_t being the stock price on date t. Although comparable event studies apply cumulative abnormal returns (McWilliams and Siegel, 1997; Warner and Kothari, 2006, i.a.), buy-and-hold returns have been shown to be particularly robust for longer holding periods (Barber and Lyon, 1997).

For the holding period, T_1 till T_2, we use two different event time spans.

[9] Note that our intend is not to measure an earnings surprise, but to analyse if financial disclosures are associated with the future earnings of a company.

Similar to Loughran and McDonald, 2011 and Jegadeesh and Wu, 2013, we base the first holding period on Griffin, 2003. Griffin, 2003, p. 447 observed an elevated return response up to day 3 after the corporate filing date. Hence, the first holding period spans from day 0 to day 3, i.e., the date of the disclosure till 3 days after the filing date. We denote this buy-and-hold-return as $\text{BHAR}_{i,\text{short}}$. For the second holding period, we employ a 30-day time horizon, from day 0 to day 30, to match the change in the monthly environmental label more closely (see Subsection 4.4.3). We denote this longer term buy-and-hold return as $\text{BHAR}_{i,\text{long}}$.

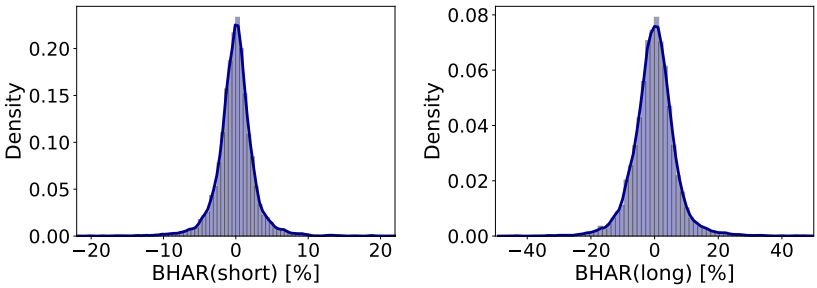
We use $\text{BHAR}_{i,\text{short}}$ for our *short-term* model (2a) and $\text{BHAR}_{i,\text{long}}$ for the *short-term* model (2b). However, since the underlying environmental data is only adjusted on a monthly basis (see Subsection 4.4.3), the $\text{BHAR}_{i,\text{short}}$ label leads to a small time mismatch for the *combined*-classifier. Nevertheless, we consider this mismatch to be negligible, since the idea is use the predicted environmental performance for the *combined*-classifier.

To test whether the buy-hold-returns are significantly different from zero for the sample of n firms, we conduct the t-test according to Barber and Lyon, 1997 as

$$\text{t}_{\text{BHAR}} = \frac{\overline{\text{BHAR}_{i,t}}}{\sigma(\text{BHAR}_{i,t})/\sqrt{n}}\,. \tag{4.3}$$

$\overline{\text{BHAR}_{i,t}}$ denotes the sample average and $\sigma(\text{BHAR}_{i,t})$ is the cross-sectional sample standard deviations of abnormal returns. Assuming the returns are independently and identically distributed, the t-test shows that the BHAR are significant to the 5% level for both $\text{BHAR}_{i,\text{short}}$ (t-value $= -1.3463$) and the 10% level for the $\text{BHAR}_{i,\text{long}}$ class labels (t-value $= -1.8482$).

Figure 4.4 plots the distribution of the buy-and-hold abnormal returns for the $\text{BHAR}_{i,\text{short}}$ class label and the $\text{BHAR}_{i,\text{long}}$ class label for the entire sample period. Equivalently to the earnings class label, we also binary encode the BHAR labels.



(a) Distribution of buy-and-hold returns for $\text{BHAR}_{i,\text{short}}$ from 2014 to 2018.

(b) Distribution of buy-and-hold returns for $\text{BHAR}_{i,\text{long}}$ from 2014 to 2018.

Fig. 4.4.: Distribution of filing period abnormal returns for the period from 2014 to 2018.

4.4.3. Environmental Performance

For the environmental performance class label, we employ ESG data from Sustainalytics as provided by Bloomberg. To avoid conceptual disagreement by combining different ESG rating providers, we decided to use just one ESG rating provider. We chose Sustainalytics due to availability and usage by investors, as Sustainalytics is one of the main big rating providers in the market (Kumar and Weiner, 2019). Sustainalytics relies on 50% qualitative as well as 50% quantitative aspects in constructing their rating (Eccles and Stroehle, 2018). This makes the rating score appropriate for our study design, as the rating relies not only on textual information but also considers numerical aspects. Hence, if a matter is reported textually, like in the MD&A section, it should be reflected in the environmental performance metric. Additionally, if a matter is only reported numerically, it should be also reflected in the rating score. Further, if a matter is not reported at all, the rating still reflects the environmental performance, since the metric does not only rely on corporate disclosures but also on external information and company surveys.

Although different results might be expected by using different rating providers, the most positive correlation of ESG rating data was observed among the environmental dimension (Gibson et al., 2021). Hence, we assume that differences will be small.

We measure the environmental performance by calculating the change in the environmental percentile score taken from Bloomberg.[10] Bloomberg writes that the top environmental performers are in the percentile 99%, i.e., they receive a score of 99, and the poor performers are in the 1%-percentile corresponding to a score of 1.[11] The environmental percentile data available on Bloomberg is adjusted on a monthly basis at the end of each month, e.g., 31.03.2016. However, the corporate filings are usually published at some point during the month, e.g., the 16.03.2016. Thus, we need to match the environmental data with the filing date. Therefore, we apply the following process: A report published at some point during the month, e.g., 16.03.2016, is assigned the environmental percentile score of the *previous* month, i.e., 28.02.2016 – provided that a rating exists. Subsequently, the according change in the environmental rating is calculated. This process ensures that no previously released information is already reflected in the environmental rating. Otherwise, a report published on the 16.03.2016 might influence the percentile score on the 31.03.2016 and the subsequent change in percentile score would be biased.[12]

The environmental class label for the *long-term* model (1) is calculated as

$$\Delta \text{Env}_{i,q} = \text{Env}_{i,q+1} - \text{Env}_{i,q}\,, \tag{4.4}$$

where $\text{Env}_{i,q}$ is the percentile score for company i in quarter q and $\text{Env}_{i,q+1}$ denotes the environmental rating of the next quarter. The environmental

[10] Bloomberg Field: SUSTAINALYTICS_ENVIRONMENT_PCT

[11] We like to note that the Sustainalytics ESG Risk Rating assesses the degree to which a company is at risk with respect to the ESG dimensions, providing *lower* percentile scores to companies with less unmanaged ESG risk. However, when asked, Bloomberg ensured that the Sustainalyitcs environmental percentile score assigns the best performing companies with respect to environmental performance in the highest percentile and the worst performing companies in the lowest percentile.

[12] Compare Armbrust et al., 2020, p. 194.

class label for both *short-term* models, (2a) and (2b), is calculated as

$$\Delta \text{Env}_{i,t} = \text{Env}_{i,t+1} - \text{Env}_{i,t}\,, \tag{4.5}$$

where $\text{Env}_{i,t}$ is the previously available percentile score for company i and $\text{Env}_{i,t+1}$ is the subsequent score.

For corporate filings published in February 2014 or March 2014, it is likely that no prior environmental percentile score is available. This leads to reports that are not considered, since no rating change can be calculated. The same applies to non-rated companies, where no environmental data is available. We binary encode the environmental class labels, i.e., converting positive $\Delta\text{Env}_{i,t}$ and $\Delta\text{Env}_{i,q}$ to 1 and negative values to 0, respectively.

4.5. Feature Extraction and Classifiers

Each model – the *long-term* model (1), the *short-term* model (2a), and the *short-term* model (2b) – consists of three classifiers: (1) A classifier predicting the financial class label, i.e., the *direct*-classifier; (2) a classifier predicting the environmental class label, i.e., the *env.*-classifier; and (3) a classifier taking into account the environmental class label and the text to predict the financial class label, i.e., the *combined*-classifier. We apply for each of these models five different feature extraction methods and classifier architectures. Consequently, we consider in total five separate *long-term* models, five separate *short-term* models (2a), and five separate *short-term* models (2b).

4.5.1. General Classifier Architectures

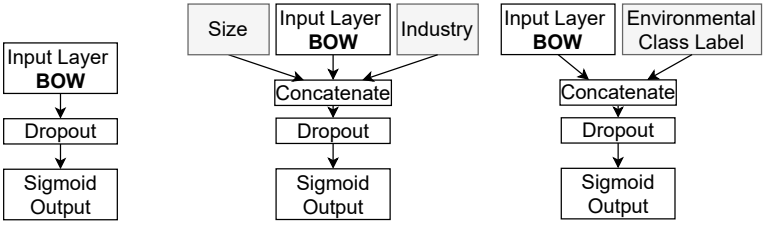
The five different feature extraction methods and classifier architectures that we use for each model are the following: (1) The first classifier architecture uses a simple Bag-of-Words method followed by drop-out and a sigmoid output layer. We call this classifier architecture **BOW**; (2) The second classifier architecture uses Bag-of-Words weighted with tf-idf followed by a dense layer with drop-out and a subsequent sigmoid output

layer. We refer to this architecture as **TF-IDF**; (3) The third classifier architecture employs pre-trained GloVe embeddings (Pennington et al., 2014) followed by a dense layer with drop-out and subsequent sigmoid output layer. We call this architecture **GloVe**; (4) The fourth classifier architecture follows the CNN architecture by Zhang and Wallace, 2017. This classifier architecture employs pre-trained GloVe embeddings with three convolutional layers, each followed by a max-pooling layer with drop-out. The resulting features are then concatenated and followed by a sigmoid output layer. The three convolutional layers have filter sizes of 2, 3, and 4, that is, the number of words that are considered together, and we apply 2 filters for each length. The size of each max-pooling window is set to 32. We refer to this architecture as **CNN**; (5) The last classifier architecture employs BERT sentence-embeddings followed by a Bidirectional-GRU, a ReLU layer, an attention layer with drop-out, and a fully-connected sigmoid output layer. The GRU are of size 50 and the attention layer size is 100. We refer to this architecture as **BERT**.

We control the *env.*-classifiers for size, since companies with a greater market capitalization tend to obtain higher ESG-ratings (Doyle, 2018, p. 9). For the company's size, we use the logarithm of the market capitalization. In addition, we use the SIC codes to control the *env.*-classifiers for industry. We one-hot-encoded the industry classification and pass it to the final sigmoid output layer. Hence, the *env.*-classifiers include a concatenate layer before the sigmoid layer to control for size and industry.

Figure 4.5 illustrates the network architectures for the **BOW** *direct*-, *env.*-, and *combined*-classifiers. Figure 4.6 shows the same for the **TF-IDF** architectures. Figure 4.7 illustrates how the GloVe embeddings are passed to the **GloVe** network. The **GloVe** architecture includes a "Flatten" layer, which reshapes the previous output of the drop-out so that it is equal to the number of elements – not including the batch size.

Similarly, the **CNN** takes GloVe embeddings as the input and passes them to the three convolutional layers, i.e., "ConvD1". Figure 4.8 illustrated the network architectures for **CNN**. Both the **GloVe** architecture

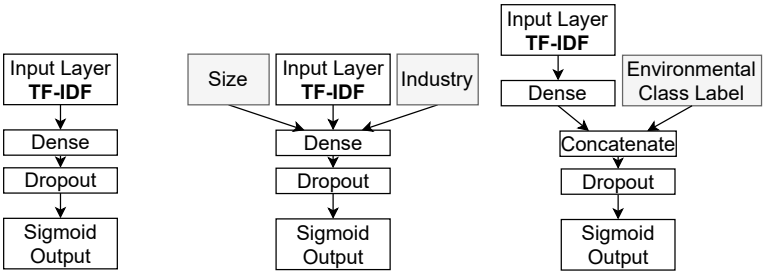


(a) *direct*-classifier, **BOW**.

(b) *env.*-classifier, **BOW**.

(c) *combined*-classifier, **BOW**.

Fig. 4.5.: Classifier architectures for **BOW**.



(a) *direct*-classifier, **TF-IDF**.

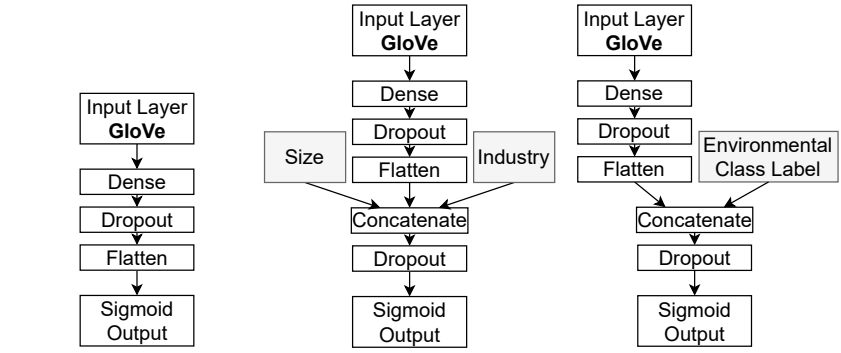
(b) *env.*-classifier, **TF-IDF**.

(c) *combined*-classifier, **TF-IDF**.

Fig. 4.6.: Classifier architectures for **TF-IDF**.

as well as the **CNN** architecture have a restriction on the maximum sequence length passed to the network, i.e., number of words per filing. We set the maximum sequence length to 20,000 so that filings with more words are cut at the 20,000-ths token. Since this affects only 2.2 percent of the filings, we assume that additional tokens would not add any information to the classification task. Shorter filings are padded to the same length.

For the **BERT** network architecture, we use pre-trained BERT to embed each sentence in a document. First, we split each document, i.e.,

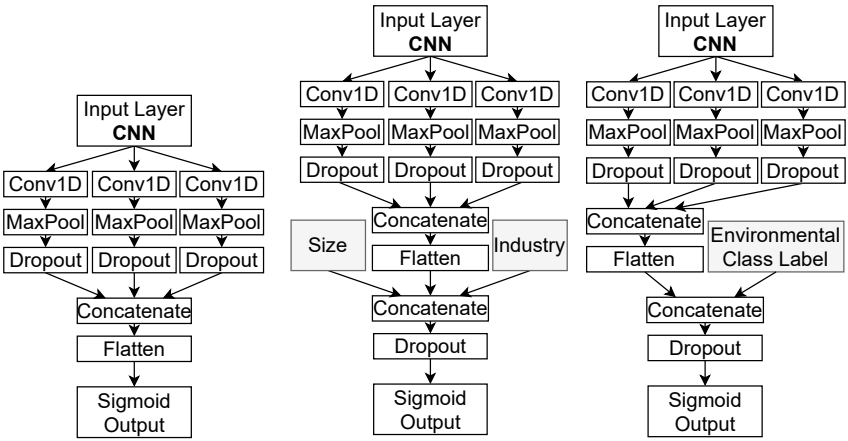


(a) *direct*-classifier, **GloVe**.

(b) *env.*-classifier, **GloVe**.

(c) *combined*-classifier, **GloVe**.

Fig. 4.7.: Classifier architectures for **GloVe**.



(a) *direct*-**CNN**-classifier.

(b) *env.*-**CNN**-classifier.

(c) *combined*-**CNN**-classifier.

Fig. 4.8.: Classifier architectures for **CNN**.

the MD&A section, into sentences using SpaCy (Honnibal et al., 2020) by considering a maximum of 1,000 sentences per document. Then, the

documents are padded to the same length. For each sentence, $\text{BERT}_{\text{BASE}}$ takes a sequence of maximum 512 tokens.[13] Therefore, sentences that are longer than 512 tokens are cut at the 510-token. The BERT tokenizer adds a [CLS]-token at the beginning and a [SEP]-token at the end of each sentence. The tokens are then converted to IDs and passed to BERT in a feedforward pass. BERT turns each token into a 1×768 hidden unit vector (see Fig. 4.9). Since these are contextualized embeddings, i.e., they contain also the context for each vector, the first [CLS]-hidden unit vector $h_{[CLS],s}$ for sentence s can be considered a representation of the sentence (Adhikari et al., 2020; Mulyar et al., 2019; Alammar, 2019, i.a.).

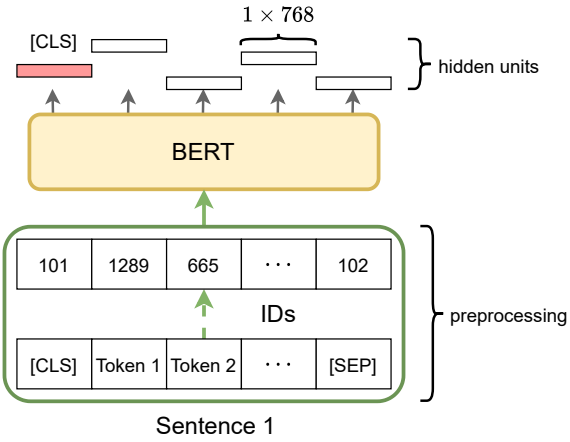


Fig. 4.9.: Exemplary depiction of how BERT processes tokens to produce contextualized embeddings, i.e., hidden units.

A document D consisting of n sentences is passed to BERT and n hidden unit vectors are produced. The hidden unit vectors are then passed to a bidirectional GRU with a consecutive attention layer to consider the sentence order for the document classification task (see Fig. 4.10). Figure 4.11 illustrates the classifier architectures for **BERT**.

[13]This includes the [CLS]-token and the [SEP]-token.

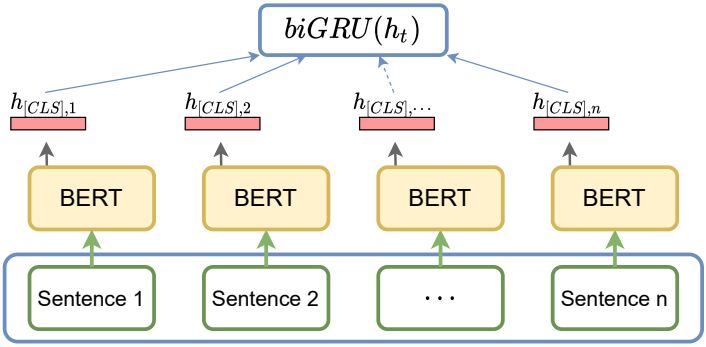
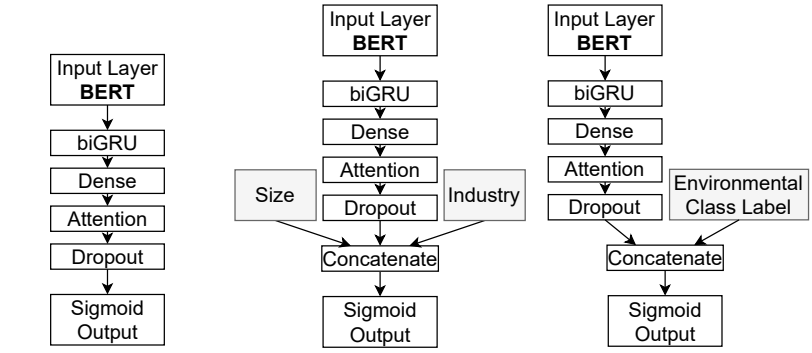


Fig. 4.10.: BERT produces contextualized embeddings for sentences, which are passed to a bidirectional GRU (biGRU).

For the classifier architectures **BOW**, **TF-IDF**, **GloVe** and **CNN**, we remove stop words using the NLTK stopword list (Bird et al., 2009), non-alphabet characters, and words that are shorter than two characters. However, we do not remove any words or characters for **BERT**, since BERT can make use of the grammatical structures. The models are fit using 50 epochs and a batch size of 32. For all models, we use the Adam optimizer (Kingma and Ba, 2015) and the binary cross-entropy as a loss function. For Adam, we use the default parameters as provided in the original paper by Kingma and Ba, 2015.

For selecting the optimal hyperparameter, we split the training set into a validation set and a training. We employ stratified k-fold to ensure that the training, test, and validation sets have approximately the same proportion of classes as the whole training set (see Subsection 3.4.6.1). Generally, we split the data into 10% test, 10% validation, and 80% training data. For the hyperparameter tuning, we set the drop-out probabilities to a probability of {50%, 60%} for each model. For the TF-IDF, GloVe, and CNN architectures, we also applied l1-regularization with regularization weights of {1, 0.1, 0.01}. The optimal drop-out probabilities are 50% for each classifier. The reported TF-IDF models apply 1-weighted l1-regularization, the GloVe models apply 0.1-weighted l1-regularization,



(a) *direct*-classifier, **BERT**.

(b) *env.*-classifier, **BERT**.

(c) *combined*-classifier, **BERT**.

Fig. 4.11.: Classifier architectures for **BERT**.

and the CNN models apply 1-weighted l1-regularization. Both the GloVe model and CNN model use GloVe word-embeddings of size 100. For all models, the dense layers are of size 100.

To implement the classifier architectures, we use the Python library Keras (Chollet et al., 2015) with tensorflow backend (Martin Abadi et al., 2015). Additionally, we leverage the classifiers by using the Scikit-learn library (Pedregosa et al., 2011). For **BERT**, we employ the Hugging Face transformer library (Wolf et al., 2019).[14] For our analysis, we select the epoch of the best performing model on the validation data.

4.6. Model Evaluation

4.6.1. Hypothesis Testing

We assume that a classifier captures the meaning of the MD&A section with respect to the according class label, if the classifier successfully predicts the according class label, that is, the classifier shows a high

[14] Our python code and the labelled data with the corresponding filing' names is available at: https://github.com/ForgeFin/DeepSustainableFinance

F_1-score. Therefore, if the *direct*-classifier shows a high F_1-score on the test set, this is an indication that the MD&A section contains meaningful information with respect to the financial performance, i.e., the MD&A section is associated with the company's financial performance – see (H1a) and (H1b). If the *env.*-classifier has a high F_1-score this indicates that the MD&A section contains meaningful information with respect to the environmental performance, i.e., the MD&A section is associated with the environmental performance of the company (H2). However, if a classifier does not capture the meaning of the MD&A section, we cannot reject the hypothesis that the MD&A section has no meaningful information content as the result could be due to the low power of the classifier. Nevertheless, a result like this might indicate that this is eventually the case.

To evaluate whether a classifier shows a sufficiently high F_1-score, we compare the classifiers performance to a majority baseline. The majority baseline is a classifier that always predicts class 1, namely an improvement in the financial and environmental class label, respectively. If a classifier has a substantially higher F_1-score than the majority baseline, we are able to predict financial and environmental improvements more precisely compared to an algorithm that always "guesses" an improvement. If the *combined*-classifier – considering both environmental performance and the company's disclosure – has a significantly greater F_1-score than the *direct*-classifier, we conclude that the environmental performance contributes to the relationship between the disclosure and financial performance (H3). We use a bootstrap approach to compare the classifiers to the majority baseline and to compare the *direct*-classifier with the *combined*-classifier (see Section 3.4.5).[15] We use the Bonferroni correction to adjust the significance level for the number of different classifier architectures. Thus, the adjusted significance level is $\alpha_{adj} = \frac{\alpha}{n}$ with n being the number of classifiers, i.e., $n = 5$. The Bonferroni-correction has the advantage to be more robust, when comparing multiple classifiers.

[15] Compare Armbrust et al., 2020.

4.6.2. LIME – Evaluation

In Subsection 3.5.3, we introduced the LIME values and how the global importance is used for the submodular pick algorithm. We employ the LIME values to evaluate the features learned by the different classifiers, i.e., what words the classifiers regard as important for the classification task. However, we do not employ LIME for **BERT**, since the **BERT** classifier architecture uses sentence embeddings instead of words. To calculate the LIME values, we select random documents from our corpus' test set. For the *env.*-classifiers, we hold the company's size and industry classification constant when calculating the LIME values. We also hold the environmental class label constant when calculating the LIME values for the *combined*-classifiers. We present the LIME values in such a way that positive LIME values show the contribution to the prediction probability as class 1, representing an improvement in the financial or environmental performance, and negative LIME values show the contribution to the prediction probability as class 0, i.e., a subsequent decline in financial or environmental performance.

However, the LIME values explain the features only locally. From the submodular pick algorithm, we know that I_j is the global importance of the feature j in the explanation space (Ribeiro et al., 2016, p. 1139). Thus, we use the importance score I_j to obtain a ranking system for the overall feature importance. Generally, the higher the importance score, the more important the feature. For text applications, Ribeiro et al., 2016 suggest calculating $I_j = \sqrt{\sum_{i=1}^{n} |W_{ij}|}$. We use this metric to calculate the overall feature importance. We assume that 10 percent of the test data is representative of the entire test set. Hence, we use 10 percent of the test data to calculate the global feature importance for the best classifier on the validation data. Generally, the importance scores should approximately reflect how the features affect the classifier globally.

