Recent Advancements in Sentiment Analysis in Finance

Constanze Bayer

Volkswirtschaftslehre (PO 2010)

A thesis submitted for the degree of
Bachelor of Science

First examiner (Erstprüfer)
Prof. Dr. Stefan Lessmann

https://doi.org/10.18452/20866

Wirtschaftswissenschaftliche Fakultät
Humboldt-Universität zu Berlin

July 28, 2019

Abstract

Sentiment analysis as deriving opinions or emotional content from written documents is an important tool in finance. It is used for prediction of stock markets, prediction of financial risk or distress, and detection of abnormal investor behavior by analyzing companies' reports, social media posts or news articles. To conduct the analysis, either word lists or machine learning algorithms are used. Taking context of sentence structures into account improves both research methods. This thesis applies these findings by analyzing German Corporate Social Responsibility reports and parts of annual reports to investigate an association with environmental, social and governance rating scores. The aim is to assess sentiment as a risk indicator for these issues. The results indicate that proportional sentiment scores proxy as metric for reporting quality rather than risk.

Declaration of Authorship

I hereby declare that the work submitted has not yet been submitted for any other examination. It was written independently. All sources, including Internet sources, which were reproduced unchanged or modified, in particular for texts, graphics, tables and pictures, were marked as such. I am aware that in the event of violations of these principles, proceedings will be instituted for attempted deception or deception.

Selbstständigkeitserklärung

Hiermit erkläre ich, dass die vorgelegte Arbeit noch für keine andere Prüfung eingereicht wurde. Sie wurde selbstständig verfasst. Sämtliche Quellen, einschließlich Internetquellen, die unverändert oder abgewandelt wiedergegeben wurden, insbesondere für Texte, Grafiken, Tabellen und Bilder, wurden als solche kenntlich gemacht wurden. Mir ist bekannt, dass bei Verstößen gegen diese Grundsätze ein Verfahren wegen Täuschungsversuch bzw. Täuschung eingeleitet wird.

Constanze Bayer
July 28, 2019

Contents

1 Introduction **3**

2 Literature review **4**
2.1 The field until 2014 . . . 4
2.2 Sources . . . 5
2.3 Sentiment Analysis . . . 6
2.3.1 Lexicon based Approaches . . . 7
2.3.2 Machine Learning Approaches . . . 11
2.4 Goals . . . 15
2.5 Missing and underrepresented research streams . . . 18

3 Empirical **19**
3.1 Data . . . 20
3.2 Methodology . . . 24

4 Results **27**
4.1 Sentiment analysis . . . 27
4.1.1 Complete sample . . . 27
4.1.2 Annual reports . . . 30
4.1.3 Sustainability reports . . . 34
4.1.4 CEO letters . . . 37
4.2 Correlation to ESG data . . . 39
4.3 Discussion . . . 44

5 Conclusions **47**

Appendices **54**

1 Introduction

The growing body of investment related texts makes it hard for investors to keep up with the information in order to make decisions. The content and tone of official disclosures have to be considered alongside news articles and social media posts regarding companies and their business situation. Automated sentiment analysis can be one tool to decrease this complexity by generating insights on the tone or the emotional content of a given text without having to read it in detail. The method makes it possible to automatically distinguish good news from bad news and place investment decisions accordingly. Big amounts of unstructured data can be processed like that.

Sentiment analysis tries to quantify opinions and attitudes expressed in texts. In the field of finance, this sentiment expressed in texts has become an important proxy for investor sentiment. In behavioral finance approaches as introduced by Shleifer (2000) [1] this yields insights in the irrationality of investors. In Efficient Market Hypothesis context as building on the work of Fama (1970) [2], sentiment of texts expresses the current state of information that fundamentally drives stock prices. In both contexts, the automated analysis of any given text sort is useful.

This thesis sets out to contribute to the field in the following: The first part summarizes major research streams in the field since approximately 2014. It goes beyond what is necessarily needed to motivate the empirical second part but aims to give the reader a comprehensive review of what has been done in finance sentiment analysis. With the amount of textual data, the application of this methodology has increased in the last years. As noted in earlier research by Kearney and Liu (2014) [3], two approaches still dominate the field: Either the analysis is lexicon-based and uses a given word list to derive the sentiment or a machine learning approach is used. Recent research on the methodology improved both approaches by including semantic information in the analysis either by handling of negations and intensifications (lexicon-based) or fine-grained feature engineering (machine learning). New applications have been developed in the finance-related sentiment research: Starting from stock market prediction the analysis is extended to risk prediction and abnormal behavior in financial markets. Major sources are social media posts as well as company disclosures of different kind. Altogether, sentiment analysis among others has evolved to be an important tool to extract risk and risk perception from texts similar to risk quantification from numerical data.

This broad literature review is used in the second empirical part of this thesis. Using sentiment analysis as a risk indicator, I focus on risks related to environmental, social and governance issues. These have become increasingly important in the last years for companies as they are more in the focus on credit rating agencies[1]. Therefore, Corporate Social Responsibility (CSR) reports[2] and sections from annual reports will be analyzed as well as letters of the Chief Executive Officer (CEO). The goal is to analyze the suitability of this text type to indicate risk related to environmental, social and governance (ESG) issues. Sentiment analysis is used to extract attitudes towards CSR and ESG related topics: Are they perceived as risks or opportunities? The major contribution is to tap a rarely used source of information: Corporate Social Responsibility reports issued by German mid cap companies. For this, a lexicon-based approach is used to perform the sentiment analysis. Results are mixed: Association with certain subcategories is strong but the relation to environmental, social and governance (ESG) risk scores is rather weak. It seems that proportional sentiment scores are able to indicate reporting quality rather than risk.

The remainder of this paper is organized as follows: Section 2 summarizes major findings in the field of finance-related sentiment analysis. Section 3 describes the two experiments conducted in terms of data retrieval and methodology whereas section 4 describes the derived results. The last section 5 concludes.

2 Literature review

2.1 The field until 2014

Sentiment Analysis has been a field of increasing research in the past years. According to Mäntyläa, Graziotin and Kuutila (2018) [4], the sheer amount of articles on the topic has been exploding since 2004 as public opinion in texts of various lengths were readily available with the emergence of the so-called Web 2.0. In their review, they find finance to be one the major fields of application of sentiment analysis. A more qualitative review of this specific field of sentiment analysis in finance up until 2014 is given by Kearney and Liu (2014) [3]. Just like Mäntyläa et al. (