5. Empirical Results

In this chapter, we present our empirical results. First, we give a general overview of the overall results, before we discuss the results of the *long-term* model and the two *short-term* models in more detail. For each model, we present the results of the *direct*-classifier, the *env.*-classifier, and the *combined*-classifier considering the **BOW**, **TF-IDF**, **GloVe**, **CNN**, and **BERT** network architecture.

5.1. Overview Results

Table 5.1 presents the overall precision, recall, and F_1-score on the test set for each model as well as the majority baseline, i.e., baseline. We always use the epoch of the best performing classifier on the validation data to show how the classifier performed on the test set. Note that for the *short-term* models the results of the *env.*-classifier are identical, since both *short-term* models share the same environmental class label. For all classifiers, the majority baseline has a recall equal to 1, since the majority baseline always predicts an improvement, i.e., $\hat{y} = 1$. Hence, the majority baseline creates no false negatives and, consequently, recall is equal to 1.[1]

From Table 5.1, we see that no classifier is superior to the corresponding majority baseline. For example, considering the *direct*-classifier for the *long-term* model (1), we see that the majority baseline achieves a F_1-score of 0.691. However, the **BOW** (.653), **TF-IDF** (.596), **GloVe** (.691), **CNN** (.616), and **BERT** (.689) *direct*-classifiers have all a F_1-score that is lower than the majority baseline. This indicates that a classifier, which always predicts an improvement in the financial performance, as measured by earnings, is superior to a trained classifier. The same applies

[1] See Subsection 3.4.4.

		long-term (1)			*short-term* (2a)			*short-term* (2b)		
	Classifier	P	R	F_1	P	R	F_1	P	R	F_1
Baseline	*direct*	.528	1	.691	.490	1	.657	.490	1	.657
	env.	.412	1	.583	.335	1	.502	.335	1	.502
	combined	.528	1	.691	.490	1	.657	.490	1	.657
BOW	*direct*	.529	.855	.653	.475	.770	.587	.503	.798	.617
	env.	.425	.760	.545	.358	.472	.407	.358	.472	.407
	combined	.523	.890	.659	.487	.786	.601	.498	.762	.603
TF-IDF	*direct*	.527	.685	.596	.478	.684	.563	.501	.689	.580
	env.	.450	.602	.515	.337	.427	.377	.337	.427	.377
	combined	.533	.773	.631	.477	.596	.530	.506	.608	.552
GloVe	*direct*	.528	1	.691	.500	.850	.630	.495	.893	.637
	env.	.424	.785	.550	.342	.885	.494	.342	.885	.494
	combined	.529	.996	.691	.490	1	.657	.489	.983	.654
CNN	*direct*	.539	.718	.616	.484	.594	.534	.525	.808	.636
	env.	.403	.511	.451	.364	.153	.215	.364	.153	.215
	combined	.530	.879	.661	.469	.637	.540	.491	.682	.571
BERT	*direct*	.533	.973	.689	.490	1	.657	.488	.980	.652
	env.	.432	.578	.494	.315	.212	.253	.315	.212	.253
	combined	.528	1	.691	.484	.711	.576	.478	.686	.564

Table 5.1.: Model results for the *direct*-classifier, the *env.*-classifier, and the *combined*-classifier on the test set with precision, recall, and F_1-score, respectively. Baseline is the majority baseline. The *long-term* model (1) uses the binary encoded ΔEPS_i class label for the *direct*- and *combined*-classifier, and the binary encoded $\Delta\text{Env}_{i,q}$ label for the *env.*-classifier. *Short-term* model (2a) uses the binary encoded $\text{BHAR}_{i,\text{short}}$ label and *short-term* model (2b) uses the binary encoded $\text{BHAR}_{i,\text{long}}$ label for the *direct*- and *combined*-classifier. Both *short-term* models use a binary encoded $\Delta\text{Env}_{i,t}$ label for the *env.*-classifier.

for the *short-term* model (2a) and (2b): No *direct*-classifier is superior to the majority baseline. Thus, we cannot reject the hypothesis that the MD&A section has no meaningful information content with respect to the financial information, that is, the corresponding null hypothesis to (H1a) and (H1b).

Table 5.1 shows similar results when comparing the *env.*-classifiers and the *combined*-classifiers to the corresponding majority baseline, i.e., no trained classifier is superior to the corresponding majority baseline. Consequently, we also cannot reject the hypothesis that the MD&A section has no meaningful information content with respect to the environmental information, that is, the null hypothesis to (H2). Accordingly, the results also indicate that there might be no support for (H3). However, these results could be due to the low performance of the classifiers, meaning that the classifiers are not able to capture enough meaningful information content within the corporate filings, *although* the information might be present. In the following sections, we thus analyse the *long-term* model and the two *short-term* models in more detail.[2]

5.2. Long-term model (1) – Empirical Results

In this section, we take a closer look at the results of the *long-term* model (1). The *long-term* model (1) uses the binary encoded earnings class label for the *direct*- and *combined*-classifier, and the binary encoded quarterly change in the environmental label for the *env.*-classifier.

Table 5.2 shows the comparison of the different classifier network architectures.[3] While the **GloVe** *direct*-classifier shows a F_1-score equal to the majority baseline, the precision and recall indicate that the **GloVe** classifier that performed best on the validation set is indeed a classifier that always predicts an improvement. All other *direct*-classifiers perform

[2]Note that we do not apply a bootstrap test when comparing the classifiers to the corresponding majority baseline, since the F_1-scores of each trained classifier are lower than the majority baseline.

[3]Note that these results are consistent with Table 5.1 but only the results for the *long-term* model (1) are shown.

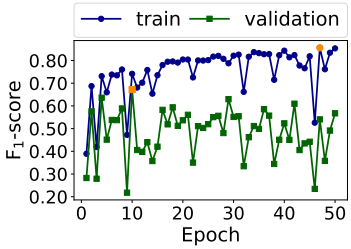
worse than the majority baseline. Hence, we find no evidence that the MD&A section contains information associated with the company's future financial performance, measured in terms of earnings.

	long-term (1) model								
	direct-classifier			*env.*-classifier			*combined*-classifier		
	P	R	F_1	P	R	F_1	P	R	F_1
Baseline	.528	1	.691	.412	1	.583	.528	1	.691
BOW	.529	.855	.653	.425	.760	.545	.523	.890	.659
TF-IDF	.527	.685	.596	.450	.602	.515	.533	.773	.631
GloVe	.528	1	.691	.424	.785	.550	.529	.996	.691
CNN	.539	.718	.616	.403	.511	.451	.530	.879	.661
BERT	.533	.973	.689	.432	.578	.494	.528	1	.691

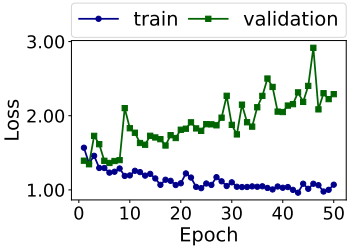
Table 5.2.: *Long-term* model (1) results for the *direct*-classifier, the *env.*-classifier, and the *combined*-classifier on the test set with precision, recall, and F_1-score, respectively. Baseline is the majority baseline. The *long-term* model (1) uses the binary encoded ΔEPS_i class label for the *direct*- and *combined*-classifier, and the binary encoded $\Delta\text{Env}_{i,q}$ label for the *env.*-classifier.

Similarly, Table 5.2 shows that the best F_1-score of a trained *env.*-classifier is achieved by **GloVe** (.550). However, no trained *env.*-classifier is able to surpass the majority baseline. Consequently, we do not find evidence that the MD&A section contains information associated with the company's future environmental performance. Likewise, the results show that no trained *combined*-classifier is superior to the majority baseline, with **BERT** having identical results to the majority baseline and **GloVe** being quite close to the majority baseline. This indicates that we might not be able to reject the hypothesis that the narrative environmental performance contributes to the financial prediction task from text. Therefore, we want to take a closer look at the learning process of each classifier.

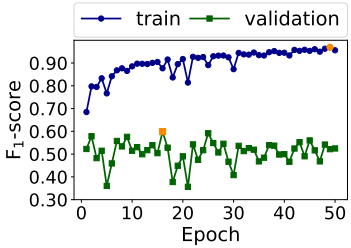
For the *long-term* model (1), we exemplary present the learning curve



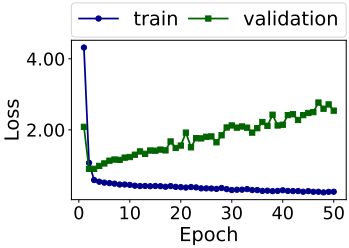
(a) F_1-score, **BOW** architecture *direct*-classifier – *long-term* model (1).



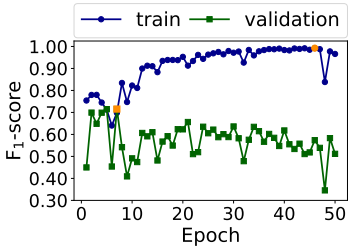
(b) Loss, **BOW** architecture *direct*-classifier – *long-term* model (1).



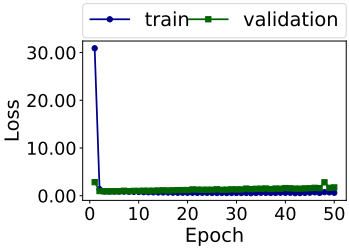
(c) F_1-score, **TF-IDF** architecture *direct*-classifier – *long-term* model (1).



(d) Loss, **TF-IDF** architecture *direct*-classifier – *long-term* model (1).

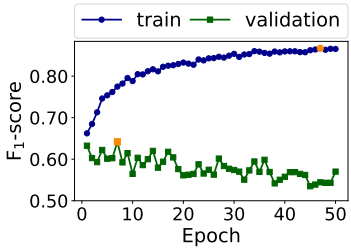


(e) F_1-score, **GloVe** architecture *direct*-classifier – *long-term* model (1).

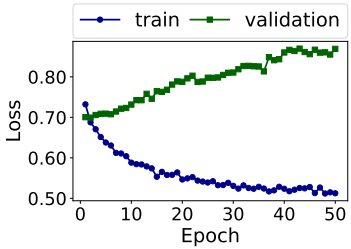


(f) Loss, **GloVe** architecture *direct*-classifier – *long-term* model (1).

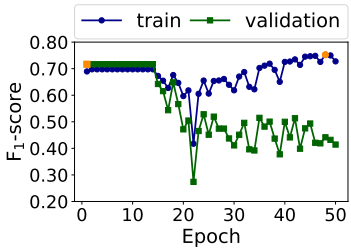
Fig. 5.1.: F_1-score and loss curve for the *direct*-classifiers – *long-term* model (1).



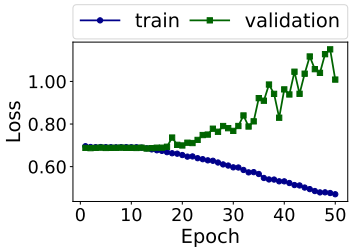
(g) F_1-score, **CNN** architecture *direct*-classifier – *long-term* model (1).



(h) Loss, **CNN** architecture *direct*-classifier – *long-term* model (1).



(i) F_1-score, **BERT** architecture *direct*-classifier – *long-term* model (1).



(j) Loss, **BERT** architecture *direct*-classifier – *long-term* model (1).

Fig. 5.1.: F_1-score and loss curve for the *direct*-classifiers – *long-term* model (1).

for the *direct*-classifier (see Fig. 5.1), the *env.*-classifier (see Fig. 5.2), and the *combined*-classifier (see Fig. 5.3). Similar results can be observed for the *short-term* model (2a) and *short-term* model (2b). For this reason, the F_1-score and loss curve for the *short-term* model (2a) and *short-term* model (2b) can be found in the appendix.

Fig. 5.1 presents the F_1-score and loss curve for the *direct*-classifiers of the *long-term* model (1). We see that the F_1-score curve on the training data (blue curve) increases for the **BOW** classifier (Fig. 5.1a), the **TF-IDF** classifier (Fig. 5.1c), the **GloVe** classifier (Fig. 5.1e), the **CNN**

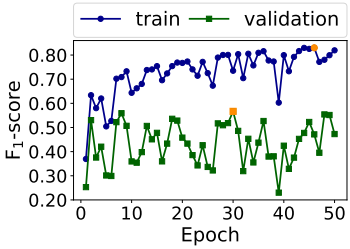
classifier (Fig. 5.1g), and the **BERT** classifier (Fig. 5.1i). For all these classifiers, the best trainings F_1-score (marked in orange on the blue curve) is achieved after the 45-ths epoch. However, the F_1-score curve on the validation data (green curve) mostly oscillates randomly, with a decreasing tendency. For the **GloVe**, **CNN**, and **BERT** classifier the decline of the validation F_1-score curve is more evident. For all classifiers, the best F_1-score on the validation data is achieved before the 17-ths epoch (marked in orange on the green curve). For example, the best result on the validation data for the **TF-IDF** *direct*-classifier is achieved at the 16-ths epoch (see Fig. 5.1c).[4]

At the same time, we see that the loss on the validation data is increasing for all classifiers, while the loss on the training data decreases (see Fig. 5.1b, Fig. 5.1d, Fig. 5.1f, Fig. 5.1h, and Fig. 5.1j). This means that the *direct*-classifiers indeed learn from the training data, but *what* the classifiers learn cannot be transferred to the validation data, i.e., the classifiers are not generalisable. Consequently, the classifiers overfit on the training data.

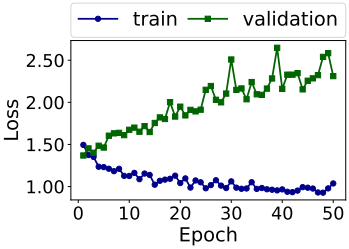
Looking at Figure 5.2, we see the F_1-score and loss curves for the *long-term* model (1) *env.*-classifiers. Again, while the F_1-score curve on the training data increases, the F_1-score curve on the validation data does not improve substantially. However, the F_1-score curves move slightly synchronous compared to the *long-term* model (1) *direct* classifiers. For example, for the **BERT** *env.*-classifier (see Fig. 5.2i), we see that initially the F_1-score curve on the training and validation data moves synchronous.

However, after the 22-ths epoch, the classifier does not achieve any improvements on the validation data. When looking at the corresponding loss curve (see Fig. 5.2j), we see that the loss on the validation data is increasing rapidly after the 22-ths epoch. The same can be observed for the other classifier architectures. Shortly after the loss on the training data is decreasing, we observe an increase in the loss for the validation

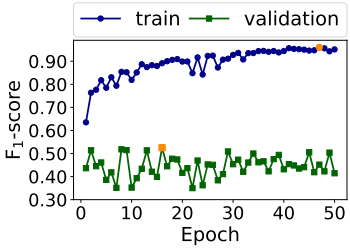
[4]Note: As previously explained, we use this best validation classifier on the test data, e.g., for the **TF-IDF** *direct*-classifier we use the trained model from the 16-ths epoch.



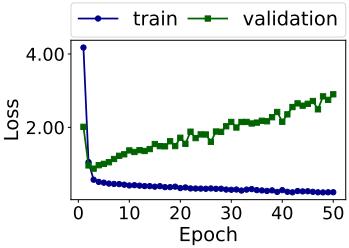
(a) F_1-score, **BOW** architecture *env.*-classifier – *long-term* model (1).



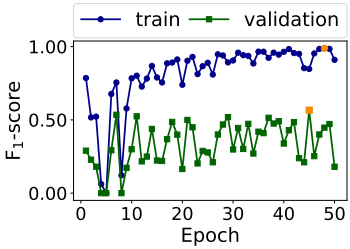
(b) Loss, **BOW** architecture *env.*-classifier – *long-term* model (1).



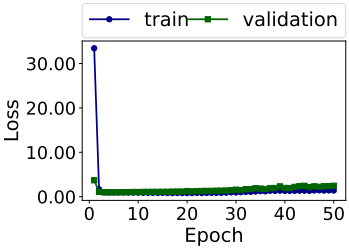
(c) F_1-score, **TF-IDF** architecture *env.*-classifier – *long-term* model (1).



(d) Loss, **TF-IDF** architecture *env.*-classifier – *long-term* model (1).

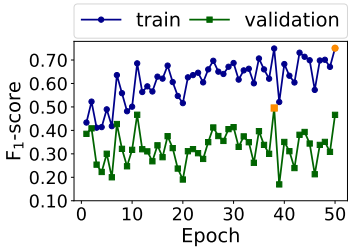


(e) F_1-score, **GloVe** architecture *env.*-classifier – *long-term* model (1).

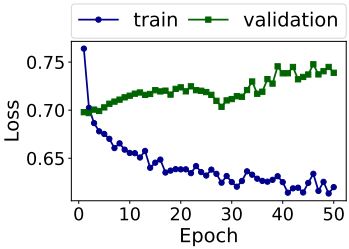


(f) Loss, **GloVe** architecture *env.*-classifier – *long-term* model (1).

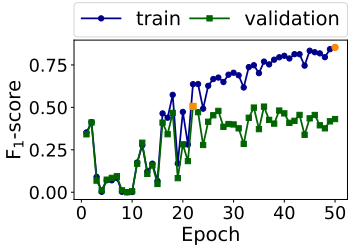
Fig. 5.2.: F_1-score and loss curve for the *env.*-classifiers – *long-term* model (1).



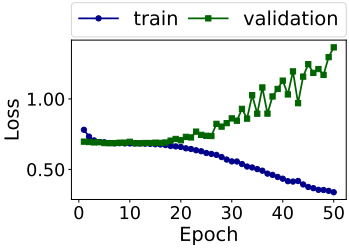
(g) F_1-score, **CNN** architecture *env.*-classifier – *long-term* model (1).



(h) Loss, **CNN** architecture *env.*-classifier – *long-term* model (1).



(i) F_1-score, **BERT** architecture *env.*-classifier – *long-term* model (1).

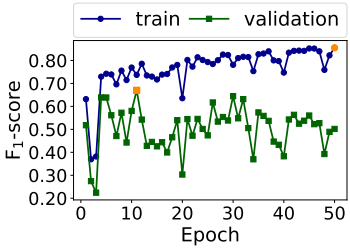


(j) Loss, **BERT** architecture *env.*-classifier – *long-term* model (1).

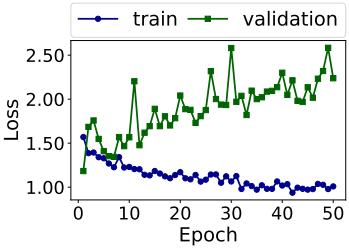
Fig. 5.2.: F_1-score and loss curve for the *env.*-classifiers – *long-term* model (1).

data. Consequently, the classifiers learn certain features from the training set, but the same pattern seems not to apply on the validation or test set.

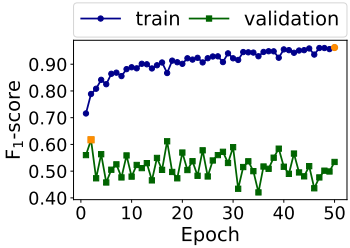
Finally, we take a closer look at the *combined*-classifiers for the *long-term* model (1) (see Fig. 5.3). Similar to the *direct*-classifiers and *env.*-classifiers of the *long-term* model (1), the F_1-score curves on the training data are increasing but there is no substantial improvement on the validation data. The corresponding loss curves show that the loss on the validation data tends to increase shortly after the first few epochs, while the loss on the training data constantly declines. Again, this indicates that the classifiers



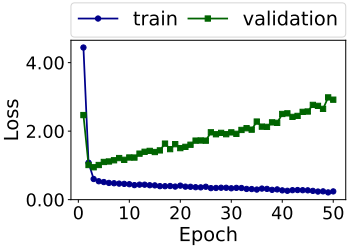
(a) F_1-score, **BOW** architecture *combined*-classifier – *long-term* model (1).



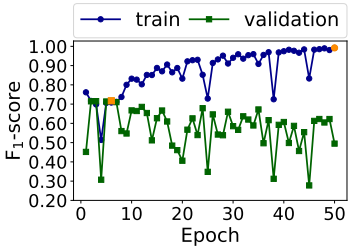
(b) Loss, **BOW** architecture *combined*-classifier – *long-term* model (1).



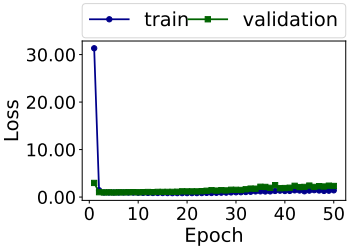
(c) F_1-score, **TF-IDF** architecture *combined*-classifier – *long-term* model (1).



(d) Loss, **TF-IDF** architecture *combined*-classifier – *long-term* model (1).

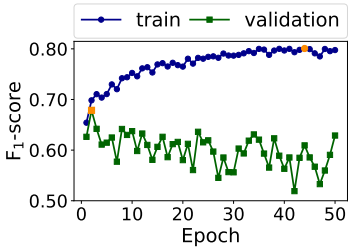


(e) F_1-score, **GloVe** architecture *combined*-classifier – *long-term* model (1).

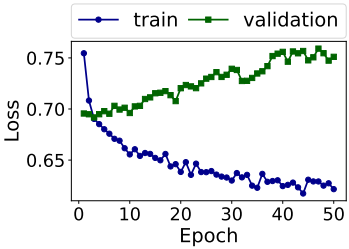


(f) Loss, **GloVe** architecture *combined*-classifier – *long-term* model (1).

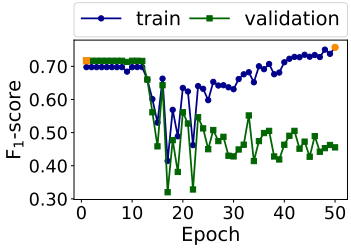
Fig. 5.3.: F_1-score and loss curve for the *combined*-classifiers – *long-term* model (1).



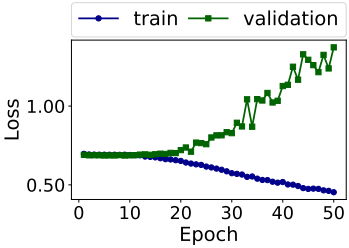
(g) F_1-score, **CNN** architecture *combined*-classifier – *long-term* model (1).



(h) Loss, **CNN** architecture *combined*-classifier – *long-term* model (1).



(i) F_1-score, **BERT** architecture *combined*-classifier – *long-term* model (1).



(j) Loss, **BERT** architecture *combined*-classifier – *long-term* model (1).

Fig. 5.3.: F_1-score and loss curve for the *combined*-classifiers – *long-term* model (1).

overfit and the learned features do not generalise well.

5.2.1. LIME values – *long-term* model (1)

We have seen that the classifiers indeed learn from the training data but the classifiers do not generalise well. Therefore, in this subsection, we analyse exemplary network architectures using LIME values to see what features the classifiers consider to be important for the classification task. Exemplary, we show how the best performing classifier on the

validation data would predict randomly selected documents from our test set according to the underlying features.[5] Furthermore, we show the global feature importance of the best classifier on the validation data for the *direct*-classifier, *env.*-classifier, and *combined*-classifier.[6] Generally, we expect the *direct*-classifier to learn finance-related words that indicate the future financial performance. Likewise, we expect the *env.*-classifier to learn words related to the environmental performance, e.g., "carbon", "environmental", or "emissions". From the *combined*-classifier, we expect words that are related to the financial performance but are potentially more specific, since we add environmental performance as an additional input to the classifier.

From Table 5.2, we saw that the best trained *direct*-classifier for the *long-term* model (1) is the **BOW** classifier. Although the **GloVe** and **BERT** classifiers achieved slightly better F_1-scores, the results of the **GloVe** classifier were identical to the majority baseline and we do not consider the **BERT** classifiers for the LIME value analysis. Accordingly, Figure 5.4 shows the global feature importance of the **BOW** *direct*-classifier from the 10-ths epoch on the test set, i.e., the **BOW** classifier that performed best on the validation data.

From Fig. 5.4, we see that the most important feature is the word "ended", followed by words such as "capital", "product", and "operating". As we would expect, the most important features are indeed words related to the financial performance. However, the previous results have shown that these features are not sufficient to lead to a classifier that is superior to the majority baseline. Consequently, we look at how the **BOW** *direct*-classifier for the *long-term* model (1) would classify randomly selected sample reports.

The first sample report is from EQT Corporation, an American producer

[5] Note that we do not consider **BERT** in the LIME value analysis, as **BERT** uses sentence embeddings instead of words.

[6] Note that we do not calculate the LIME values for all classifiers since all classifiers were not able to surpass the majority baseline. For this reason, we only show the top ten LIME values for the classifier that achieved the highest F_1-score compared to the majority baseline.

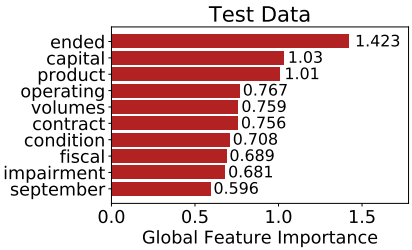


Fig. 5.4.: Global feature importance – top ten absolute LIME values for the *long-term* model (1) **BOW** *direct*-classifier calculated on ten percent of the test set.

of natural gas. Fig. 5.5 shows the top ten LIME values for the 10-Q report of EQT Corporation that was filed on the 27.10.2016. From Fig. 5.5, we see that the classifier correctly predicted that the financial performance would not improve, i.e., Prediction = True label. The classifier predicted that the probability for an improvement is only 0.0765, i.e., 7.65%. Since positive LIME values increase the probability that the classifier predicts an improvement, we see that the words "ended" and "volumes" – words that also rank highly in the global feature importance – are associated with a financial improvement. On the other hand, words such as "partly" and "september" contribute more strongly to the classification as class 0, that is, a decline in earnings.

Looking into the EQT filing, we find that "mvp" stands for the "Mountain Valley Pipeline Project". Hence, the classifier learned that the Mountain Valley Pipeline Project positively contributes to the financial performance. However, "eqm", which stands for EQT Midstream Partners, LP, is associated with a decline in financial performance. This probably relates to asset sales of the MVP project to EQM in 2016. Nevertheless, words such as "mvp" and "eqm" are very company-specific and probably do not generalise well. Furthermore, it is possible that the calculation of LIME values is influenced by higher attribution features. This means for **BOW** that if a word appears more often in a report, it can influence the

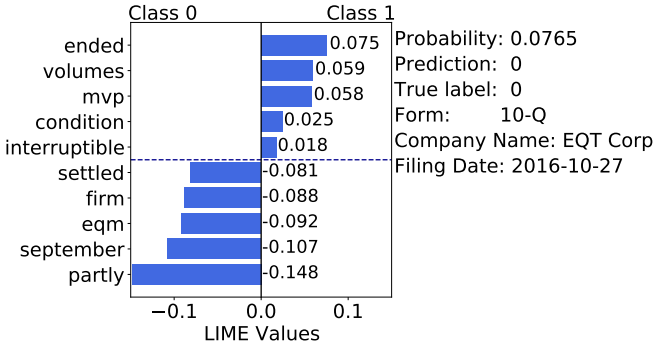


Fig. 5.5.: LIME values for sample report (I.) – *long-term* model (1) **BOW** *direct*-classifier.

calculation of the LIME values. Apparently, this is also the case for the EQT report: The word "eqm" is among the top 15 most occurring words in the filing – after cleaning for stopwords and non-alphabetic characters.

Fig. 5.6 shows the LIME values for another sample report. We see that that the classifier predicted the financial performance for the 10-K filing from Union Pacific, a railway company, incorrectly. With a probability of 0.9698, the classifier was confident that the filing indicates a financial improvement. Nevertheless, the consecutive earnings did not improve. Although the classifier learned that words like "ton", "track", "corridors", and "density" indicate an earnings decline, the classifier was tricked by the words "capital", "fuel", "shipments", "opeb"[7], "multiplying", and "unions".

Similarly, Figure 5.7 shows the LIME values for a third randomly selected report from our test set. For this 10-K report from International Paper, a packaging company, the classifier predicted no consecutive financial improvement, although in reality the financial performance did improve. The classifier appears to be offset by the spinoff of the so called

[7]In the report, "opeb" is used as an abbreviation for "pension and medical and life insurance benefits".

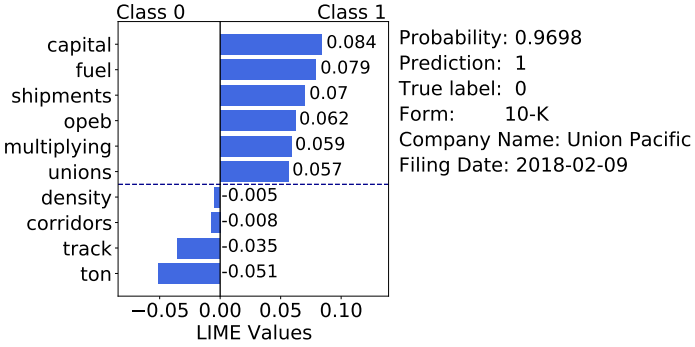


Fig. 5.6.: LIME values for sample report (II.) – *long-term* model (1) **BOW** *direct*-classifier.

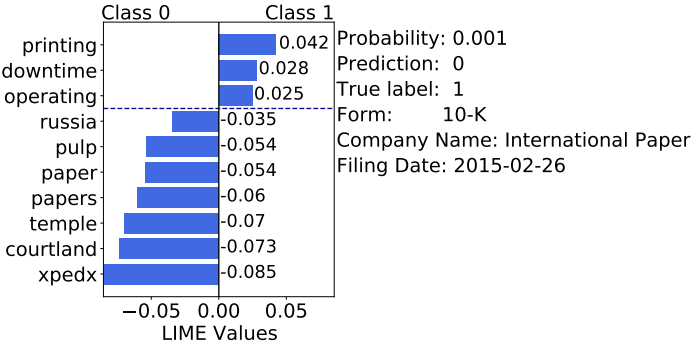


Fig. 5.7.: LIME values for sample report (III.) – *long-term* model (1) **BOW** *direct*-classifier.

"xpedx" division in 2014 as well as other words such as "courtland" and "temple". While there are some words like "printing", "downtime", and "operating" that the classifier associates with a financial improvement, the words associated with a consecutive earnings decline prevail. Additionally, the features selected by the LIME algorithm show again that the words are quite company-specific. This might explain why the **BOW** *direct*-classifier

fails to generalise and tends to overfit.

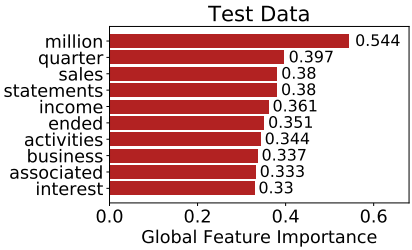


Fig. 5.8.: Global feature importance – top ten absolute LIME values for the *long-term* model (1) **GloVe** *env.*-classifier calculated on ten percent of the test set.

For the *long-term* model (1) *env.*-classifier, **GloVe** achieved the best F_1-score (.550). Thus, Figure 5.8 shows the global feature importance of the **GloVe** classifier. Contrary to what we would expect from a classifier predicting the environmental performance, the most important features for the **GloVe** *env.*-classifier are finance-related words. Figure 5.8 shows no words that could be related to the environmental performance.

We look again at some sample reports from our test set. Figure 5.9 presents the LIME values for a randomly selected sample report. Although the **GloVe** *env.*-classifier was 97.14% certain that the report belongs to class 1, the true class is indeed class 0. This means that in reality no environmental improvement followed the publication of the filing, despite the classifier's prediction of an environmental improvement. Additionally, we see no words related to the environmental performance. Instead, it is surprising to see that the classifier learned that a word like "obligation" is positively associated with an improvement in environmental performance – a word that we would rather expect in the financial context. Taken together, the words appear to be quite random, given that the classifier predicted the filing incorrectly and did also not outperform the majority baseline.

Selecting another sample report for the env.-classifier shows similar

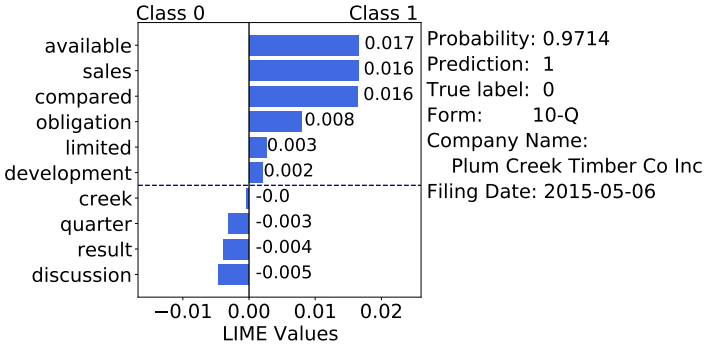


Fig. 5.9.: LIME values for sample report (I.) – *long-term* model (1) **GloVe** *env.*-classifier.

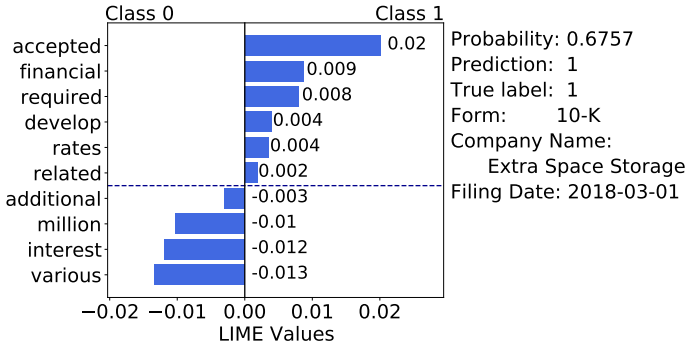


Fig. 5.10.: LIME values for sample report (II.) – *long-term* model (1) **GloVe** *env.*-classifier.

results (see Fig. 5.10). For the report from Extra Space Storage, a real estate investment trust, we find no words that are potentially related to the environmental performance of the company. Although the classifier predicted the environmental performance for the 10-K report correctly, it appears that the *env.*-classifier learns features that correlate with the environmental performance but purely by chance. This might explain

why the *env.*-classifier fails to surpass the majority baseline and, as a result, overfits.

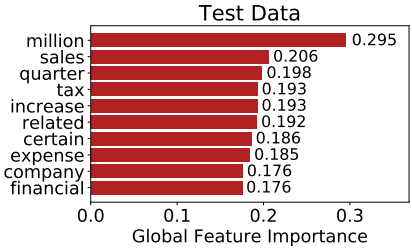


Fig. 5.11.: Global feature importance – top ten absolute LIME values for the *long-term* model (1) **GloVe** *combined*-classifier calculated on ten percent of the test set.

Lastly, we present the global feature importance of the best *combined*-classifier. We choose again the **GloVe** network architecture as an example, since the F_1-score of the **GloVe** *combined*-classifier is closest to the corresponding majority baseline. From Figure 5.11, we see again that the classifier picked up words such as "million", "sales", and "quarter". However, these words are simply among the top 15 most occurring words in the corpus. Consequently, the larger the feature attributions the larger the effect on the overall model predictions. For this reason, we also calculate the LIME values for randomly selected sample reports.

Figure 5.12 presents the LIME values for the 10-K filing from Occidental Petroleum published on the 23.02.2017. We see that the classifier predicted the consecutive earnings improvement correctly. However, all shown features contribute positively to the prediction probability, even words like "oil" and "decreases". While "decreases" could also relate to expenses that are decreasing, we would not expect a word like "oil" to contribute positively to the prediction probability of class 1, since the classifier also received the environmental class label as an input. It is more likely that the classifier simply always predicts an improvement in the financial performance, as Table 5.2 showed **GloVe** to be very close to the majority

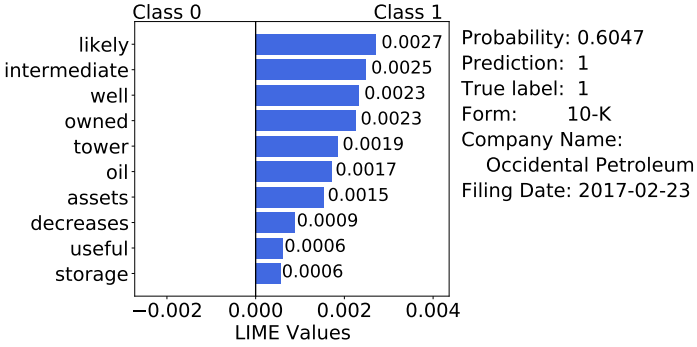


Fig. 5.12.: LIME values for sample report (I.) – *long-term* model (1) **GloVe** *combined*-classifier.

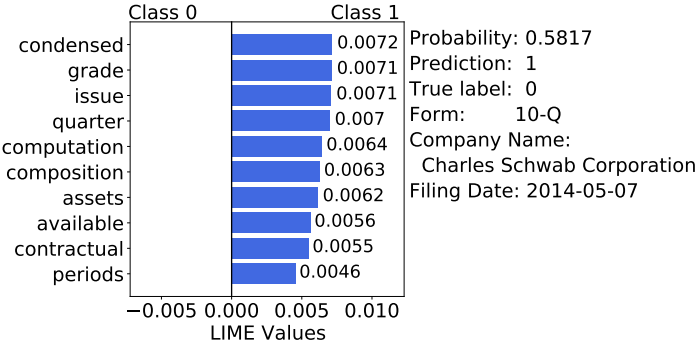


Fig. 5.13.: LIME values for sample report (II.) – *long-term* model (1) **GloVe** *combined*-classifier.

baseline.

As we can see from Figure 5.13, the classifier also predicted a financial improvement for Charles Schwab's filing. However, the prediction does not coincide with the true label. Instead, we see again that all features contribute positively to the prediction of class 1. Since the F_1-score of the **GloVe** *combined*-classifier is almost identical to the majority baseline,

it is reasonable to assume that the classifier just learned to always predict an improvement in the company's financial performance.

5.2.2. Comparison *direct*-classifier and *combined*-classifier – *long-term* model (1)

For the *long-term* model (1), we discussed the results of the *direct*-classifier, the *env.*-classifier, and the *combined*-classifier in relation to the majority baseline. However, we have not discussed whether the narrative environmental performance contributes to the financial classification task from text. Therefore, Table 5.3 shows the comparison of the *direct*-classifier and the *combined*-classifier. The *p-value* was created from 10,000 bootstrap re-samples.[8]

	long-term (1) model						
	direct-classifier			*combined*-classifier			
	P	R	F_1	P	R	F_1	*p-value*
BOW	.529	.855	.653	.523	.890	.659	0.3563
TF-IDF	.527	.685	.596	.533	.773	.631	0.0052*
GloVe	.528	1	.691	.529	.996	.691	—
CNN	.539	.718	.616	.530	.879	.661	0.0024*
BERT	.533	.973	.689	.528	1	.691	0.1520

Table 5.3.: Comparison of the *long-term* model (1) *direct*-classifier and the *combined*-classifier with precision, recall, and F_1-score, respectively. The *p-value* is created from 10,000 bootstrap re-samples (see Subsection 3.4.5). * indicates statistical significance at the 5% level ($\alpha_{adj} = 0.01$) using the Bonferroni corrected significance level.

For the **TF-IDF** and the **CNN** network architectures, we find that the *combined*-classifier is significantly better than the *direct*-classifier.

[8] Since the F_1-score of the **GloVe** *combined*-classifier is not greater than the F_1-score of the **GloVe** *direct*-classifier, we do not examine whether the *combined*-classifier is superior to the *direct*-classifier.

Hence, we reject the corresponding null hypothesis that the *combined*-classifier is not better than the *direct*-classifier. However, for the **BOW**, **GloVe**, and **BERT** network architecture, we cannot reject the null hypothesis. Consequently, we can also not reject the corresponding null hypothesis with respect to our proposed hypothesis (H3) that the narrative environmental performance contained within the MD&A section contributes to the company's financial performance. Only the **TF-IDF** and the **CNN** network architectures showed that the *combined*-classifier is significantly better than the *direct*-classifier. However, all classifiers – even the **TF-IDF** and the **CNN** *combined*-classifier – are not superior to the majority baseline. Thus, we do not find statistical evidence for (H3) when employing the *long-term* model (1).

5.3. Short-term model (2a) – Empirical Results

In this section, we look at the results of the *short-term* model (2a) in more detail. The *short-term* model (2a) employs the binary encoded $\text{BHAR}_{i,\text{short}}$ label for the *direct*- and *combined*-classifier, and the binary encoded $\Delta\text{Env}_{i,t}$ label for the *env.*-classifier. Table 5.4 shows the comparison of the different classifier network architectures.[9]

From Table 5.4, we see that no *direct*-classifier is superior to the majority baseline. Although **BERT** achieved a F_1-score equal to the majority baseline, recall and precision indicate that the best performing **BERT** *direct*-classifier on the validation data is indeed a classifier that always predicts class 1, i.e., an improvement in the stock price performance. The second best *direct*-classifier is the **GloVe** network architecture. However, the **GloVe** classifier also achieved no F_1-score superior to the majority baseline. Thus, we do not find evidence that the MD&A section has meaningful information content with respect to the financial information for the *short-term* model (2a).

For the *short-term* model (2a) *env.*-classifiers, we see that the F_1-scores

[9]Note that these results are consistent with Table 5.1 but only the results for the *short-term* model (2a) are shown.

	short-term (2a) model								
	direct-classifier			*env.*-classifier			*combined*-classifier		
	P	R	F_1	P	R	F_1	P	R	F_1
Baseline	.490	1	.657	.335	1	.502	.490	1	.657
BOW	.475	.770	.587	.358	.472	.407	.487	.786	.601
TF-IDF	.478	.684	.563	.337	.427	.377	.477	.596	.530
GloVe	.500	.850	.630	.342	.885	.494	.490	1	.657
CNN	.484	.594	.534	.364	.153	.215	.469	.637	.540
BERT	.490	1	.657	.315	.212	.253	.484	.711	.576

Table 5.4.: *Short-term* model (2a) results for the *direct*-classifier, the *env.*-classifier, and the *combined*-classifier on the test set with precision, recall, and F_1-score, respectively. Baseline is the majority baseline. The *short-term* model (2a) uses the binary encoded $\text{BHAR}_{i,\text{short}}$ label and the binary encoded $\Delta\text{Env}_{i,t}$ label for the *env.*-classifier.

are generally lower, with the **CNN** and **BERT** classifiers performing quite poorly. Similar to the *long-term* model (1) *env.*-classifiers, the trained *env.*-classifier that performed best for the *short-term* model (2a) is the **GloVe** *env.*-classifier. However, no trained *env.*-classifier is able to find enough meaningful features to outperform the majority baseline. Hence, for the *short-term* model (2a), we cannot reject the the hypothesis that the MD&A section has no meaningful information content with respect to the environmental information, that is, the corresponding null hypothesis to (H2).

For the *combined*-classifiers, we see from Table 5.4 that the **GloVe** network architecture simply always predicts an improvement in the stock price performance. The second best *short-term* model (2a) *combined*-classifier is indeed a simple Bag-of-Words approach. However, no trained *combined*-classifier is superior to the majority baseline. For the *short-term* model (2a), this indicates that we might not be able to reject the hypothesis that the narrative environmental performance contained

within the MD&A section does not contributes to a company's financial performance, i.e., the corresponding null hypothesis to (H3).[10]

5.3.1. LIME values – *short-term* model (2a)

Although no trained classifier is able to surpass the majority baseline, we exemplary analyse the best performing *direct*-classifier, *env.*-classifier, and *combined*-classifier, that is, the classifier that achieved a F_1-score closest to the majority baseline, to understand the reasoning behind the classification. Therefore, we show the global feature importance for each of these classifiers and the top ten LIME values on some randomly selected sample reports from our test set.

For the *short-term* model (2a) *direct*-classifier, we saw that the best classifier is the **GloVe** *direct*-classifier. Hence, Figure 5.14 shows the global feature importance of the **GloVe** *direct*-classifier on ten percent of the test set. We see that the most important features globally are indeed words that are among the most common words in the corpus. Hence, we pick again some random sample reports to analyse the LIME values for the **GloVe** *direct*-classifier.

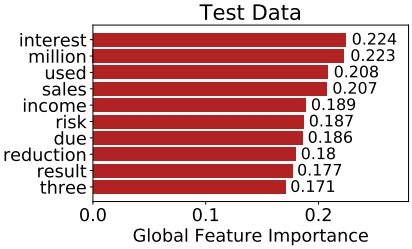


Fig. 5.14.: Global feature importance – top ten absolute LIME values for the *short-term* model (2a) **GloVe** *direct*-classifier calculated on ten percent of the test set.

[10]The F_1-score and loss curve for the *short-term* model (2a) can be found in the appendix.

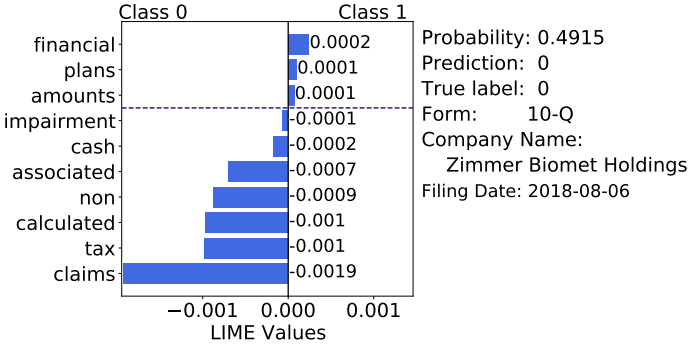


Fig. 5.15.: LIME values for sample report (I.) – *short-term* model (2a) **GloVe** *direct*-classifier.

The first randomly selected report is from Zimmer Biomet Holdings, a medical devices manufacturer. Figure 5.15 presents the according LIME values for the filing published on the 06.08.2018. We see that the classifier correctly predicted the financial stock price performance following the publication of the filing. However, the classifier predicted the stock price decline only with a probability of 0.5085% (= probability − 1), which indicates that the classifier was not so confident whether the filing belongs to class 1 or class 0. The features influencing the prediction probability the most are words related to the financial performance. Indeed, words such as "claims", "tax", and "impairment" are commonly associated with a financial decline. On the other hand, words such as "financial" and "plans" can be associated with class 1, i.e., a stock price improvement.

Figure 5.16 shows that the **GloVe** *direct*-classifier predicted the 10-Q Form of Pfizer Inc also correctly. However, the feature that contributed the most to the prediction as stock price improvement is the word "uncertainties". Usually, in the financial context, a word like "uncertainties" is negatively connoted. Additionally, a self-referencing feature like "pfizer" probably does not generalize well and hinders the classifier to achieve better F_1-scores.

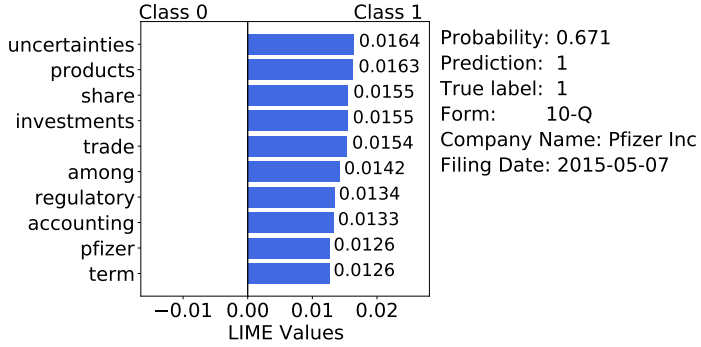


Fig. 5.16.: LIME values for sample report (II.) – *short-term* model (2a) **GloVe** *direct*-classifier.

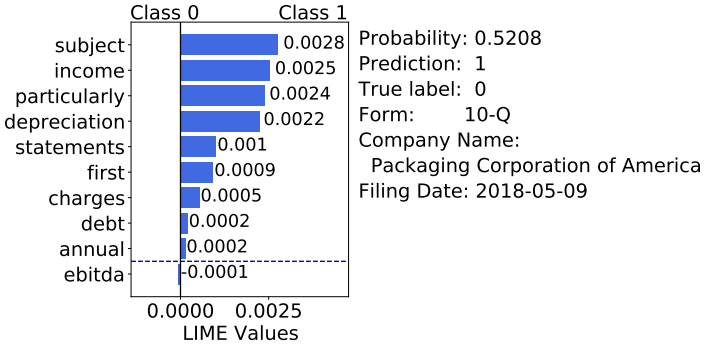


Fig. 5.17.: LIME values for sample report (III.) – *short-term* model (2a) **GloVe** *direct*-classifier.

We look at one last sample report for the *short-term* model (2a) **GloVe** *direct*-classifier (see Fig. 5.17). The report is from the Packaging Corporation of America, a packaging company. Although the classifier learned that words like "subject", "income", "particularly", and "depreciation" are associated with an increasing stock price in relation to the market, this was not true for the filing from Packaging Corporation of America.

We see that the classifier incorrectly predicted a stock price improvement compared to the market. However, the **GloVe** *direct*-classifier shows that it predicted class 1, i.e., a consecutive financial improvement, only with a probability of 52.08%.

For the *short-term* model (2a) *env.*-classifier, we also look at the **GloVe** network architecture, since the **GloVe** *env.*-classifier achieved the highest F_1-score compared to the majority baseline. Figure 5.18 shows the global feature importance of the *short-term* model (2a) **GloVe** *env.*-classifier. Apparently, among the most important features globally are no words related to the environemental performance except maybe the word "impact", e.g., impact investments. However, the words are similar to the global feature importance from the *direct*-classifier and predominantly the most common words in the corpus.

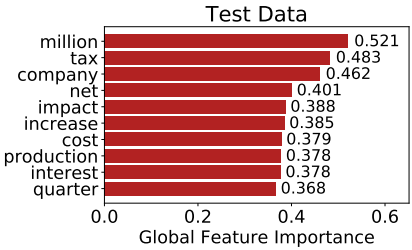


Fig. 5.18.: Global feature importance – top ten absolute LIME values for the *short-term* model (2a) **GloVe** *env.*-classifier calculated on ten percent of the test set.

For the *short-term* model (2a) **GloVe** *env.*-classifier, we randomly selected again a sample report to calculate the LIME values. Figure 5.19 presents the LIME values for the Occidental Petroleum 10-K Form that was published on the 23.02.2017. The classifier predicted class 1 with a certainty of 95.38%. While the classifier learned that the shown features contribute positively to the prediction of class 1, the actual class is class 0, meaning that the consecutive environmental performance did not improve. Similar to our analysis of the *long-term* model (1) *env.*-classifier, we

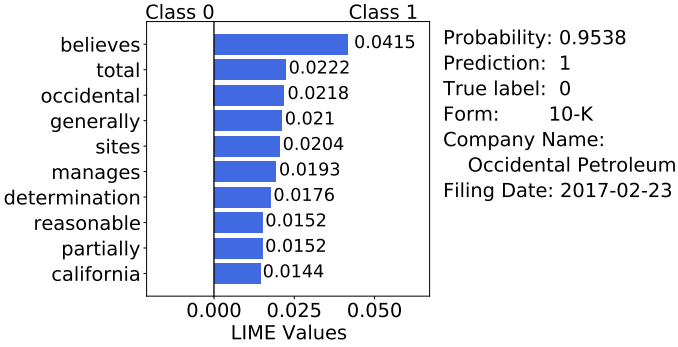


Fig. 5.19.: LIME values for sample report (I.) – *short-term* model (2a) **GloVe** *env.*-classifier.

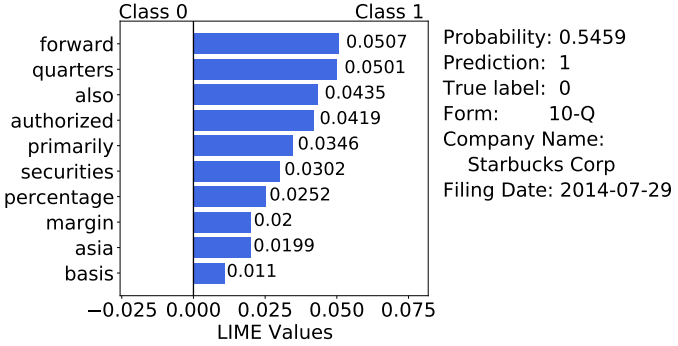


Fig. 5.20.: LIME values for sample report (II.) – *short-term* model (2a) **GloVe** *env.*-classifier.

also find no words that are potentially related to the environmental performance. Instead, the classifier focuses on apparently random words from the report, failing to capture enough meaningful features to explain the environmental performance decline following the report.

For the 10-Q filing by Starbucks Corp, Figure 5.20 shows the according LIME values. The classifier learned that, among others, the features

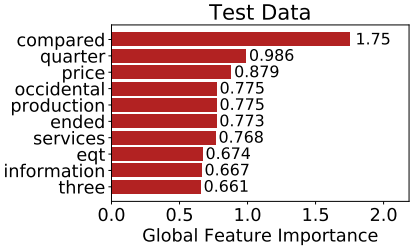


Fig. 5.21.: Global feature importance – top ten absolute LIME values for the *short-term* model (2a) **BOW** *combined*-classifier calculated on ten percent of the test set.

"forward", "quarters", and "also" contribute to the prediction probability of class 1. However, the filing was indeed followed by an environmental performance decline. While the **GloVe** *env.*-classifier predicts class 1 only with a probability of 54.59%, the classifier still learns no environmentally related words and predicts the environmental performance incorrectly. This leads us to believe that the **GloVe** *env.*-classifier is unable to extract enough meaningful features to correctly predict the reports consecutive environmental performance.

For the *short-term* model (2a) *combined*-classifier, we choose the **BOW** network architecture to explain what the classifier learns. The **BOW** *combined*-classifier achieved the highest F_1-score compared to the majority baseline. Note that we do not choose the **GloVe** *combined*-classifier, since the **GloVe** *combined*-classifiers results are identical to the majority baseline. Figure 5.21 presents the global feature importance of the **BOW** *combined*-classifier. Interestingly, among the features that have the most influence globally are two company-specific terms: "occidental" and "eqt". It appears that among the 10% of the test set these features have a strong feature attribution, potentially setting the classifier off. Thus, we want to look at the LIME values for some randomly selected sample reports.

Figure 5.22 presents the LIME values for the **BOW** *combined*-classifier on the 10-Q report of Pfizer Inc. The classifier learned that the drug

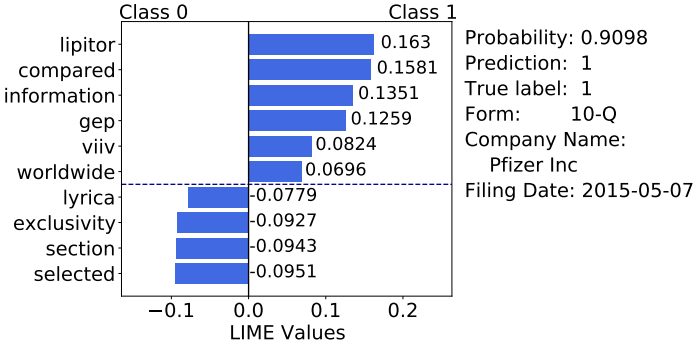


Fig. 5.22.: LIME values for sample report (I.) – *short-term* model (2a) **BOW** *combined*-classifier.

"lipitor" is positively associated with the stock price performance compared to the market, which most likely relates to a reported strong growth for Lipitor in developed markets. Also "gep", which stands for "global established pharmaceutical segment", and "viiv", i.e., "ViiV healthcare limited", are associated with a financial improvement measured in terms of buy-and-hold returns. However, the drug "lyrica" is negatively associated with the stock price performance compared to the market, which probably relates to the loss of exclusivity for "Lyrica" in the "gep" segment. However, while these features explain Pfizer's report quite well, they probably do not generalise well.

Figure 5.23 presents the LIME values for the Textron's report, a multi-industry company, published on the 26.07.2018. The **BOW** *combined*-classifier found that, among others, features like "bell", "arctic", "half", and "finance" indicate that the report would be followed by a decline in the financial performance. However, the consecutive financial performance did indeed improve, although the **BOW** *combined*-classifier predicted class 0 with a probability of 84.78%.

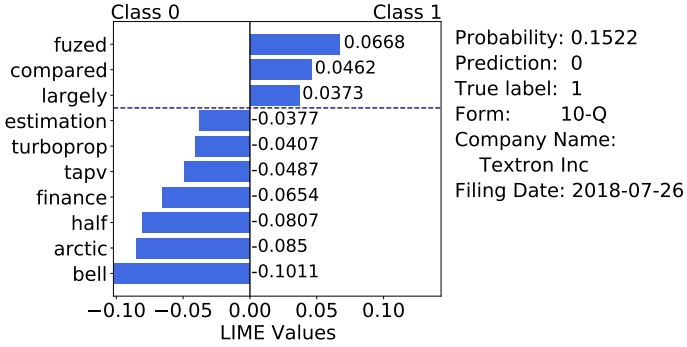


Fig. 5.23.: LIME values for sample report (II.) – *short-term* model (2a) **BOW** *combined*-classifier.

5.3.2. Comparison *direct*-classifier and *combined*-classifier – *short-term* model (2a)

In Table 5.5, we present the results of the *short-term* model (2a) *direct*-classifier and *combined*-classifier next to each other. Both classifiers predict the consecutive company's financial performance measured by buy-and-hold-returns for a 4-day event window. However, the *combined*-classifier also uses the environmental class label as an input. To see whether the *combined*-classifier is superior to the *direct*-classifier, we calculated the *p-value* each time the *combined*-classifier achieved better results than the *direct*-classifier.

From Table 5.5, we see that the **BOW** *combined*-classifier as well as the **GloVe** *combined*-classifier achieve better F_1-scores than the corresponding *direct*-classifier. Nevertheless, for both *combined*-classifiers the *p-value* indicates that the classifiers are not significantly better than the corresponding *direct*-classifier at the 5% significance level. Consequently, we do not find statistical evidence for (H3) when considering the *short-term* model (2a).

	short-term (2a) model						
	direct-classifier			*combined*-classifier			
	P	R	F_1	P	R	F_1	*p-value*
BOW	.475	.770	.587	.487	.786	.601	0.1477
TF-IDF	.478	.684	.563	.477	.596	.530	—
GloVe	.500	.850	.630	.490	1	.657	—
CNN	.484	.594	.534	.469	.637	.540	0.5950
BERT	.490	1	.657	.484	.711	.576	—

Table 5.5.: Comparison of the *short-term* model (2a) *direct*-classifier and the *combined*-classifier with precision, recall, and F_1-score, respectively. The *p-value* is created from 10,000 bootstrap re-samples (see Subsection 3.4.5). * indicates statistical significance at the 5% level ($\alpha_{adj} = 0.01$) using the Bonferroni corrected significance level.

5.4. Short-term model (2b) – Empirical Results

Lastly, we present the results of the *short-term* model (2b). Like the *short-term* model (2a), the *short-term* model (2b) uses buy-hold-returns to measure the financial performance of the company and the monthly change in the environmental class label to measure the company's environmental performance. However, the *short-term* model (2b) uses a longer period to measure the buy-hold-returns, that is, a monthly event window.

Table 5.6 presents the results for the *short-term* model (2b) *direct*-classifier, *env.*-classifier, and *combined*-classifier for all network architectures. The best *direct*-classifier is the **BERT** classifier, followed by the **GloVe** and the **CNN** classifier. However, no trained *direct*-classifier is superior to the majority baseline. Thus, we cannot reject the hypothesis that the MD&A section has no meaningful information content with respect to the financial information for the *short-term* model (2b), that is, the corresponding null hypothesis to (H1b).

For the *env.*-classifiers, the results are identical to the *short-term*

	short-term (2b) model								
	direct-classifier			*env.*-classifier			*combined*-classifier		
	P	R	F_1	P	R	F_1	P	R	F_1
Baseline	.490	1	.657	.335	1	.502	.490	1	.657
BOW	.503	.798	.617	.358	.472	.407	.498	.762	.603
TF-IDF	.501	.689	.580	.337	.427	.377	.506	.608	.552
GloVe	.495	.893	.637	.342	.885	.494	.489	.983	.654
CNN	.525	.808	.636	.364	.153	.215	.491	.682	.571
BERT	.488	.980	.652	.315	.212	.253	.478	.686	.564

Table 5.6.: *Short-term* model (2b) results for the *direct*-classifier, the *env.*-classifier, and the *combined*-classifier on the test set with precision, recall, and F_1-score, respectively. Baseline is the majority baseline. The *short-term* model (2a) uses the binary encoded $BHAR_{i,long}$ label and the binary encoded $\Delta Env_{i,t}$ label for the *env.*-classifier.

model (2a), since the *short-term* model (2b) uses the same environmental class label. Accordingly, the **GloVe** *env.*-classifier achieves the highest F_1-score compared to the majority baseline. Again, no trained *env.*-classifier is superior to the majority baseline. Thus, we cannot reject the the hypothesis that the MD&A section has no meaningful information content with respect to the environmental information for the *short-term* model (2b), that is, the corresponding null hypothesis to (H2).[11]

Considering the *combined*-classifiers, we see from Table 5.6 that the **GloVe** *combined*-classifier achieves the highest F_1-score. However, no classifier learns enough meaningful features to outperform the majority baseline. For the *short-term* model (2b), this indicates that we might not be able to reject the hypothesis that the narrative environmental performance contained within the MD&A section does not contributes to a company's financial performance, i.e., the corresponding null hypothesis

[11] For a discussion see Section 5.3.

to (H3).[12]

5.4.1. LIME values – *short-term* model (2b)

In this subsection, we show what features are most important for the classification task for some exemplary *short-term* model (2b) classifiers, to understand why the classifiers fail to surpass the majority baseline. We choose the best performing *direct*-classifier, *env.*-classifier, and *combined*-classifier on the validation set, to evaluate the corresponding LIME values on the test set. Additionally, we show the global feature importance for each of these classifiers.

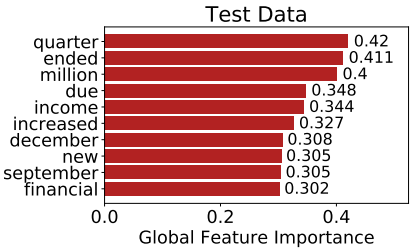


Fig. 5.24.: Global feature importance – top ten absolute LIME values for the *short-term* model (2b) **GloVe** *direct*-classifier calculated on ten percent of the test set.

For the *short-term* model (2b), the best result for the *direct*-classifier was achieved by **GloVe**. Fig. 5.24 shows the global feature importance for the **GloVe** *direct*-classifier calculated on ten percent of the test set. We see that "quarter", "ended", and "million" are the features that influence the classification the most globally. Also the months "december" and "september" seem to play a crucial role for the classification task. Hence, let us see how the **GloVe** *direct*-classifier would classify some randomly selected sample reports from our test set.

[12]The F_1-score and loss curve for the *short-term* model (2b) can be found in the appendix.

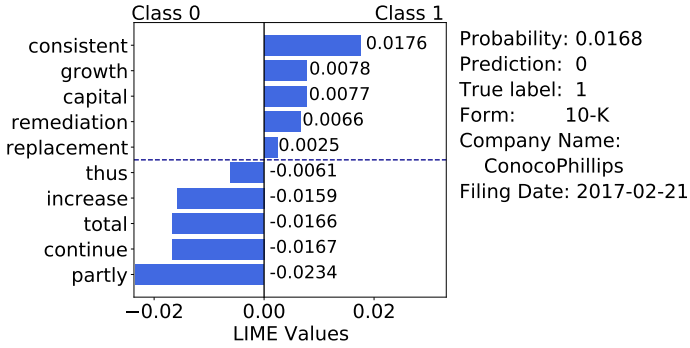


Fig. 5.25.: LIME values for sample report (I.) – *short-term* model (2b) **GloVe** *direct*-classifier.

The first sample report we consider is from ConocoPhillips, a large oil producer. While the **GloVe** *direct*-classifier classifier learned that "consistent", "growth", and "capital" are associated with a financial improvement, the classifier predicted that the report would not be followed by a stock price increase compared to the market. Apparently, the features that misled the classifier are words like "partly", "continue", and "increase", among others, which have a higher feature weight.

The next sample report is the 10-Q Form from Electronics Arts, a video game company. While the classifier was 56.36% certain that the financial performance did not improve, the classifier predicted the financial performance incorrectly. The **GloVe** *direct*-classifier learned to associate the word "generation", referring to Electronics Arts new generation of consoles, with a financial decline. However, when looking at an example sentence from the report, we see that the related investments into the new generation does not have to be an economically bad decision.[13] Probably, the classifier should have set a higher feature weight on the

[13] "EA delivered five major products for each of these new-generation console systems around the time of their launch, and we are continuing to make significant investments in products and services for these new consoles.", Electronic Arts Inc., 2014, p. 33

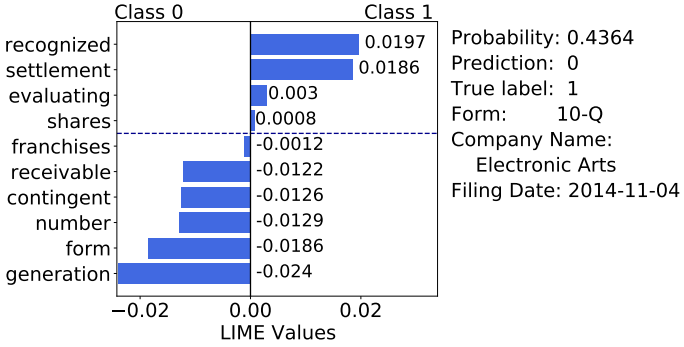


Fig. 5.26.: LIME values for sample report (II.) – *short-term* model (2b) **GloVe** *direct*-classifier.

words "recognized" and "settlement" instead.

For the *env.*-classifier, we already showed some sample reports for the best performing *env.*-classifier, that is, the **GloVe** *env.*-classifier (see Subsection 5.3.1). This is due to the fact that the *short-term* model (2a) and *short-term* model (2b) share the same environmental class label. Therefore, we only show some sample reports for the second best performing *env.*-classifier, that is, the **TF-IDF** classifier.

Figure 5.27 shows the LIME values for the 10-K Form from ONEOK, a midstream service provider. The classifier correctly predicted that the report is not followed by an environmental improvement. From the words "lpg" and "wtlpg", we can see that the classifier learned that the West Texas LPG Pipeline is associated with an environmental decline. Also other oil related words such as "mbbl", which stands for "thousand barrels of oil per day", "bighorn", for the "Bighorn Gas Gathering system", and "fractionator" are associated with a decline in environmental performance. However, these words are very specific for oil companies and likely do not apply for companies from different industries.

Let us consider again the report from Starbucks that was published on the 29.07.2014. This is the same report that we also used for the *short-*

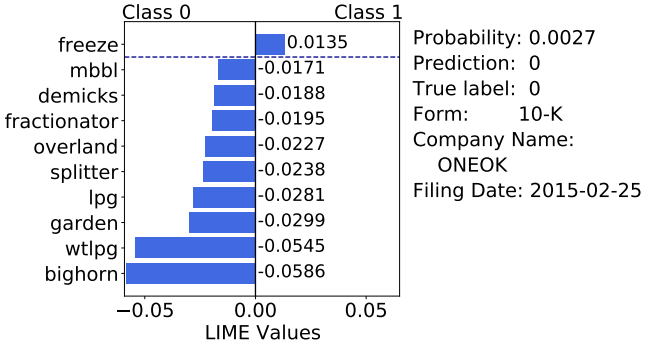


Fig. 5.27.: LIME values for sample report (I.) – *short-term* model (2b) **TF-IDF** *env.*-classifier.

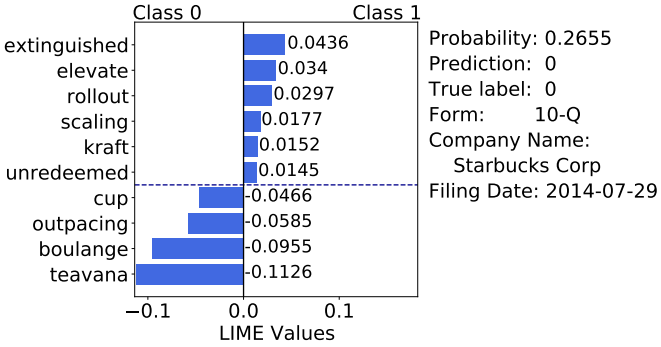


Fig. 5.28.: LIME values for sample report (II.) – *short-term* model (2b) **TF-IDF** *env.*-classifier.

term model (2a) **GloVe** *env.*-classifier. Figure 5.28 shows the top ten LIME values for the report from Starbucks for the **TF-IDF** *env.*-classifier. We see that the **TF-IDF** *env.*-classifier predicted the consecutive environmental performance this time correctly. However, we also see no features that would be expected from a classifier predicting the environmental performance. Thus, it is more likely that the classifier learned those features

that coincide with a consecutive environmental performance decline but that are not necessarily related to the environmental performance.

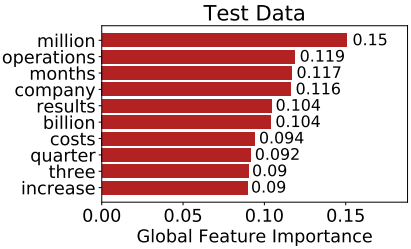


Fig. 5.29.: Global feature importance – top ten absolute LIME values for the *short-term* model (2b) **GloVe** *combined*-classifier calculated on ten percent of the test set.

For the *combined*-classifier, we consider again the classifier that achieved the F_1-score closest to the majority baseline. For the *short-term* model (2b) *combined*-classifier, this is the **GloVe** *combined*-classifier. Figure 5.29 shows the global feature importance. We can see that the words influencing the classifiers decision the most are financial words like "million", "operations", "results", and "billion", among others. We select again some sample reports to analyse how the features affect the **GloVe** *combined*-classifier locally.

The first sample report for the **GloVe** *combined*-classifier is from Whole Foods, a grocery chain. Figure 5.30 shows that the **GloVe** *combined*-classifier incorrectly predicted the financial performance following the report. All shown LIME values are contributing to the prediction of class 1, i.e., a financial improvement. While the classifier was only 53.77% confident with this decision, the prediction is still false. As we have previously discussed, words like "year" and "sales" are among the most common words in the corpus. This probably misled the classifier. Thus, we consider one more sample report for the **GloVe** *combined*-classifier.

The last sample report is the 10-K report from MT Bank Corp, a bank. This time, Figure 5.31 shows that all shown features contributed to the

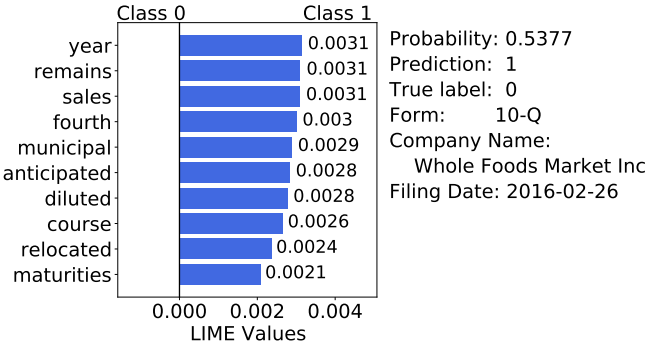


Fig. 5.30.: LIME values for sample report (I.) – *short-term* model (2b) **GloVe** *combined*-classifier.

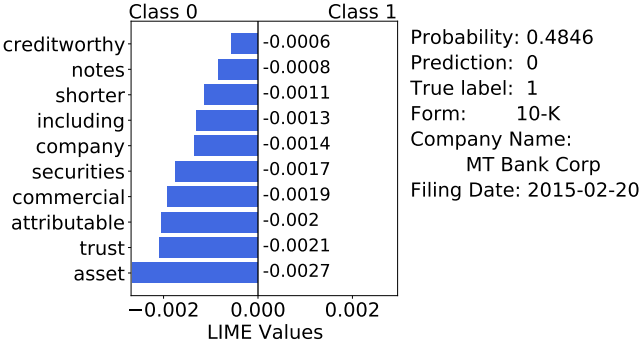


Fig. 5.31.: LIME values for sample report (II.) – *short-term* model (2b) **GloVe** *combined*-classifier.

classifiers prediction as class 0, that is, a financial decline. While we see finance-related words, it is surprising that the classifier associates words like “asset”, “trust”, and “creditworthy” with a financial decline – words that are usually positively connotated in the financial context.

5.4.2. Comparison *direct*-classifier and *combined*-classifier – *short-term* model (2b)

In this subsection, we compare the results of the *short-term* model (2b) *direct*-classifiers and the *short-term* model (2b) *combined*-classifiers. Table 5.7 presents the precision, recall, and F_1-score for the classifiers and the according *p-value*.

From Table 5.7, we see that only the **GloVe** *combined*-classifier achieves a F_1-score higher than the corresponding *direct*-classifier. However, the *p-value* indicates that the *combined*-classifier is not significantly better than the *direct*-classifier. Since the other *combined*-classifiers are also not superior to the corresponding *direct*-classifiers, we do not find statistical evidence for (H3) when considering the *short-term* model (2b).

	short-term (2b) model						
	direct-classifier			*combined*-classifier			
	P	R	F_1	P	R	F_1	*p-value*
BOW	.503	.798	.617	.498	.762	.603	—
TF-IDF	.501	.689	.580	.506	.608	.552	—
GloVe	.495	.893	.637	.489	.983	.654	0.0273
CNN	.525	.808	.636	.491	.682	.571	—
BERT	.488	.980	.652	.478	.686	.564	—

Table 5.7.: Comparison of the *short-term* model (2b) *direct*-classifier and the *combined*-classifier with precision, recall, and F_1-score, respectively. The *p-value* is created from 10,000 bootstrap re-samples (see Subsection 3.4.5). * indicates statistical significance at the 5% level ($\alpha_{adj} = 0.01$) using the Bonferroni corrected significance level.

5.5. Environmental Classifier Troubleshooting – Environmental Word List

For all models, we find no evidence that the MD&A section is associated with the future environmental performance. While we expected that the *env.*-classifiers learn environmental words, the LIME value analysis showed that the classifiers learn words that coincide with the environmental performance but do not actually indicate the environmental performance. This finding is rather unexpected, as we expected that the environmental classifiers learn at least some environmental-related words. However, it might be the case that there are simply no such words in our corpus. In this section, we analyse whether environmental words are present and also how they are distributed among our dataset, to better understand the results of the environmental classifiers. For this reason, we construct a word list of environmental search terms that is based on other authors. Table 5.8 presents the environmental search terms that we would expect an environmental classifier to learn.[14]

First, we examine our entire corpus for the environment-related search terms to obtain an overview of which word is most frequently used among the environmental search terms. Figure 5.32 shows the term frequency in relation to the total number of environmental search terms in our corpus, i.e., for all 10-K and 10-Q filings from 2014 to 2018.[15] We see that the term "environmental" is by far the most common term. The second most occurring term is "epa", which stands for the United States Environmental Protection Agency, shortly followed by "renewable", "emissions", and "waste".[16] The number of times companies make reference

[14] Note that we have only added the term "greenhouse gases".

[15] Note that the term frequency is not to be confused with the term frequency per document. Instead, the frequency simply shows, which environmental search term is used the most in the corpus.

[16] Note that we subtracted the appearances of "climate change" from "climate" and also the appearances of "global warming" from "warming", since otherwise this would lead to a double count of "climate" and "warming". Similarly, we account for double counts of the word "greenhouse" in "greenhouse gas" and "greenhouse gases".

Search Terms	Authors
biodiversity	Baier et al., 2020; Verbeeten et al., 2016
carbon	Feng and Gao, 2020
climate	Baier et al., 2020; Feng and Gao, 2020
climate change	Doran and Quinn, 2009
contamination	Baier et al., 2020
emission	Baier et al., 2020
emissions	Baier et al., 2020; Feng and Gao, 2020 Verbeeten et al., 2016
environmental	Baier et al., 2020; Feng and Gao, 2020
epa	Baier et al., 2020; Feng and Gao, 2020
ghg	Baier et al., 2020; Feng and Gao, 2020 Doran and Quinn, 2009
ghgs	Baier et al., 2020
global warming	Doran and Quinn, 2009
greenhouse	Feng and Gao, 2020
greenhouse gas	Doran and Quinn, 2009
greenhouse gases	
pollution	Baier et al., 2020
recycle	Baier et al., 2020
recycled	Verbeeten et al., 2016
renewable	Baier et al., 2020
sustainability	Baier et al., 2020; Feng and Gao, 2020
waste	Baier et al., 2020; Verbeeten et al., 2016
wastes	Baier et al., 2020
warming	Baier et al., 2020; Feng and Gao, 2020

Table 5.8.: Environmental search terms.

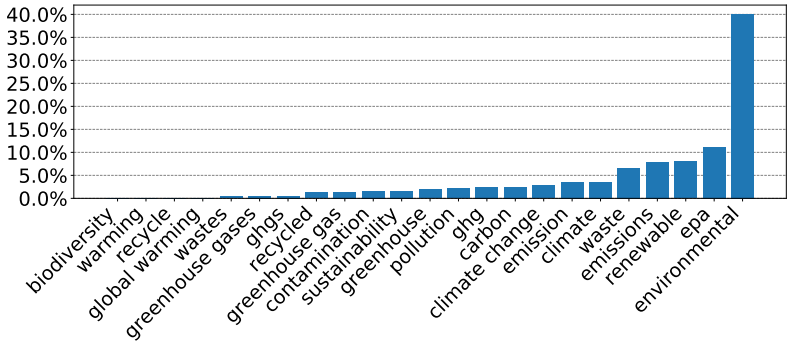


Fig. 5.32.: Frequency of the environmental search terms in the corpus. The term frequency tf for term j is calculated as $\text{tf}_j = \frac{\text{t}_j}{\sum_{i=1}^{n} \text{t}_i}$ with t_j being the number of times term j is mentioned in the corpus, t_i is the number of times term i occurred with $i = 1, \ldots, n$, and n is the total number of search terms used.

to "biodiversity", "warming", "recycle", and "global warming" is on the other hand quite low.

In a second step, we analyse whether there is a trend in mentions of environmental search terms over time and if mentions of environmental search terms are approximately equally distributed among the 10-Q and 10-K reports' MD&A section. Figure 5.33 shows how many reports use at least one or more of the environmental search terms and how many reports use ten or more of the environmental search terms, subdivided into 10-Q and 10-K reports.

Except for the year 2014, we see that on average 57.3% of the 10-K reports use at least one of the environmental search terms. Similarly, on average only 49% of the 10-Q reports mention at least one environmental-related word. When we look at how many reports mention at least ten or more of the environmental search terms, we see a substantial drop for both 10-K and 10-Q reports. Although on average 19.4% of the 10-K reports mention at least ten or more of the search terms, only 11.4%

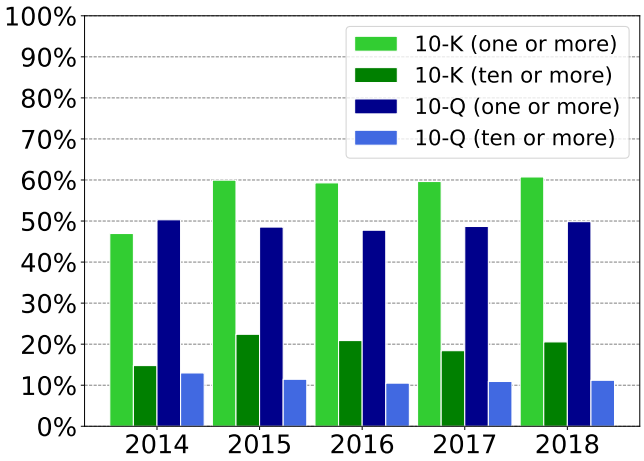


Fig. 5.33.: S&P500 10-K and 10-Q MD&A sections from 2014 to 2018 with mentions of environmental search terms, subdivided into 10-K and 10-Q filings with *one or more mentions* and *ten or more mentions* of environmental search terms.

of the 10-Q reports use ten or more of these words. Furthermore, we see no clear trend over time on how many reports make reference to the environmental search terms.

Since the use of the environment-related terms might not be evenly distributed among different industries, we also analyse the occurrence of environmental search terms with respect to industries. Figure 5.34 shows how many reports use at least one of the environmental search terms. We clearly see that among construction companies, the use of environment-related terms in the MD&A section is the highest. Almost all construction companies in our sample make reference to at least one of the terms. Additionally, in the transportation and public utilities sector the usage of the environmental search terms is quite high. 68% to 78% of transportation and public utilities companies include at least one environmental search term and roughly 71% to 77% in the mining industry.

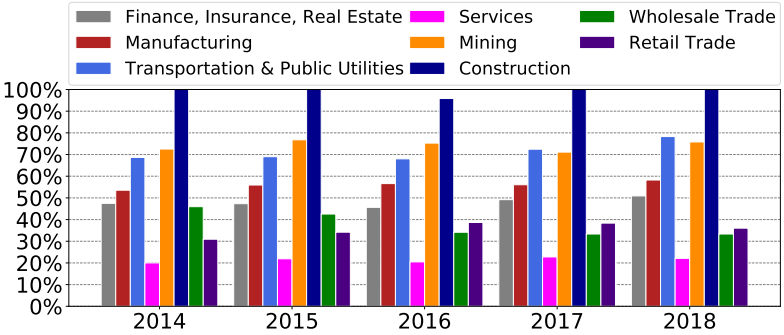


Fig. 5.34.: S&P500 10-K and 10-Q MD&A sections from 2014 to 2018 with *one or more mentions* of environmental search terms, subdivided by industry.

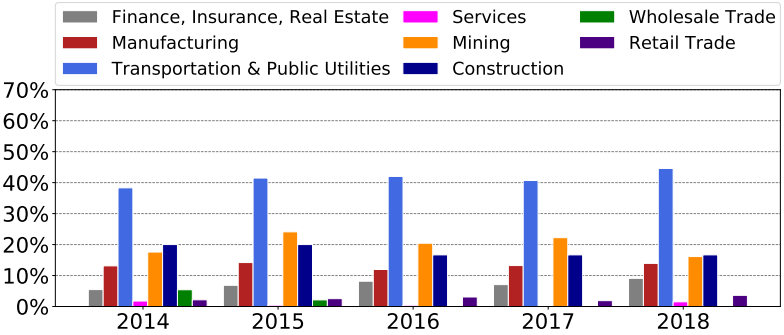


Fig. 5.35.: S&P500 10-K and 10-Q MD&A sections from 2014 to 2018 with *ten or more mentions* of environmental search terms, subdivided by industry.

The industry with the least mentions of environmental search terms is the service industry with only 20% to 23% of the reports containing at least one environment-related term.

This picture drastically shifts when we look at how many reports use at least 10 or more of the environmental search terms. From Figure 5.35, we see that only 16% to 20% of the reports in the construction industry include at least ten of the search term with a decreasing tendency over time. 38% to 45% of the transportation and public utilities MD&A sections make use of ten or more of the environmental terms, but with a slight incline from 2014 to 2018. However, only a small fraction of reports from the service industry, wholesale trade, and retail trade use ten or more environment-related terms. Companies that report the most environmental information are in the "Transportation & Public Utilities" industry – on average 41% of the 10-K and 10-Q filings contain at least 10 or more environmental terms, followed by 20% in the "Mining" industry and 18% in the "Construction" industry. All other industries – "Manufacturing" (13%), "Finance, Insurance, Real Estate" (7%), "Retail Trade" (2.6%), "Wholesale Trade" (1.5%), and "Services" (0.8%) – report less environmental terms.

5.6. Summary

In this chapter, we presented our empirical results. First, we gave a general overview of our results (Section 5.1). We showed that no trained classifier was able to surpass the majority baseline providing no empirical evidence for our hypotheses. Next, we looked at the classifiers in detail. Starting with the *long-term* model (1), we showed that the classifiers learn from the training data but the classifiers do not show a meaningful F_1-score increase on the validation data (Section 5.2). The same holds for the *short-term* model (2a) and (2b) (Section 5.3 and Section 5.4). Consequently, the classifiers do not find enough meaningful features to classify the MD&A section better than the majority baseline. The LIME value analysis confirmed that the *direct*-classifiers and the *combined*-classifiers

learn financial terms but the most relevant features are quite company-specific and probably do not generalise well. Since the *env.*-classifiers did not learn environment-related terms as expected, we analysed whether there are any environmental terms in the MD&A section. For this reason, we constructed an environmental word list based on other authors. It turns out that the usage of environmental terms is unevenly distributed among 10-K and 10-Q reports as well as among different industries with few companies reporting more than 10 environment-related words (Section 5.5). For the *long-term* model (1) and the *short-term* model (2a) and (2b), we also analysed whether the narrative environmental performance contributes to the financial performance prediction task from text in comparing the *direct*-classifiers to the *combined*-classifiers. We find no evidence that the environmental performance contributes to the financial performance prediction task from text.

6. Discussion

In the previous chapter, we presented our empirical findings. For the 10-K and 10-Q reports from 2014 to 2018, we find no evidence that the MD&A section contains meaningful information with respect to the future financial performance, measured in terms of earnings as well as in terms of buy-and-hold returns, and we also do not find evidence for the MD&A section to contain meaningful information with respect to the company's future environmental performance. Additionally, we find no statistical evidence for the environmental performance to contribute to the company's future financial performance. In this chapter, we thus discuss our findings, first, in comparison to other studies (Section 6.1), and, second, in comparison to the underlying theory (Section 6.2). Then, we conclude this chapter by listing and discussing the limitations of our study (Section 6.3).

6.1. Interpretation and Comparison

Our study employs a unique, holistic approach to explain the future financial performance and the future environmental performance from the MD&A section, as well as the contribution of environmental performance to financial performance on a narrative basis. Unlike similar studies, we do not rely on tailored word lists (Feng and Gao, 2020; Jegadeesh and Wu, 2013; Loughran and McDonald, 2011; Doran and Quinn, 2009; Feldman et al., 2008, i.a.) nor do we derive a disclosure score, on which we regress a financial variable (F. Li, 2010b; Clarkson et al., 1999, i.a.) or an environmental variable (Cong et al., 2020; Coburn and Cook, 2014, i.a.). Instead, we employ end-to-end neural networks to draw conclusions about the informativeness of the MD&A section. Additionally, in comparison to

similar studies, which only include 10-K reports in their analysis (Muslu et al., 2015; Jegadeesh and Wu, 2013; F. Li, 2006, i.a.), we also include 10-Q reports in our analysis, as both types of filings should contain market relevant information.

6.1.1. Interpretation and Comparison of Financial Results

The empirical results show that the MD&A section is not associated with the company's future financial performance, measured in terms of earnings and also in terms of buy-and-hold returns. The corresponding majority baseline – a classifier that always predicts an improvement in the financial performance – proves to be superior to all of our trained classifiers. This indicates that the MD&A section does not provide the relevant information to predict the future financial performance more precisely and persistently than just always predicting an improvement. Consequently, it appears that the MD&A section is not informative with respect to the company's own future financial performance and also not useful to investors.

For the *long-term* model (1), we find no association between the MD&A section and the binary encoded earnings class label. This holds for all *direct*-classifier network architectures. Consequently, the general informativeness of the MD&A section appears to be quite low. However, the results could be due to a low performance of the classifiers. Thus, we analysed the classifiers in detail, to find that the classifiers indeed learn from the training data but the same pattern cannot be transferred to the validation or test data. Although we applied state-of-the-art classifiers plus various regularization techniques, which minimize overfitting, the classifiers do not generalise. This indicates that either there is not enough meaningful information in the reports or that the reports are too heterogeneous to extract meaningful information.

The LIME value analysis showed that the classifiers indeed extract meaningful information for some reports but fail on others. This supports the argument that some companies fail to disclose meaningful information

with respect to the company's future financial performance. Furthermore, some companies might be too optimistic, when providing a short-term and long-term analysis of the company's future performance, or simply fail to assess their future performance correctly, which might hinder the classifier to learn meaningful associations. This would be in line with Loughran and McDonald, 2019, who noted that corporate reports show an ambiguity of sentiment and low polarisation, as business reports rarely use negative words unless it is unavoidable.[1] While we cannot reject that meaningful information is *not* present – at least in some reports –, the results show that the information is difficult to locate. Accordingly, if meaningful information is present, the relevant information appears to be surrounded by a lot of boilerplate text.

The same arguments hold for the *short-term* model (2a) and *short-term* model (2b). For the *short-term* model (2a) and *short-term* model (2b), we also find no association between the MD&A section and the binary encoded buy-and-hold return class label. This indicates that the usefulness of the MD&A section to investors is quite limited. Instead, investors seem to focus on the numerical information, since the buy-and-hold returns are indeed significant but not associated with the narrative information content. Consequently, investors either do not read the MD&A section or they do not find meaningful narrative information within the MD&A section. However, the classifiers tendency to overfit as well as the LIME value analysis show that the classifiers learn from some reports but the same pattern cannot be transferred to other filings. Accordingly, not all reports appear to be equally useful for investors.

Our findings are in line with Kothari et al., 2009; Pava and Epstein, 1993; Schroeder and Gibson, 1990, who also find the 10-K and 10-Q reports (Kothari et al., 2009) or, explicitly, the MD&A section (Pava and Epstein, 1993; Schroeder and Gibson, 1990) to be boilerplate. However, as opposed to F. Li, 2006, who showed that an increase in the filings' risk

[1] While companies might exaggerate positiveness, making untrue statements would be a violation of the SEC guidelines, except the reports include modal language, e.g., words like "may", "could", and "probably".

sentiment is associated with lower future earnings, and F. Li, 2010b, who associated the forward-looking statements of the MD&A section to future earnings, we find no such association for earnings. Moreover, in contrast to Jegadeesh and Wu, 2013 and Loughran and McDonald, 2011, we find no association between the filings and market returns.

The differences could stem from the fact that F. Li, 2006, Jegadeesh and Wu, 2013, and Loughran and McDonald, 2011 use the entire 10-K filings and also incorporate background information through a customized word list. Furthermore, contrary to F. Li, 2010b, our approach is neither supported by humanly annotation nor do we filter for sentences that include modal language. This shows that meaningful information might be present in the filings but not exclusively within the MD&A section. In addition, we conclude that our classifiers could be improved by pre-filtering the reports for forward-looking paragraphs or sentences, e.g., by filtering for modal language or keywords that imply forward-looking statements like "next quarter" or "in the future". Otherwise, without providing additional background information, the MD&A section appears to be too long and obscuring.

Already when filtering the reports for the MD&A section, it became apparent that the 10-K and 10-Q forms are not as standardized as we expected. For some reports, the sections' content appears in later parts of the filings making the relevant sections hard to identify. Also fluent transitions into the QQD section make it difficult to differentiate the sections. Furthermore, we observed that 10-Q reports are quite brief and frequently refer to the 10-K report for more information. On the other hand, 10-K reports are quite extensive and often stretch up to more than 100 pages.[2]

6.1.2. Interpretation and Comparison of Environmental Results

Our findings show that the MD&A section is not associated with the environmental performance, and that the narrative environmental per-

[2]Compare Armbrust et al., 2020, p. 188.

formance does not contribute to the financial performance prediction task. For the *long-term* model (1) as well as for both *short-term* models (2a) and (2b), we find that the *env.*-classifiers are not superior to a classifier that always predicts an improvement in the environmental performance. Furthermore, for all models, we find that the *combined*-classifiers, which include the environmental performance to predict the financial performance, are not significantly better than the *direct*-classifiers nor are they superior to the corresponding majority baseline. We observe no difference in predicting the long-term environmental performance or the short-term environmental performance from the MD&A section. For both classification tasks the *env.*-classifiers do not learn a meaningful association between the environmental performance and the MD&A section. This indicates that the reports do not contain meaningful information with respect to the companies' future environmental performance.

Even for those reports where the classifiers correctly predict the environmental performance, the LIME value analysis showed that the classifiers learn no terms that indicate the environmental performance. Instead, the classifiers focus on words that coincide with the environmental performance rather than those words that truly indicate the performance. This also implies that either the environmental information is not meaningful in predicting the environmental performance or not present at all. Consequently, we searched the reports for environment-related terms. Our results show that 47% to 61% of the filings include at least one environmental search term and only 10% to 22% include more than ten of those terms. Additionally, the reports appear to be quite heterogeneous, as the 10-Q reports include roughly 10% less environmental terms than the 10-K reports and the usage varies widely among industries. The environmental word analysis shows that most companies do not report their environmental performance in the MD&A section. In addition, the reports' heterogeneity with respect to the environmental content might explain why the classifiers fail to generalise.

Our findings are in line with Doran and Quinn, 2009, who show that only a small percentage of companies include discussions on climate-related

risks and opportunities in their 10-K filings. Our findings are also in line with Eccles et al., 2012, who find that the majority of companies either include boilerplate statements or no climate-related information at all. Furthermore, our findings support research from Coburn and Cook, 2014 and DiSalvio and Dorata, 2014, which find a huge variation in company's environmental disclosures and a decline in average quality from 2010. Our research also supports arguments by the Government Accountability Office, 2020, which showed that the included ESG-related information is not always clear and useful. Furthermore, according to Cong et al., 2020 even when company's report on their environmental performance, some firms use the climate-change disclosures to obscure their environmental performance. This would also explain, why the classifiers do not learn a meaningful association between the environmental performance and the MD&A section: Poor performing companies and environmental improvers might use the same language to describe their environmental performance. Thus, the classifiers fail to learn a meaningful relationship.

Secondly, our findings indicate that the narrative environmental performance does not contribute to the financial classification task from the MD&A section. This holds for both the *long-term*-model (1) and the *short-term*-models (2a) and (2b). Adding environmental performance as an input to the classifiers predicting the financial performance resulted only in marginal improvements that are not statistically significant. Apparently, the environmental performance does neither contribute to predicting the company's own financial performance nor does it appear to be useful for investors. Consequently, the information in its current form does not support the prediction of the companies long-term performance nor does it support investors in their decision-making process.

This finding is in contrast to Matsumura et al., 2018, who find that the filings' climate-related disclosures are associated with the firms' cost of equity. However, Matsumura et al., 2018 use the entire 10-K reports and filter them manually for climate-relevant sections. Additionally, Matsumura et al., 2018 provide background knowledge by using a tailored word list. Furthermore, our findings are in contrast to Berkman et al., 2019.

While Berkman et al., 2019 find that the tone of climate-risk disclosure within 10-K filings is associated with firm value, Berkman et al., 2019 use also the entire report, a tailored word list, and pre-selected only sections with environmental information. This supports our argument that not all MD&A sections do include useful environmental information, however, other sections might. Similarly, the findings by Kölbel et al., 2020, who use only the filings' risk section, support arguments for meaningful climate-related information being contained in other parts of the filings. In addition, the findings by Kölbel et al., 2020, Berkman et al., 2019, and Matsumura et al., 2018 show that our classifiers might be improved by humanly annotation or other additional background information.

6.2. Comparison to Theory

Our findings show that the MD&A section's quasi-discretionary disclosure setting allows companies to withhold meaningful information. The discretionary disclosure setting does indeed appear to result in boilerplate disclosures. While some companies seem to provide meaningful information, other companies do not if they are not obliged to do so. This supports arguments made by Verrecchia, 2001, 1983, that is, firms withhold information if the disclosure-related cost exceeds the benefits of disclosure. Our findings suggest that particularly the disclosure of forward-looking statements related to the company's financial and environmental performance might be associated with proprietary costs. Consequently, non-disclosure or imprecise disclosure may lead to strategic advantages, especially if the information is of proprietary nature (Richarson, 2001). Additionally, our findings indicate that the long-term and short-term analysis of the MD&A section does not relate to legal liabilities, as the long- and short-term analysis mostly contains conjectures of the company.

Since we find no association between the MD&A section and environmental performance, we cannot make any definite statements regarding the voluntary disclosure theory and legitimacy theory. However, we have focused only on the section's narrative information content. According

to Hummel and Schlick, 2016, narrative information is most likely used by poor performing companies to obscure their environmental performance, while good performing companies are using numerical data to describe their environmental performance. Therefore, it is likely that those companies, which do report on their environmental performance, use the narrative information to deceive the readers. However, we find no evidence that the environmental information contributes to the future financial performance. This might support arguments by the voluntary disclosure theory that only precise and comprehensible environmental information can lead to a strategic advantage. Consequently, our results show that more research with respect to the actual narrative information content is necessary.

6.3. Limitations

Before concluding the discussion, we discuss the limitations of our study. The first limitation is that our study focuses solely on the MD&A section as part of the 10-K and 10-Q filings. In addition, we included only the QQD section in our analysis. Consequently, we cannot make statements about the informativeness of the entire 10-K and 10-Q filings. This is particularly important when discussing the environmental information and the environmental performance. The SEC guidelines allow companies to publish meaningful information also in other parts of the filing. As we pointed out, other authors find meaningful associations between the environmental information and financial variables (Berkman et al., 2019; Matsumura et al., 2018) by using the entire reports (see Subsection 6.1.1). Furthermore, authors find also meaningful associations between the financial information and market returns (Jegadeesh and Wu, 2013; Loughran and McDonald, 2011) and the financial information and earnings (F. Li, 2006) in using the entire reports. Therefore, meaningful environmental and financial information might be present in other parts of the filings. Furthermore, we focused only on the narrative information content and, thus, we neglect any potential informative numerical information in the reports.

In addition, we assumed that 10-K and 10-Q reports are comparable. While the SEC guidelines require companies to report meaningful information with respect to the financial performance and environmental performance within both filings, we observe differences in length and 10-Q reports, which often refer to 10-K reports for more information. Thus, the information content – and consequently 10-K and 10-Q reports – might not be comparable.

Another limitation is that we assumed that the change in the environmental percentile rating from Sustainalytics reflects the environmental performance of the company. However, Sustainalytics might simply not react to the information published in the MD&A section. Additionally, the rating of a company could simply change if another company improves in percentile rank. While we assumed that these changes are small and do not affect our binary encoding, this must not apply for all companies. Moreover, we focus solely on one ESG-rating provider. While empirical evidence shows that the correlation among the environmental dimension is the highest (Gibson et al., 2021), there might be substantial differences.

We also need to point out that for the *combined*-classifiers, we measure the narrative environmental information with the actual environmental performance. While we would have used the predicted environmental performance, the predicted environmental performance proves to be unreliable (see Section 5.1). Additionally, for the *short-term* model (2b), we used monthly buy-and-hold returns, although the returns only show elevated returns at the 10% significance level. Consequently, the information might be already priced-in shortly after the dissemination of the filing. However, this does not affect the results of the *long-term* model (1) or the *short-term* model (2a).

While we trained five different classifier architectures for each model and most of our classifiers use state-of-the-art NLP methods, the classifiers are subject to some *technical limitations*. If the underlying texts are quite long and heterogeneous, which we also observe for the MD&A section, the classifiers might fail to extract the meaningful information, although meaningful information is present. This is particularly true if reports do

not have a strong polarity with respect to the financial or environmental performance. The fact that the classifiers do not generalise well shows that the underlying data quality could be improved. Possible ways to overcome the generalisation issue are either to pre-filter the reports for forward-looking paragraphs or sentences, e.g., for modal language or keywords that imply forward-looking statements like "next quarter" or "in the future", or by employing a larger corpus that provides more learning material for the classifiers and lessens the effect of the reports' diversity. Furthermore, to overcome the generalisation issue, researchers could use Multi-Task Learning, i.e., training the classifier on multiple tasks to improve generalization (C. Lee et al., 2010). For instance, researchers could improve classifiers by training the classifiers on the reports' sentiment *and* the future financial performance.

While we do analyse the classifiers in detail using the LIME value analysis, LIME has also some limitations. The LIME values present only local explanations and, thus, only local interpretability. In addition, local explanations are shown to be unstable (Molnar, 2019; Alvarez-Melis and Jaakkola, 2018), i.e., if sampling around the same local data is repeated several times to generate the explanations, it can lead to different results. Moreover, the global feature importance calculated from the LIME values is influenced by the local feature weights and features which occur more often (van der Linden et al., 2019). This means that the global feature importance is affected by the amount of features per explanation and that the features, which occur more often, have a larger effect on the model predictions (van der Linden et al., 2019). Other researchers might employ different interpretation methods that are also globally stable like SHAP (SHapley Additive exPlanations) (Lundberg and Lee, 2017).

7. Conclusion

This final chapter concludes this thesis. We begin by recapitulating the key findings of our study and answering each of our research questions (Section 7.1). Next, we show the practical implications of our study (Section 7.2) and, finally, we highlight potential areas for future research (Section 7.3).

7.1. Study Recapitulation

Our research objective was threefold: First, we analysed whether corporate disclosures contain meaningful information with respect to the future corporate financial performance and the future stock price performance of a company. Second, we examined whether corporate disclosures contain meaningful environmental information to predict the company's future environmental performance. Third, we analysed whether the narrative environmental information contributes to the narrative financial information in predicting the company's financial performance. For this reason, we focused on the MD&A section of the 10-K and 10-Q filings and designed a comprehensive model that encompasses three different machine learning classifiers: A classifier predicting the future financial performance from the MD&A section, i.e., the *direct*-classifier, a classifier predicting the future environmental performance from the MD&A section, i.e. the *env.*-classifier, and a classifier predicting the financial performance using the environmental performance and the MD&A section, i.e., the *combined*-classifier. For each of these classifiers, we employed five different neural network architectures.

To separate the earnings prediction task from the stock price prediction task, we created two separate models: A *long-term* model (1), which

employed earnings to measure the financial performance, and a *short-term* model (2), using buy-and-hold returns. We further subdivided the *short-term* model (2) into a model using a four-day event window for the buy-and-hold returns, i.e., *short-term* (2a), and a model using a monthly event window, i.e., *short-term* (2b). For each of these models, we analysed the three hypotheses whether the MD&A section is associated with the financial performance, the environmental performance, and whether the environmental performance contributes to the financial prediction task. We used earnings to measure the general informativeness of the corporate filings and the buy-and-hold returns to measure the informativeness with respect to investors. To measure the environmental information content, we employed environmental percentile ratings from Sustainalytics.

In the course of deriving our hypotheses, we presented the current regulatory framework of the SEC with respect to the MD&A section. This included discussing the discretionary nature of the MD&A section as well as the required climate-disclosures. Furthermore, we discussed current research on financial and environmental disclosures including how environmental performance is measured. Then, we introduced machine learning and deep learning as well as the different Natural Language Processing (NLP) methods that we apply. We also discussed the prevailing methods for analysing text in the financial domain and explained our proposed holistic model before presenting our empirical results and discussing them in detail. In the following, we answer each research question that we posed in Section 1.4.

7.1.1. Answering the Research Questions

RQ-1: **Do corporate disclosures contain meaningful information with respect to (1) the future corporate financial performance and (2) the future stock price performance?**

Answer-1: Generally, we find that the MD&A section appears to be uninformative – uninformative with respect to the companies own financial performance and also uninformative to investors. Our research

shows that not all companies provide useful financial information in the MD&A section – at least no information that allows conclusions to be drawn about the future financial performance. We conclude that the general informativeness of the MD&A section is quite low, since we find that the companies' disclosures are not informative with respect to the companies' own financial performance. This implies that companies either do not assesses their future performance correctly or they do not disclose the information in the first place. Furthermore, our results indicate that the legal consequence for misstatements in the future long- and short-term analysis are unlikely to be severe when the disclosure is of narrative nature. Otherwise, companies would provide more meaningful information to avoid litigations.

In addition, we find that the MD&A section is not informative with respect to the long- and short-term stock price performance. This indicates that investors either do not read the filing or they do not find information related to the future performance within the MD&A section. Instead, investors seem to focus on the reports' numerical information. Accordingly, we conclude that not all companies report decision-useful information in their filings.

RQ-2: **Do corporate disclosures contain meaningful environmental information to predict the future environmental performance of the company?**

Answer-2: In the MD&A section, we find no evidence for meaningful environmental disclosures. Indeed, the environmental information within in MD&A section appears to be boilerplate. This applies for both long-term future environmental performance as well as short-term future environmental performance. While some companies include at least a few environmental terms, we also find that the disclosure among industries varies. Companies that report the most environmental information are in the "Transportation & Public Utilities" industry – on average 41% of the 10-K and 10-Q filings contain at least 10 or more environmental terms, followed by 20% in the "Mining" industry and 18% in the "Construction"

industry. All other industries – "Manufacturing" (13%), "Finance, Insurance, Real Estate" (7%), "Retail Trade" (2.6%), "Wholesale Trade" (1.5%), and "Services" (0.8%) – report less environmental terms. However, the few information the companies report is not useful to predict the future environmental performance of the company. These findings are in line with the existing literature, which shows that only a only a small percentage of companies include climate-disclosures and that the information is not always clear and useful (Government Accountability Office, 2020; Coburn and Cook, 2014; DiSalvio and Dorata, 2014; Eccles et al., 2012, i.a.).

RQ-3: **Does the narrative environmental information contribute to the narrative financial information in predicting the company's financial performance?**

Answer-3: While we do not find meaningful environmental information in the MD&A section to predict the future environmental performance, we assumed that the actual environmental performance reflects the narrative environmental performance. However, we do not find that the environmental performance contributes to the company's financial performance. This means that the merit of including the environmental information in the financial prediction task is quite low – at least considering the environmental information in its current form. This is true for predicting the company's own financial performance as well as for predicting the stock price performance. While there is research, which suggests that meaningful environmental information affecting the financial performance might be present in other parts of the filings (Kölbel et al., 2020; Berkman et al., 2019; Matsumura et al., 2018, i.a.), we find that the inclusion of environmental information in the MD&A section does not result in improved predictions of the financial performance.

7.1.2. Main Contributions

In this subsection, we highlight our main contributions to the literature. Our contributions are the following:

- We introduced a novel, holistic approach to analyse the informativeness of corporate disclosures. This approach provides a comprehensive picture of the disclosures' information content with respect to:
  - the future corporate financial performance and the future stock price performance,
  - the future environmental performance, and
  - the contribution of environmental performance to financial performance on narrative basis.
- We address the criticism of the disclosures being boilerplate (Bloomfield, 2008) – boilerplate with respect to the financial information content but also with respect to the environmental information content (Wasim, 2019; Palmiter, 2015).
- We contribute to the discussion on the informational merit of the SEC's climate-change disclosure guideline (SEC, 2010), i.e., whether the current disclosure guideline elicits sufficient material environmental information.
- Our research also contributes to the existing research on environmental disclosures. While most research focusing on environmental disclosures applies disclosures scores, measuring the quantity of information published (Cong et al., 2020; Reverte, 2016; Clarkson et al., 2008, i.a.), or straightforward word-count based measures (Berkman et al., 2019; Matsumura et al., 2018; Verbeeten et al., 2016, i.a.), we focus on the actual narrative information content by using state-of-the-art NLP methods.

- We provide researchers with an introduction to machine and deep learning and introduce modern NLP methods to the financial domain.

7.2. Practical Implications

In this section, we address the practical implications of our results. As our findings indicate that the MD&A section is uninformative with respect to the company's future financial performance, this implies that companies either do not correctly assess their future performance or that they do not disclose the relevant information. While the MD&A section might have informational merit with respect to the company's past performance, the SEC explicitly states that companies should include discussions on prospective matters and material trends (SEC, 2003). Consequently, the discretionary nature of the disclosures might cause companies to not disclose the required information if the cost of disclosure exceeds the benefits. However, investors as well as other stakeholders would benefit from more informative disclosures. Therefore, we argue that the SEC might consider more explicit guidance for the MD&A section. For example, the SEC might consider providing a numerical threshold as to when a matter is considered material. This would provide companies and investors with a number to evaluate disclosure or non-disclosure.

Furthermore, we find that the filings are quite inconsistent. In particular, the 10-Q filings are quite brief and often refer to the 10-K reports for more information. In addition, we find that companies often report the MD&A section's content under another headline of the filings. Therefore, the SEC might consider more explicit guidelines with respect to the appearance of the disclosures. Nevertheless, the SEC has continuously undertaken steps to improve filings and make them more accessible.[1]

Our findings with respect to the environmental performance and also the additional analysis with respect to environment-related terms (see

[1] Examples include the use of Markup Language and reporting in the Inline XBRL format (SEC, 2018).

Section 5.5) support arguments for a clear and coherent ESG disclosure framework. While we see that investors ask for more meaningful narrative ESG information (WBCSD, 2018, p. 7), companies are missing the guidance to include the ESG information. We argue that it would be to the benefit of all market participants, if the SEC provides more explicit guidance on what information companies must report and how to report that information in a uniform manner. This includes precise instructions on the required information that needs to be included in each section as well as the calculation, presentation, and description of ESG-related numbers. If companies do not report on the environmental information in a uniform manner, the usefulness of the information is quite limited.

7.3. Future Research

This last section presents an outlook as well as recommendations for future research. The Local Interpretable Model-agnostic Explanations (LIME) analysis has shown that our approach can be indeed useful and successful, and that the classifiers learn at least some relevant information. While standard word list-based approaches provide only insights into existing knowledge, our approach can provide new insights on unknown patterns. Consequently, in using end-to-end neural networks researchers could learn more about the inherent information and the reasons of disclosure.

As we have discussed in Section 6.3, the contradiction with existing, word-count-based research (Loughran and McDonald, 2011; Jegadeesh and Wu, 2013; F. Li, 2006, i.a.) points towards the technical limitations of our classifiers. We observed that the classifiers fail to generalize on the underlying data, i.e., on the MD&A section of the 10-K and 10-Q reports, despite applying various regularization techniques and state-of-the-art classifiers. The reason lies in the heterogeneity of the MD&A sections. We find that the MD&A sections vary greatly in size and content, especially the 10-K and 10-Q filings among themselves. Long filings with weak polarity complicate the classifiers' training processes. Therefore, classifiers without additional background knowledge, e.g., in the form of

word lists, have difficulties finding meaningful information even if the information might be present.

Based on the results of previous research, we propose to improve our classifiers through pre-filtering the sections with tailored word lists (Berkman et al., 2019; Loughran and McDonald, 2011, i.a.), humanly annotating sentences for the classifiers (Kölbel et al., 2020; F. Li, 2010b), pre-filtering the sections manually (Matsumura et al., 2018), or by focusing on sentences with modal language and other keywords (Muslu et al., 2015; F. Li, 2010b). In focusing on paragraphs with on modal language, researchers could improve the forward-looking nature of the classifiers, as modal language could be an indication of forward-looking statements. Additionally, researchers could focus on those paragraphs or sentences, which include keywords that imply forward-looking statements, e.g., "next quarter" or "in the future". Considering the environmental information, researchers could also filter to paragraphs or sentences, which address climate-related issues, e.g., by pre-filtering the texts for environmental search terms. Researchers could use terms similar to our environmental search term list (see Subsection 5.5). Consequently, we propose that future researchers may improve our preprocessing steps to create more successful classifiers.

To overcome the generalisation issue, researchers could also use Multi-Task Learning, i.e., training the classifier on multiple tasks to improve generalization (C. Lee et al., 2010). For instance, researchers could improve the classifiers by training the classifiers on the reports' sentiment *and* the future financial performance. Another possible solution could be a bigger corpus that provides more learning material for the classifier and that lessens the effect of the reports' diversity.

Furthermore, future research might use different ESG rating providers to measure the environmental performance. The usage of just one ESG provider is error prone. Ratings for one and the same company tend to vary among different data vendors. To better understand the implications of our results, future studies could also train the classifiers directly on the company's carbon emission, waste generated, or water usage.

We belief that future research might build on our holistic approach

to analyse disclosures with respect to the financial performance and environmental performance as well as the interaction of financial and environmental performance. Only by obtaining an encompassing picture of financial and environmental performance, we can understand what information is truly material and gain new insights into the usefulness of corporate disclosures. Therefore, we belief that applications of Natural Language Processing in the financial domain are a promising and exiting field for future research.

Bibliography

Abu Bakar, A. S. and Ameer, R. (2011). Readability of Corporate Social Responsibility communication in Malaysia. *Corporate Social Responsibility and Environmental Management*, *18*(1), 50–60.

Adhikari, A., Ram, A., Tang, R., Hamilton, W. L. and Lin, J. (2020). Exploring the Limits of Simple Learners in Knowledge Distillation for Document Classification with DocBERT. *Proceedings of the 5th Workshop on Representation Learning for NLP*, 72–77. https://doi.org/10.18653/v1/2020.repl4nlp-1.10

Alammar, J. (2019). *A visual guide to using BERT for the first time*. Retrieved January 9, 2020, from http://jalammar.github.io/a-visual-guide-to-using-bert-for-the-first-time/

Al-Tuwaijri, S. A., Christensen, T. E. and Hughes Ii, K. (2004). The relations among environmental disclosure, environmental performance, and economic performance: A simultaneous equations approach. *Accounting, organizations and society*, *29*(5-6), 447–471.

Alvarez-Melis, D. and Jaakkola, T. S. (2018). Towards Robust Interpretability with Self-Explaining Neural Networks. *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, 7786–7795.

Amel-Zadeh, A. and Serafeim, G. (2018). Why and How Investors Use ESG Information: Evidence from a Gglobal Survey. *Financial Analysts Journal*, *74*(3), 87–103.

Armbrust, F., Schäfer, H. and Klinger, R. (2020). A Computational Analysis of Financial and Environmental Narratives within Financial Reports and its Value for Investors. *Proceedings of the 1st*

Joint Workshop on Financial Narrative Processing and MultiLing Financial Summarisation, 181–194.

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

Bahdanau, D., Cho, K. and Bengio, Y. Neural machine translation by jointly learning to align and translate [3rd International Conference on Learning Representations, ICLR 2015 ; Conference date: 07-05-2015 Through 09-05-2015]. English (US). In: 3rd International Conference on Learning Representations, ICLR 2015 ; Conference date: 07-05-2015 Through 09-05-2015. 2015, January.

Baier, P., Berninger, M. and Kiesel, F. (2020). Environmental, social and governance reporting in annual reports: A textual analysis. *Financial Markets, Institutions & Instruments*, *29*(3), 93–118.

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

Baxamusa, M., Jalal, A. and Jha, A. (2018). It pays to partner with a firm that writes annual reports well. *Journal of Banking & Finance*, *92*, 13–34.

Bengio, Y., Ducharme, R., Vincent, P. and Janvin, C. (2003). A neural probabilistic language model. *J. Mach. Learn. Res.*, *3*(null), 1137–1155.

Bennington, A. (2020). *Recommendation from the Investor-as-Owner Subcommittee of the SEC Investor Advisory Committee Relating to ESG Disclosure*. Retrieved December 22, 2020, from https://corpgov.law.harvard.edu/2020/05/28/recommendation-from-the-investor-as-owner-subcommittee-of-the-sec-investor-advisory-committee-relating-to-esg-disclosure/

Berg, F., Koelbel, J. F. and Rigobon, R. (2020). *Aggregate Confusion: The Divergence of ESG Ratings*. Retrieved January 30, 2021, from https://dx.doi.org/10.2139/ssrn.3438533

Berg-Kirkpatrick, T., Burkett, D. and Klein, D. (2012). An empirical investigation of statistical significance in NLP. *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, 995–1005. https://www.aclweb.org/anthology/D12-1091

Berkman, H., Jona, J. and Soderstrom, N. S. (2019). *Firm-Specific Climate Risk and Market Valuation*. Retrieved February 18, 2021, from https://dx.doi.org/10.2139/ssrn.2775552

Bertsekas, D. P. and Tsitsiklis, J. N. (1996). *Neuro-dynamic programming*. Athena Scientific.

Bird, S., Klein, E. and Loper, E. (2009). *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O'Reilly Media, Inc.

Blei, D. M., Ng, A. Y. and Jordan, M. I. (2003). Latent dirichlet allocation. *the Journal of machine Learning research*, *3*, 993–1022.

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

Brown, S. V. and Tucker, J. W. (2011). Large-Sample Evidence on Firms' Year-over-Year MD&A Modifications. *Journal of Accounting Research*, *49*(2), 309–346.

Brownlee, J. (2018). *Difference between a batch and an epoch in a neural network*. Retrieved July 20, 2018, from https://machinelearningmastery.com/difference-between-a-batch-and-an-epoch/

Butler, M. and Kešelj, V. (2009). Financial Forecasting Using Character N-Gram Analysis and Readability Scores of Annual Reports. *Canadian Conference on Artificial Intelligence*, 39–51.

Cambridge Dictionary. (2020). *Polysemy*. Retrieved April 8, 2020, from https://dictionary.cambridge.org/dictionary/english/polysemy

Campbell, J. L., Chen, H., Dhaliwal, D. S., Lu, H.-m. and Steele, L. B. (2014). The information content of mandatory risk factor disclosures in corporate filings. *Review of Accounting Studies*, *19*(1), 396–455.

Cannon, J. N., Ling, Z., Wang, Q. and Watanabe, O. V. (2020). 10-K Disclosure of Corporate Social Responsibility and Firms' Competitive Advantages. *European Accounting Review*, *29*(1), 85–113.

CDSB. (2019). *CDSB Framework for reporting environmental & climate change information*. Retrieved January 15, 2021, from https://www.cdsb.net/sites/default/files/cdsb_framework_2019_v2.2.pdf

Center, R. (2014). *10-K Filings Guide: Traps and Mistakes*. Retrieved April 9, 2020, from https://businessjournalism.org/2014/02/10-k-filings-guide-traps-and-mistakes/

CFA Institute. (2017). *ESG Survey: Global Perceptions of Environmental, Social, and Governance Issues in Investing*. Retrieved January 6, 2021, from https://www.cfainstitute.org/en/research/survey-reports/esg-survey-2017

Chakrabarty, B., Seetharaman, A., Swanson, Z. and Wang, X. (2018). Management Risk Incentives and the Readability of Corporate Disclosures. *Financial Management*, *47*(3), 583–616.

Chatterji, A. K., Durand, R., Levine, D. I. and Touboul, S. (2016). Do ratings of firms converge? Implications for managers, investors and strategy researchers. *Strategic Management Journal*, *37*(8), 1597–1614.

Chaturvedi, I., Ong, Y.-S., Tsang, I. W., Welsch, R. E. and Cambria, E. (2016). Learning word dependencies in text by means of a deep recurrent belief network. *Knowledge-Based Systems*, *108*, 144–154.

Cho, C. H. and Patten, D. M. (2007). The role of environmental disclosures as tools of legitimacy: A research note. *Accounting, organizations and society*, *32*(7-8), 639–647.

Cho, C. H., Patten, D. M. and Roberts, R. W. (2006). Corporate political strategy: An examination of the relation between political expenditures, environmental performance, and environmental disclosure. *Journal of Business Ethics*, *67*(2), 139–154.

Cho, K., van Merriënboer, B., Bahdanau, D. and Bengio, Y. (2014). On the Properties of Neural Machine Translation: Encoder–Decoder Approaches. *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, 103–111. https://doi.org/10.3115/v1/W14-4012

Chollet, F. et al. (2015). Keras.

Christensen, D., Serafeim, G. and Sikochi, A. (2019). Why is Corporate Virtue in the Eye of The Beholder? The Case of ESG Ratings.

Chung, J., Gulcehre, C., Cho, K. and Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. *NIPS 2014 Workshop on Deep Learning, December 2014*.

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

Clarkson, P. M., Kao, J. L. and Richardson, G. D. (1999). Evidence that management discussion and analysis (MD&A) is a part of a firm's overall disclosure package. *Contemporary accounting research*, *16*(1), 111–134.

Clarkson, P. M., Li, Y., Richardson, G. D. and Vasvari, F. P. (2008). Revisiting the Relation Between Environmental Performance and Environmental Disclosure: An Empirical Analysis. *Accounting, organizations and society*, *33*(4-5), 303–327.

Clarkson, P. M., Overell, M. B. and Chapple, L. (2011). Environmental reporting and its relation to corporate environmental performance. *Abacus*, *47*(1), 27–60.

Coburn, J. and Cook, J. (2014). *Cool response: The SEC and corporate climate change Reporting*. Retrieved February 20, 2021, from https://www.ceres.org/sites/default/files/reports/2017-03/Ceres_SECguidance-append_020414_web.pdf

Cohen, J. R., Gaynor, L. M., Holder-Webb, L. L. and Montague, N. (2008). Management's Discussion and Analysis: Implications for

Audit Practice and Research. *Current Issues in Auditing*, *2*(2), A26–A35.

Cole, C. J. and Jones, C. L. (2004). The usefulness of md&a disclosures in the retail industry. *Journal of Accounting, Auditing & Finance*, *19*(4), 361–388.

Cong, Y., Freedman, M. and Park, J. D. (2020). Mandated Greenhouse Gas Emissions and Required SEC Climate Change Disclosures. *Journal of Cleaner Production*, *247*, 119111.

Corporate Register. (2020). Retrieved April 7, 2020, from https://www.corporateregister.com/

Dahl, G. E., Sainath, T. N. and Hinton, G. E. (2013). Improving deep neural networks for LVCSR using rectified linear units and dropout. *2013 IEEE international conference on acoustics, speech and signal processing*, 8609–8613.

Dai, A. M. and Le, Q. V. (2015). Semi-supervised sequence learning. *Proceedings of the 28th International Conference on Neural Information Processing Systems-Volume 2*, 3079–3087.

Datamaran. (2020). *Global Insights Report: The Rise of ESG Regulations*. Retrieved January 28, 2021, from https://www.datamaran.com/global-insights-report/

Datamaran. (2021). *Getting started with double materiality: A 5-step plan*. Retrieved September 19, 2021, from https://www.datamaran.com/double-materiality-ebook/

Datta, A., Sen, S. and Zick, Y. (2016). Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems. *2016 IEEE symposium on security and privacy (SP)*, 598–617.

Davies, P. A., Dudek, P. M. and Wyatt, K. S. (2020). Recent Developments in ESG Reporting, 161–179.

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

De Villiers, C. and Van Staden, C. J. (2006). Can less environmental disclosure have a legitimising effect? Evidence from Africa. *Accounting, organizations and society*, *31*(8), 763–781.

Deegan, C. (2002). The legitimising effect of social and environmental disclosures - a theoretical foundation. *Accounting, Auditing & Accountability Journal*, 282–311.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K. and Harshman, R. (1990). Indexing by Latent Semantic Analysis. *Journal of the American society for information science*, *41*(6), 391–407.

Depoers, F., Jeanjean, T. and Jérôme, T. (2016). Voluntary Disclosure of Greenhouse Gas Emissions: Contrasting the Carbon Disclosure Project and Corporate Reports. *Journal of Business Ethics*, *134*(3), 445–461.

Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *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)*, 4171–4186. https://doi.org/10.18653/v1/N19-1423

Dhaliwal, D., Li, O. Z., Tsang, A. and Yang, Y. G. (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.

Dhaliwal, D. S., Li, O. Z., Tsang, A. and Yang, Y. G. (2011). Voluntary nonfinancial disclosure and the cost of equity capital: The initiation of corporate social responsibility reporting. *The accounting review*, *86*(1), 59–100.

DiSalvio, J. and Dorata, N. T. (2014). SEC guidance on climate change risk disclosures: An assessment of firm and market responses. *Accounting for the environment: More talk and little progress*. Emerald Group Publishing Limited.

Doran, K. L. and Quinn, E. L. (2009). Climate change risk disclosure: A sector by sector analysis of SEC 10-K filings from 1995-2008. *North Carolina Journal of International Law and Commercial Regulation*, *34*.

Doyle, T. M. (2018). *Ratings that don't rate: The subjective world of ESG ratings agencies*. Retrieved February 19, 2021, from https://accfcorpgov.org/wp-content/uploads/2018/07/ACCF_RatingsESGReport.pdf

Dror, R., Peled-Cohen, L., Shlomov, S. and Reichart, R. (2020). Statistical Significance Testing for Natural Language Processing. *Synthesis Lectures on Human Language Technologies*, *13*(2), 1–116.

Dumais, S. T., Furnas, G. W., Landauer, T. K., Deerwester, S. and Harshman, R. (1988). Using latent semantic analysis to improve access to textual information. *Proceedings of the SIGCHI conference on Human factors in computing systems*, 281–285.

Dumay, J., Bernardi, C., Guthrie, J. and Demartini, P. (2016). Integrated Reporting: A Structured Literature Review. *Accounting Forum*, *40*(3), 166–185.

Dyck, A., Lins, K. V., Roth, L. and Wagner, H. F. (2019). Do Institutional Investors drive Corporate Social Responsibility? International Evidence. *Journal of Financial Economics*, *131*(3), 693–714.

Dye, R. A. (1985). Strategic accounting choice and the effects of alternative financial reporting requirements. *Journal of accounting research*, 544–574.

Dye, R. A. (2001). An evaluation of "essays on disclosure" and the disclosure literature in accounting. *Journal of accounting and economics*, *32*(1-3), 181–235.

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

Eccles, R. G. and Stroehle, J. C. (2018). *Exploring social origins in the construction of ESG measures*. Retrieved September 15, 2018, from https://dx.doi.org/10.2139/ssrn.3212685

Efron, B. and Tibshirani, R. J. (1993). *An introduction to the bootstrap. 57 boca raton*. FL: CRC Press.

Electronic Arts Inc. (2014). *SEC 10-Q Form for Electronic Arts Inc. filed 2014-11-04*. Retrieved August 23, 2021, from https://www.sec.gov/Archives/edgar/data/0000712515/000071251514000063/ea9302014-q2fy1510qdoc.htm

Ernst & Young. (2012). Now is the time to address disclosure overload [(accessed: 2020-12-19)].

Escrig-Olmedo, E., Fernández-Izquierdo, M. Á., Ferrero-Ferrero, I., Rivera-Lirio, J. M. and Muñoz-Torres, M. J. (2019). Rating the raters: Evaluating how esg rating agencies integrate sustainability principles. *Sustainability*, *11*(3), 915.

Esty, D. and Cort, T. (2017). Corporate sustainability metrics: What investors need and don't get. *Journal of Environmental Investing*, *8*(1), 11–53.

Esty, D., Cort, T., Strauss, D., Kristina, W. and Yeargain, T. (2020). *Toward Enhanced Sustainability Disclosure: Identifying Obstacles to Broader and More Actionable ESG Reporting*. Retrieved December 19, 2020, from https://envirocenter.yale.edu/toward-enhanced-sustainability-disclosure-identifying-obstacles-broader-and-more-actionable-esg

European Commission. (2020). Regulation (EU) 2020/852 of the European Parliament and of the Council of 18 June 2020 on the establishment of a framework to facilitate sustainable investment, and amending Regulation (EU) 2019/2088.

European Commission. (2021a). Factsheet: EU Sustainable Finance: April Package.

European Commission. (2021b). Proposal for a DIRECTIVE OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL amending Directive 2013/34/EU, Directive 2004/109/EC, Directive 2006/43/EC and Regulation (EU) No 537/2014, as regards corporate sustainability reporting COM/2021/189 final.

Farewell, S., Fisher, I. and Daily, C. (2014). The lexical footprint of sustainability reports: A pilot study of readability. *American Accounting Association Annual Meeting and Conference on Teaching and Learning in Accounting*.

FASB. (2010). Conceptual Framework for Financial Reporting [Statement of Financial Accounting Concepts No. 8, September 2010].

Federation of European Accountants. (2016). EU Directive on Disclosure of Non-Financial and Diversity Information. Achieving Good Quality and Consistent Reporting.

Feldman, R., Govindaraj, S., Livnat, J. and Segal, B. (2010). Management's tone change, post earnings announcement drift and accruals. *Revieew of Accounting Studies*, *15*, 915–953.

Feldman, R., Govindaraj, S., Livnat, J. and Segal, B. (2008). *The incremental information content of tone change in management discussion and analysis*. Retrieved January 28, 2021, from https://dx.doi.org/10.2139/ssrn.1126962

Feng, S. and Gao, L. S. (2020). The Verbal Tone in Mandatory Environmental Disclosures: Evidence from Changes in Disclosures Following SEC Guidance. *Social and Environmental Accountability Journal*, *40*(2), 116–139.

Fisher, I. E., Garnsey, M. R. and Hughes, M. E. (2016). Natural Language Processing in Accounting, Auditing and Finance: A Synthesis of the Literature with a Roadmap for Future Research. *Intelligent Systems in Accounting, Finance and Management*, *23*(3), 157–214.

Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology*, *32*(3), 221–233. https://doi.org/10.1037/h0057532

Freedman, M. and Park, J. D. (2014). Mandated Climate Change Disclosures by Firms Participating in the Regional Greenhouse Gas Initiative. *Social and Environmental Accountability Journal*, *34*(1), 29–44.

Freeman, R. E. (2010). *Strategic Management: A Stakeholder Approach*. Cambridge University Press.

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

Frunza, O. (2020). Information Extraction from Federal Open Market Committee Statements. *Proceedings of the 1st Joint Workshop on Financial Narrative Processing and MultiLing Financial Summarisation*, 195–203. https://www.aclweb.org/anthology/2020.fnp-1.32

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

Gibson, R., Krueger, P., Riand, N. and Schmidt, P. S. (2021). ESG Rating Disagreement and Stock Returns. *Financial Analyst Journal*, *77*(4).

Glorot, X., Bordes, A. and Bengio, Y. (2011). Deep sparse rectifier neural networks. *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, 315–323.

Goodfellow, I., Bengio, Y. and Courville, A. (2017). *Deep learning* (Vol. 1). MIT Press, Cambridge, MA.

Government Accountability Office. (2020). *Disclosure of Environmental, Social, and Governance Factors and Options to Enhance Them*. Retrieved December 19, 2020, from https://www.gao.gov/assets/710/707949.pdf

GRI. (2016). *GRI 101: Foundation 2016*. Retrieved January 15, 2021, from https://www.globalreporting.org/standards/media/1036/gri-101-foundation-2016.pdf

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

Hackett, D., Demas, R., Sanders, D., Wicha, J. and Fowler, A. (2020). Growing ESG Risks: The Rise of Litigation. *Envtl. L. Rep.*, *50*, 10849.

Hahn, R., Reimsbach, D. and Schiemann, F. (2015). Organizations, climate change, and transparency: Reviewing the literature on carbon disclosure. *Organization & Environment*, *28*(1), 80–102.

Hamilton, W. L., Clark, K., Leskovec, J. and Jurafsky, D. (2016). Inducing domain-specific sentiment lexicons from unlabeled corpora. *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, *2016*, 595.

Hanks, P. (2005). Lexicography. In R. Mitkov (Ed.), *The oxford handbook of computational linguistics* (pp. 48–69). Oxford University Press.

Henry, E. (2008). Are investors influenced by how earnings press releases are written? *The Journal of Business Communication (1973)*, *45*(4), 363–407.

Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I. and Salakhutdinov, R. R. (2012). Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580*.

Hirji, Z. (2013). *Most U.S. companies ignoring SEC rule to disclose climate risks*. Retrieved June 8, 2020, from https://insideclimatenews.org/news/20130919/most-us-companies-ignoring-sec-rule-disclose-climate-risks

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

Hoffmann, T. (2011). *Unternehmerische Nachhaltigkeitsberichterstattung: Eine Analyse des GRI G3. 1-Berichtsrahmens* (Vol. 30). BoD–Books on Demand.

Hong, H. and Kostovetsky, L. (2012). Red and blue investing: Values and finance. *Journal of Financial Economics*, *103*(1), 1–19.

Honnibal, M., Montani, I., Van Landeghem, S. and Boyd, A. (2020). spaCy: Industrial-strength Natural Language Processing in Python [https://spacy.io/].

Howard, J. and Ruder, S. (2018). Universal language model fine-tuning for text classification. *Proceedings of the 56th Annual Meeting of*

the Association for Computational Linguistics (Volume 1: Long Papers), 328–339. https://doi.org/10.18653/v1/P18-1031

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

Hummel, K. and Schlick, C. (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.

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

IIRC. (2013). *The International <IR> Framework*. Retrieved January 15, 2021, from https://integratedreporting.org/wp-content/uploads/2015/03/13-12-08-THE-INTERNATIONAL-IR-FRAMEWORK-2-1.pdf

Ilhan, E., Krueger, P., Sautner, Z. and Starks, L. T. (2019). Institutional investors' views and preferences on climate risk disclosure [Swiss Finance Institute Research Paper, No. 19-66].

IMP. (2020). *Statement of intent to work together towards comprehensive corporate reporting*. Retrieved January 28, 2021, from https://www.impactmanagementproject.com/structured-network/statement-of-intent-to-work-together-towards-comprehensive-corporate-reporting/

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

Jones, K. S. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*.

Jurafsky, D. and Martin, J. H. (2020). *Speech and language processing (thrid edition draft)*. https://web.stanford.edu/~jurafsky/slp3/ed3book.pdf.

Kahn, R. (2019a). *How do Tensorflow and Keras implement Binary Classification and the Binary Cross-Entropy function?* Retrieved November 14, 2019, from https://rafayak.medium.com/how-do-tensorflow-and-keras-implement-binary-classification-and-the-binary-cross-entropy-function-e9413826da7

Kahn, R. (2019b). *Nothing but numpy: Understanding & creating binary classification neural networks with computational graphs from scratch*. Retrieved November 14, 2019, from https://towardsdatascience.com/nothing-but-numpy-understanding-creating-binary-classification-neural-networks-with-e746423c8d5c

Kalchbrenner, N., Grefenstette, E. and Blunsom, P. (2014). A Convolutional Neural Network for Modelling Sentences. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 655–665. https://doi.org/10.3115/v1/P14-1062

Kamath, U., Liu, J. and Whitaker, J. (2019). *Deep Learning for NLP and Speech Recognition* (Vol. 84). Springer.

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

Kelly, E. F. and Stone, P. J. (1975). *Computer Recognition of English Word Senses* (Vol. 13). North-Holland.

Kim, Y. (2014). Convolutional Neural Networks for Sentence Classification. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1746–1751. https://doi.org/10.3115/v1/D14-1181

Kinderman, D. (2015). Corporate Social Responsibility – Der Kampf um die EU-Richtlinie. *Nomos Verlagsgesellschaft mbH & Co. KG*, *68*(8), 613–621.

Kingma, D. P. and Ba, J. (2015). Adam: A Method for Stochastic Optimization [Contribution to International Conference on Learning Representations, May 7-9, 2015, San Diego]. *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*.

Kohavi, R. et al. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. *Ijcai*, *14*(2), 1137–1145.

Kölbel, J. F., Leippold, M., Rillaerts, J. and Wang, Q. (2020). *Does the CDS market reflect regulatory climate risk disclosures?* Retrieved August 10, 2020, from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3616324

Kothari, S. P., Li, X. and Short, J. E. (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.

Kravet, T. and Muslu, V. (2013). Textual risk disclosures and investors' risk perceptions. *Review of Accounting Studies*, *18*(4), 1088–1122.

Krüger, P. (2015). Corporate goodness and shareholder wealth. *Journal of financial economics*, *115*(2), 304–329.

Kumar, R. and Weiner, A. (2019). *The ESG Data Challenge*. Retrieved September 21, 2021, from https://www.ssga.com/investment-topics/environmental-social-governance/2019/03/esg-data-challenge.pdf

Lansford, B. (2006). *Strategic coordination of good and bad news disclosures: The case of voluntary patent disclosures and negative earnings surprises*. Retrieved August 10, 2020, from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=830705

Lawrence, A. (2013). Individual investors and financial disclosure. *Journal of Accounting and Economics*, *56*(1), 130–147.

Lee, C., Jung, S., Kim, K. and Lee, G. G. (2010). Hybrid approach to robust dialog management using agenda and dialog examples. *Computer Speech & Language*, *24*(4), 609–631. https://doi.org/https://doi.org/10.1016/j.csl.2009.08.003

Lee, Y.-J. (2012). The effect of quarterly report readability on information efficiency of stock prices. *Contemporary Accounting Research*, *29*(4), 1137–1170.

Lehavy, R., Li, F. and Merkley, K. (2011). The Effect of Annual Report Readability on Analyst Following and the Properties of Their Earnings Forecasts. *The Accounting Review*, *86*(3), 1087–1115.

Leuz, C. and Wysocki, P. D. (2008). *Economic consequences of financial reporting and disclosure regulation: A review and suggestions for future research*. Retrieved July 19, 2020, from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1105398

Li, F.-F. (2019). CS231n: Convolutional Neural Networks for Visual Recognition. Chapter: Neural Networks and Backpropagation [Stanford University. Lecture. http://cs231n.stanford.edu/slides/2019/cs231n_2019_lecture04.pdf].

Li, F. (2006). *Do stock market investors understand the risk sentiment of corporate annual reports?* Retrieved June 17, 2020, from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=898181

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

Li, F. (2010a). Textual Analysis of Corporate Disclosures: Survey of the Literature. *Journal of accounting literature*, *29*, 143–165.

Li, F. (2010b). The Information Content of Forward-Looking Statements in Corporate Filings – A Naïve Bayesian Machine Learning Approach. *Journal of Accounting Research*, *48*(5), 1049–1102.

Lo, K., Ramos, F. and Rogo, R. (2017). Earnings management and annual report readability. *Journal of Accounting and Economics*, *63*(1), 1–25.

Loughran, T. and McDonald, B. (2009). Plain English, Readability, and 10-K Filings. *University of Notre Dame*.

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

Loughran, T. and McDonald, B. (2014). Measuring readability in financial disclosures. *The Journal of Finance*, *69*(4), 1643–1671.

Loughran, T. and McDonald, B. (2016). Textual Analysis in Accounting and Finance: A Survey. *Journal of Accounting Research*, *54*(4), 1187–1230.

Loughran, T. and McDonald, B. (2019). *Textual Analysis in Finance*. Retrieved August 14, 2020, from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3470272

Loughran, T., McDonald, B. and Yun, H. (2009). A wolf in sheep's clothing: The use of ethics-related terms in 10-K reports. *Journal of Business Ethics*, *89*(1), 39–49.

Luccioni, A. and Palacios, H. (2019). Using Natural Language Processing to Analyze Financial Climate Disclosures. *Proceedings of the 36th International Conference on Machine Learning, Long Beach, California*.

Luhn, H. P. (1957). A Statistical Approach to Mechanized Encoding and Searching of Literary Information. *IBM Journal of research and development*, *1*(4), 309–317.

Lundberg, S. M. and Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. *Proceedings of the 31st international conference on neural information processing systems*, 4768–4777.

Martin 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, Jia, Y., Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, ... And Xiaoqiang Zheng. (2015). TensorFlow: Large-scale machine learning on heterogeneous systems [Software available from tensorflow.org]. https://www.tensorflow.org/

Matsumura, E. M., Prakash, R. and Vera-Muñoz, S. C. (2014). Firm-Value Effects of Carbon Emissions and Carbon Disclosures. *The Accounting Review*, *89*(2), 695–724.

Matsumura, E. M., Prakash, R. and Vera-Muñoz, S. C. (2018). *Capital Market Expectations of Risk Materiality and the Credibility of*

Managers' Risk Disclosure Decisions. Retrieved February 18, 2021, from https://dx.doi.org/10.2139/ssrn.2983977

McDonald, B. (2019). *Software Repository for Accounting and Finance*. Retrieved April 9, 2020, from https://sraf.nd.edu/

McEnery, T. (2005). Corpus Linguistics. In R. Mitkov (Ed.), *The oxford handbook of computational linguistics* (pp. 449–463). Oxford University Press.

McWilliams, A. and Siegel, D. (1997). Event studies in management research: Theoretical and empirical issues. *Academy of management journal*, *40*(3), 626–657.

Merkley, K. J. (2014). Narrative disclosure and earnings performance: Evidence from R&D disclosures. *The Accounting Review*, *89*(2), 725–757.

Mikheev, A. (2005). Text Segmentation. In R. Mitkov (Ed.), *The oxford handbook of computational linguistics* (pp. 48–69). Oxford University Press.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S. and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 3111–3119.

Mikolov, T., Yih, W.-t. and Zweig, G. (2013). Linguistic Regularities in Continuous Space Word Representations. *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 746–751. https://www.aclweb.org/anthology/N13-1090

Minnaar, A. (2015). *Word2Vec Tutorial Part I: The Skip-Gram Model*. Retrieved March 17, 2021, from http://alexminnaar.com/2015/04/12/word2vec-tutorial-skipgram.html

Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D. and Riedmiller, M. (2013). Playing Atari with Deep Reinforcement Learning. *arXiv preprint arXiv:1312.5602*.

Molnar, C. (2019). *Interpretable machine learning: A guide for making black box models explainable* [https://christophm.github.io/interpretable-ml-book/].

Moreno-Sandoval, A., Gisbert, P. A. H. A., Guerrero, M. and Montoro, H. (2019). Tone Analysis in Spanish Financial Reporting Narratives. *Proceedings of the Second Financial Narrative Processing Workshop (FNP 2019)*, 42–50.

Mulyar, A., Schumacher, E., Rouhizadeh, M. and Dredze, M. (2019). Phenotyping of Clinical Notes with Improved Document Classification Models using Contextualized Neural Language Models. *arXiv preprint arXiv:1910.13664*.

Muslu, V., Radhakrishnan, S., Subramanyam, K. and Lim, D. (2015). Forward-Looking MD&A Disclosures and the Information Environment. *Management Science*, *61*, 931–948. https://doi.org/10.1287/mnsc.2014.1921

Nair, V. and Hinton, G. E. (2010). Rectified Linear Units Improve Restricted Boltzmann Machines. *Icml*.

Nguyen, D. (2018). Comparing Automatic and Human Evaluation of Local Explanations for Text Classification. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 1069–1078. https://doi.org/10.18653/v1/N18-1097

O'donovan, G. (2002). Environmental Disclosures in the Annual Report. *Accounting, Auditing & Accountability Journal*.

Othman, I. W., Hasan, H., Tapsir, R., Rahman, N. A., Tarmuji, I., Majdi, S., Masuri, S. A. and Omar, N. (2012). Text readability and fraud detection. *2012 IEEE Symposium on Business, Engineering and Industrial Applications*, 296–301.

Palmiter, A. R. (2015). Climate change disclosure: A failed SEC mandate [https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2639181].

Pang, B., Lee, L. and Vaithyanathan, S. (2002). Thumbs up? Sentiment Classification using Machine Learning Techniques. *EMNLP*.

Patten, D. M. (2002). The relation between environmental performance and environmental disclosure: A research note. *Accounting, organizations and Society*, *27*(8), 763–773.

Pava, M. L. and Epstein, M. J. (1993). How good is MD&A as an investment tool? *Journal of Accountancy*, *175*(3), 51.

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

Pennington, J., Socher, R. and Manning, C. D. (2014). GloVe: Global Vectors for Word Representation. *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 1532–1543.

Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K. and Zettlemoyer, L. (2018). Deep Contextualized Word Representations. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 2227–2237. https://doi.org/10.18653/v1/N18-1202

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K. and Zettlemoyer, L. (2018). Deep Contextualized Word Representations. *Proceedings of NAACL-HLT*, 2227–2237.

Plumlee, M., Brown, D., Hayes, R. M. and Marshall, R. S. (2015). Voluntary Environmental Disclosure Quality and Firm Value: Further Evidence. *Journal of accounting and public policy*, *34*(4), 336–361.

Porter, M. F. (1980). An Algorithm for Suffix Stripping. *Program*.

Radford, A., Narasimhan, K., Salimans, T. and Sutskever, I. (2018). Improving Language Understanding by Generative Pre-Training

[https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf].

Raschka, S. (2015). *Python Machine Learning*. Packt Publishing Ltd.

Reverte, C. (2016). Corporate Social Responsibility Disclosure and Market Valuation: Evidence from Spanish Listed Firms. *Review of Managerial Science*, *10*(2), 411–435.

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

Ribeiro, M. T., Singh, S. and Guestrin, C. (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144. https://doi.org/10.1145/2939672.2939778

Richarson, S. (2001). Discretionary Disclosure: A Note. *Abacus*, *37*(2), 233–247.

Riezler, S. and Maxwell, J. T. (2005). On Some Pitfalls in Automatic Evaluation and Significance Testing for MT. *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, 57–64. https://www.aclweb.org/anthology/W05-0908

Rogers, R. K. and Grant, J. (1997). Content Analysis of Information Cited in Reports of Sell-side Financial Analysts. *Journal of Financial Statement Analysis*, *3*, 17–31.

Ruder, S. (2016). An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747*.

Rumelhart, D. E., Hinton, G. E. and Williams, R. J. (1986). Learning representations by back-propagating errors. *nature*, *323*(6088), 533–536.

Salton, G. and Lesk, M. E. (1968). Computer evaluation of indexing and text processing. *Journal of the ACM (JACM)*, *15*(1), 8–36.

Salton, G. and McGill, M. J. (1971). The SMART Retrieval System - Experiments in Automatic Document Retrieval. *Prentice Hall Inc., Englewood Cliffs, NJ*.

Salton, G. and McGill, M. J. (1983). *Introduction to Modern Information Retrieval*. McGrawHill.

Sanderson, G. and Pullen, J. (2021). *The ESG Data Challenge*. Retrieved September 25, 2021, from https://www.3blue1brown.com/lessons/backpropagation-calculus

SASB. (2016). *Business and Financial Disclosure Required by Regulatin S-K - The SEC's Concept Release and Its Implication*. Retrieved January 8, 2021, from https://www.sasb.org/wp-content/uploads/2016/09/Reg-SK-Comment-Bulletin-091416.pdf

SASB. (2017). SASB Conceptual Framework, 1–75. https://www.sasb.org/standard-setting-process/conceptual-framework/.

SASB. (2020). *IIRC and SASB announce intent to merge in major step towards simplifying the corporate reporting system*. https://www.sasb.org/wp-content/uploads/2020/11/IIRC-SASB-Press-Release-Web-Final.pdf

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

Schroeder, N. and Gibson, C. (1990). Readability of Management's Discussion and Analysis. *Accounting Horizons*, *4*(4), 78–87.

SEC. (1987). Securities Act Release No. 33-6711. April 24.

SEC. (1989). Securities Act Release No. 33-6835. May 18 [https://www.sec.gov/rules/interp/33-6835.htm].

SEC. (1999). *17 cfr part 211, release no. sab 99. august 12*. https://www.sec.gov/interps/account/sab99.htm

SEC. (2003). Commission Guidance Regarding Management's Discussion and Analysis of Financial Condition and Results of Operations.

SEC. (2010). *Commission Guidance Regarding Disclosure Related to Climate Change*. Retrieved September 20, 2021, from https://www.govinfo.gov/content/pkg/FR-2010-02-08/pdf/2010-2602.pdf

SEC. (2011). *How to read a 10-k*.

SEC. (2016). *Concept Release: Businessand Financial Disclosure Required by Regulation S-K*. Retrieved January 14, 2021, from https://www.sec.gov/rules/concept/2016/33-10064.pdf

SEC. (2018). Securities act release no. 33-10514. june 18. inline xbrl filing of tagged data [https://www.sec.gov/rules/final/2018/33-10514.pdf].

SEC. (2019). Modernization of Regulation S-K Items 101, 103, and 105 [SEC Release Nos. 33-10668, 34-86614. Proposed Rule].

SEC. (2020). Securities Act Release No. 33-10890. February 10 [https://www.sec.gov/rules/final/2020/33-10890.pdf].

SEC. (2021). *Sec announces enforcement task force focused on climate and esg issues*. Retrieved September 10, 2021, from https://www.sec.gov/news/press-release/2021-42

Semenova, N. and Hassel, L. G. (2008). Financial outcomes of environmental risk and opportunity for us companies. *Sustainable development*, *16*(3), 195–212.

Shrikumar, A., Greenside, P. and Kundaje, A. (2017). Learning important features through propagating activation differences. In D. Precup and Y. W. Teh (Eds.), *Proceedings of the 34th international conference on machine learning* (pp. 3145–3153). PMLR. http://proceedings.mlr.press/v70/shrikumar17a.html

Silva, W., Fernandes, K., Cardoso, M. J. and Cardoso, J. S. (2018). Towards Complementary Explanations Using Deep Neural Networks. In D. Stoyanov, Z. Taylor, S. M. Kia, I. Oguz, M. Reyes, A. Martel, L. Maier-Hein, A. F. Marquand, E. Duchesnay, T. Löfstedt, B. Landman, M. J. Cardoso, C. A. Silva, S. Pereira and R. Meier (Eds.), *Understanding and interpreting machine learning in medical image computing applications* (pp. 133–140). Springer International Publishing.

Smeuninx, N., De Clerck, B. and Aerts, W. (2020). Measuring the readability of sustainability reports: A corpus-based analysis through standard formulae and NLP. *International Journal of Business Communication*, *57*(1), 52–85.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. and Salakhutdinov, R. (2014). Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *The journal of machine learning research*, *15*(1), 1929–1958.

Stolowy, H. and Paugam, L. (2018). The expansion of non-financial reporting: An exploratory study. *Accounting and Business Research*, *48*(5), 525–548.

Sutton, R. S. and Barto, A. G. (1998). *Introduction to Reinforcement Learning* (Vol. 135). MIT press Cambridge.

Tai, Y.-J. and Kao, H.-Y. (2013). Automatic domain-specific sentiment lexicon generation with label propagation. *Proceedings of International Conference on Information Integration and Web-based Applications & Services*, 53–62.

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

TCFD. (2019). *Task Force on Climate-related Financial Disclosures: 2019 Status Report*. Retrieved January 16, 2021, from https://www.fsb-tcfd.org/wp-content/uploads/2019/06/2019-TCFD-Status-Report-FINAL-053119.pdf

TCFD. (2020). *Guidance on risk management integration and disclosure*. Retrieved December 18, 2020, from https://assets.bbhub.io/company/sites/60/2020/09/2020-TCFD_Guidance-Risk-Management-Integration-and-Disclosure.pdf

The Economist. (2020). In the soup - Measuring Sustainability [The Economist,(2020, October 3), Vol. 437, No. 9214 https://www.economist.com/business/2020/10/03/the-proliferation-of-sustainability-accounting-standards-comes-with-costs].

Tzoukermann, E., Klavans, J. L. and Strzalkowski, T. (2005). Information Retrieval. In R. Mitkov (Ed.), *The oxford handbook of computational linguistics* (pp. 530–544). Oxford University Press.

van der Linden, I., Haned, H. and Kanoulas, E. (2019). Global Aggregations of Local Explanations for Black Box Models [https://arxiv.org/abs/1907.03039].

van Rijsbergen, C. J. (1975). *Information retrieval*. Butterworths.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. and Polosukhin, I. (2017). Attention is all you need. *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 6000–6010.

Verbeeten, F. H., Gamerschlag, R. and Möller, K. (2016). Are CSR disclosures relevant for investors? Empirical evidence from Germany. *Management Decision*.

Verrecchia, R. E. (1983). Discretionary disclosure. *Journal of Accounting and Economics*, *5*, 179–194. https://doi.org/https://doi.org/10.1016/0165-4101(83)90011-3

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

Warner, J. B. and Kothari, S. P. (2006). Econometrics of event studies. In E. Eckbo (Ed.), *Handbook of corporate finance: Empirical corporate finance* (pp. 3–36). North-Holland, Elsevier.

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

WBCSD. (2018). *Enhancing the Credibility of Non-financial Information: The Investor Perspective.* Retrieved January 9, 2021, from https://docs.wbcsd.org/2018/10/WBCSD_Enhancing_Credibility_Report.pdf

Windolph, S. E. (2011). Assessing corporate sustainability through ratings: Challenges and their causes. *Journal of Environmental sustainability*, *1*(1), 5.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M. and Brew, J. (2019). Huggingface's transformers: State-of-the-art natural language processing [https://arxiv.org/abs/1910.03771].

World Economic Forum. (2020). Measuring stakeholder capitalism: Towards common metrics and consistent reporting of sustainable value creation [(accessed on 2020-12-18)].

Xing, F., Malandri, L., Zhang, Y. and Cambria, E. (2020). Financial Sentiment Analysis: An Investigation into Common Mistakes and Silver Bullets. *Proceedings of the 28th International Conference on Computational Linguistics*, 978–987. https://www.aclweb.org/anthology/2020.coling-main.85

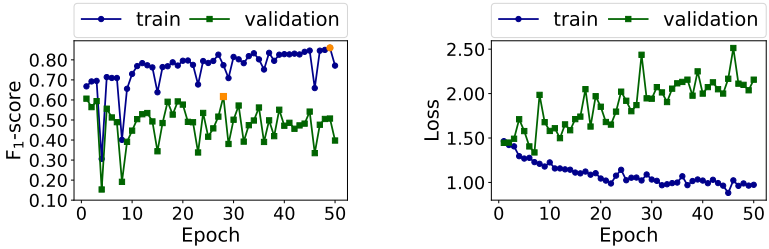
Xing, F. Z., Cambria, E. and Welsch, R. E. (2018). Natural Language based Financial Forecasting: A Survey. *Artificial Intelligence Review*, *50*(1), 49–73.

Yang, Z., Yang, D., Dyer, C., He, X., Smola, A. and Hovy, E. (2016). Hierarchical attention networks for document classification. *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, 1480–1489.

Zhang, Y. and Wallace, B. C. (2017). A Sensitivity Analysis of (and Practitioners' Guide to) Convolutional Neural Networks for Sentence Classification. *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 253–263.

A. Appendix

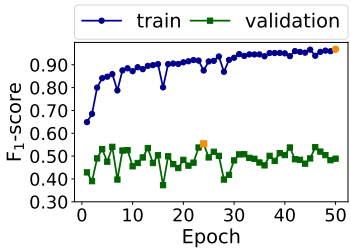
A.1. F-score and Loss Curve Plots



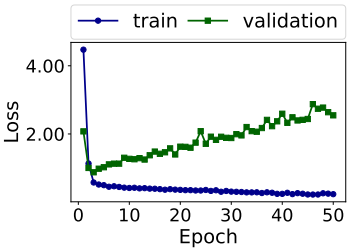
(a) F_1-score, **BOW** architecture *direct*-classifier – *short-term* model (2a).

(b) Loss, **BOW** architecture *direct*-classifier – *short-term* model (2a).

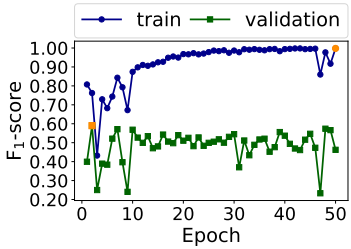
Fig. A.1.: F_1-score and loss curve for the *direct*-classifiers – *short-term* model (2a).



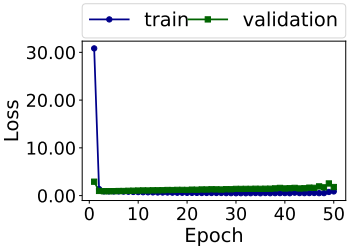
(c) F_1-score, **TF-IDF** architecture *direct*-classifier – *short-term* model (2a).



(d) Loss, **TF-IDF** architecture *direct*-classifier – *short-term* model (2a).

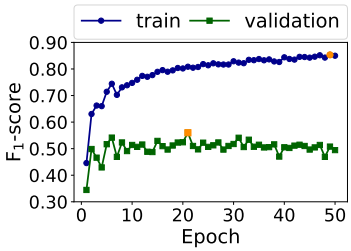


(e) F_1-score, **GloVe** architecture *direct*-classifier – *short-term* model (2a).

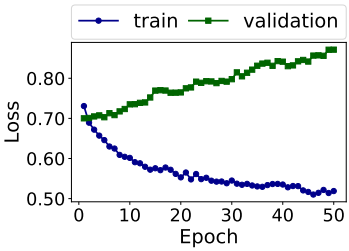


(f) Loss, **GloVe** architecture *direct*-classifier – *short-term* model (2a).

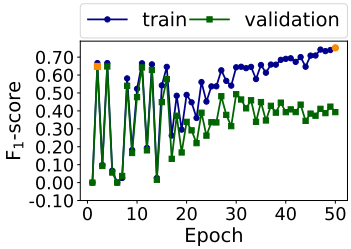
Fig. A.1.: F_1-score and loss curve for the *direct*-classifiers – *short-term* model (2a).



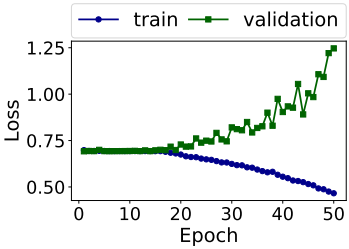
(g) F_1-score, **CNN** architecture *direct*-classifier – *short-term* model (2a).



(h) Loss, **CNN** architecture *direct*-classifier – *short-term* model (2a).

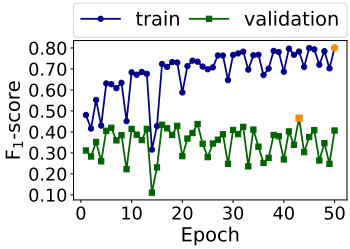


(i) F_1-score, **BERT** architecture *direct*-classifier – *short-term* model (2a).

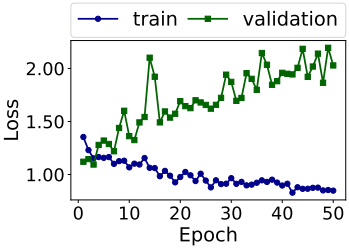


(j) Loss, **BERT** architecture *direct*-classifier – *short-term* model (2a).

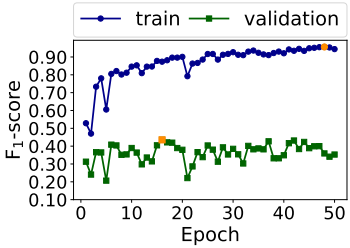
Fig. A.1.: F_1-score and loss curve for the *direct*-classifiers – *short-term* model (2a).



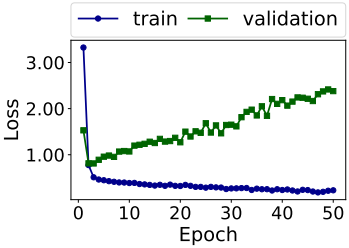
(a) F_1-score, **BOW** architecture *env.*-classifier – *short-term* model (2a) and (2b).



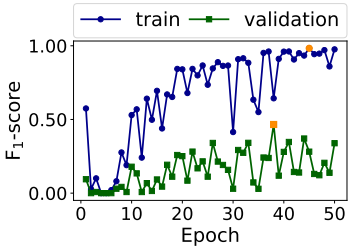
(b) Loss, **BOW** architecture *env.*-classifier – *short-term* model (2a) and (2b).



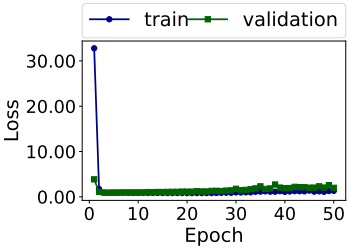
(c) F_1-score, **TF-IDF** architecture *env.*-classifier – *short-term* model (2a) and (2b).



(d) Loss, **TF-IDF** architecture *env.*-classifier – *short-term* model (2a) and (2b).

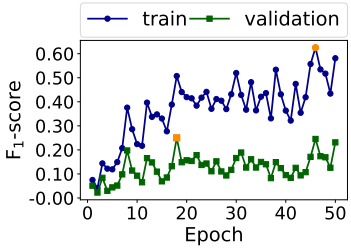


(e) F_1-score, **GloVe** architecture *env.*-classifier – *short-term* model (2a) and (2b).

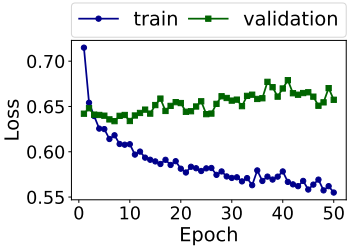


(f) Loss, **GloVe** architecture *env.*-classifier – *short-term* model (2a) and (2b).

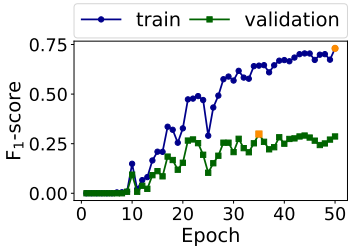
Fig. A.2.: F_1-score and loss curve for the *env.*-classifiers – *short-term* model (2a) and (2b).



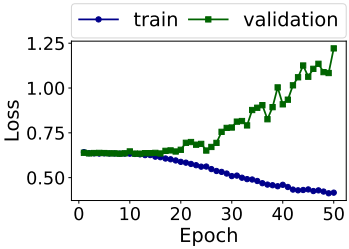
(g) F_1-score, **CNN** architecture *env.*-classifier – *short-term* model (2a) and (2b).



(h) Loss, **CNN** architecture *env.*-classifier – *short-term* model (2a) and (2b).

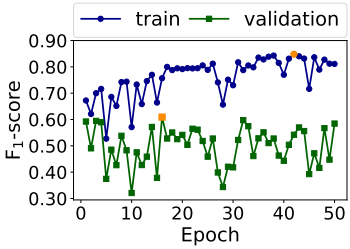


(i) F_1-score, **BERT** architecture *env.*-classifier – *short-term* model (2a) and (2b).

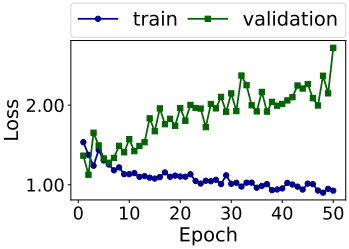


(j) Loss, **BERT** architecture *env.*-classifier – *short-term* model (2a) and (2b).

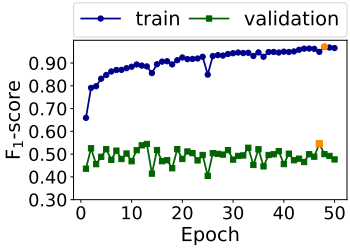
Fig. A.2.: F_1-score and loss curve for the *env.*-classifiers – *short-term* model (2a) and (2b).



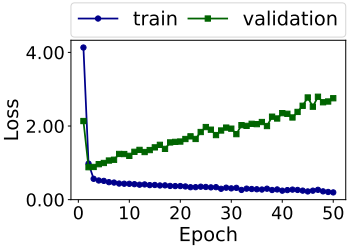
(a) F_1-score, **BOW** architecture *combined*-classifier – *short-term* model (2a).



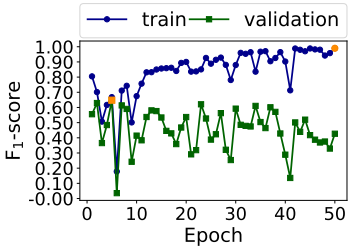
(b) Loss, **BOW** architecture *combined*-classifier – *short-term* model (2a).



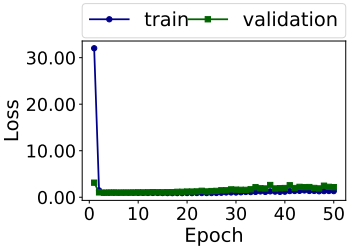
(c) F_1-score, **TF-IDF** architecture *combined*-classifier – *short-term* model (2a).



(d) Loss, **TF-IDF** architecture *combined*-classifier – *short-term* model (2a).

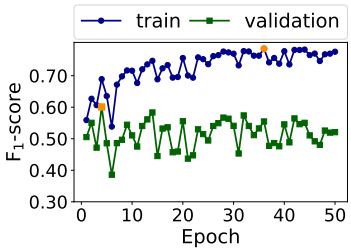


(e) F_1-score, **GloVe** architecture *combined*-classifier – *short-term* model (2a).

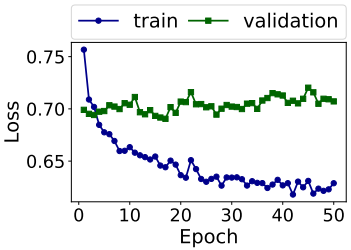


(f) Loss, **GloVe** architecture *combined*-classifier – *short-term* model (2a).

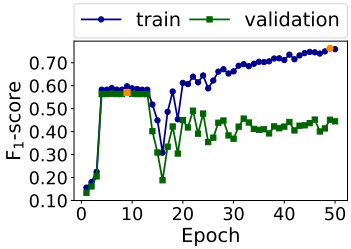
Fig. A.3.: F_1-score and loss curve for the *combined*-classifiers – *short-term* model (2a).



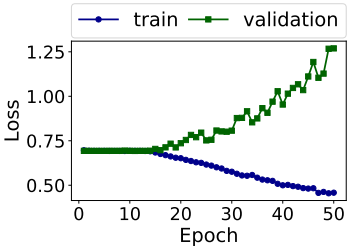
(g) F_1-score, **CNN** architecture *combined*-classifier – *short-term* model (2a).



(h) Loss, **CNN** architecture *combined*-classifier – *short-term* model (2a).

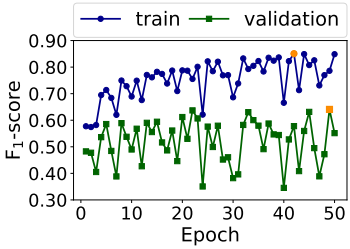


(i) F_1-score, **BERT** architecture *combined*-classifier – *short-term* model (2a).

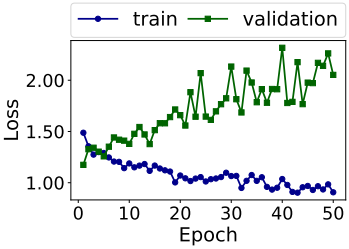


(j) Loss, **BERT** architecture *combined*-classifier – *short-term* model (2a).

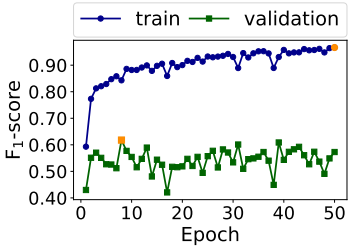
Fig. A.3.: F_1-score and loss curve for the *combined*-classifiers – *short-term* model (2a).



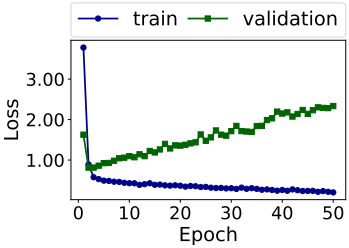
(a) F_1-score, **BOW** architecture *direct*-classifier – *short-term* model (2b).



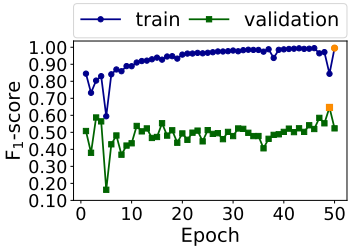
(b) Loss, **BOW** architecture *direct*-classifier – *short-term* model (2b).



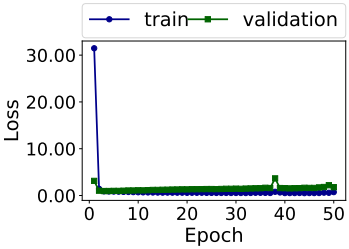
(c) F_1-score, **TF-IDF** architecture *direct*-classifier – *short-term* model (2b).



(d) Loss, **TF-IDF** architecture *direct*-classifier – *short-term* model (2b).

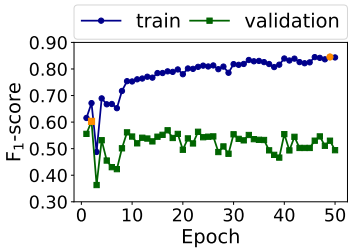


(e) F_1-score, **GloVe** architecture *direct*-classifier – *short-term* model (2b).

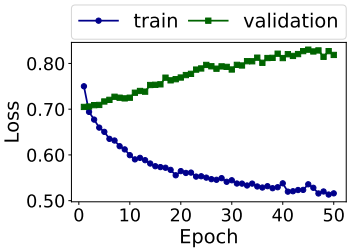


(f) Loss, **GloVe** architecture *direct*-classifier – *short-term* model (2b).

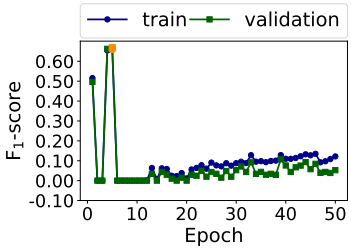
Fig. A.4.: F_1-score and loss curve for the *direct*-classifiers – *short-term* model (2b).



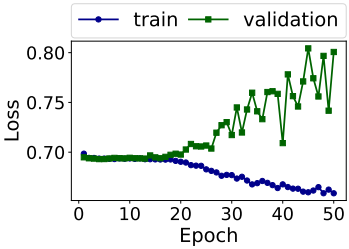
(g) F_1-score, **CNN** architecture *direct*-classifier – *short-term* model (2b).



(h) Loss, **CNN** architecture *direct*-classifier – *short-term* model (2b).

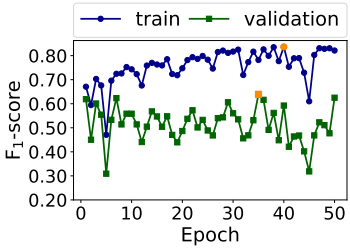


(i) F_1-score, **BERT** architecture *direct*-classifier – *short-term* model (2b).

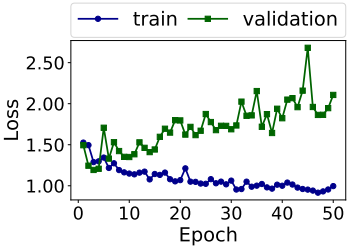


(j) Loss, **BERT** architecture *direct*-classifier – *short-term* model (2b).

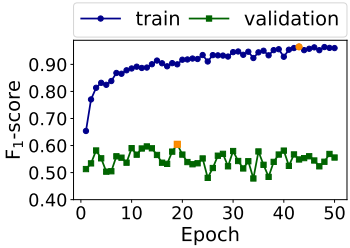
Fig. A.4.: F_1-score and loss curve for the *direct*-classifiers – *short-term* model (2b).



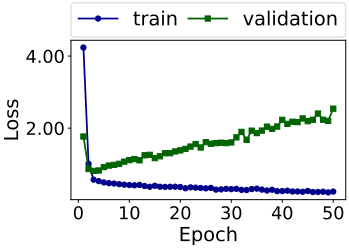
(a) F_1-score, **BOW** architecture *combined*-classifier – *short-term* model (2b).



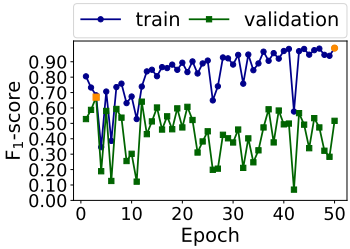
(b) Loss, **BOW** architecture *combined*-classifier – *short-term* model (2b).



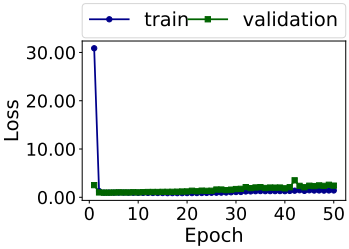
(c) F_1-score, **TF-IDF** architecture *combined*-classifier – *short-term* model (2b).



(d) Loss, **TF-IDF** architecture *combined*-classifier – *short-term* model (2b).

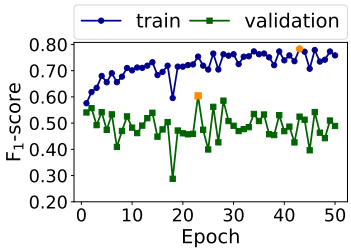


(e) F_1-score, **GloVe** architecture *combined*-classifier – *short-term* model (2b).

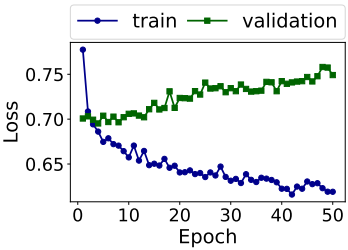


(f) Loss, **GloVe** architecture *combined*-classifier – *short-term* model (2b).

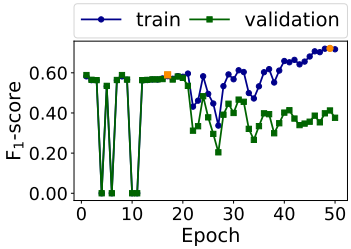
Fig. A.5.: F_1-score and loss curve for the *combined*-classifiers – *short-term* model (2b).



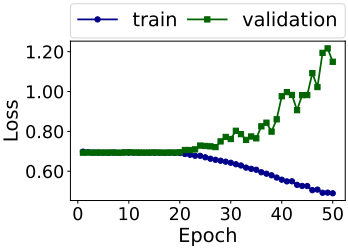
(g) F_1-score, **CNN** architecture *combined*-classifier – *short-term* model (2b).



(h) Loss, **CNN** architecture *combined*-classifier – *short-term* model (2b).



(i) F_1-score, **BERT** architecture *combined*-classifier – *short-term* model (2b).



(j) Loss, **BERT** architecture *combined*-classifier – *short-term* model (2b).

Fig. A.5.: F_1-score and loss curve for the *combined*-classifiers – *short-term* model (2b).

Declaration

Eidesstattliche Erklärung

Ich erkläre mit meiner Unterschrift, dass ich diese Arbeit selbstständig verfasst habe und keine anderen als die angegebenen Quellen benutzt habe. Alle Stellen dieser Arbeit, die dem Wortlaut, dem Sinn oder der Argumentation nach anderen Werken entnommen sind, habe ich unter Angabe der Quellen vollständig kenntlich gemacht.

Stuttgart, den 16. Juni 2022 Felix Arnsdorf .....

Declaration of Authorship

I declare that I completed this thesis on my own and that passages and ideas which have been directly or indirectly taken from other sources have been noted as such. Neither this nor a similar work has been presented to an examination committee.

Stuttgart, 16th June 2022 Felix Arnsdorf .....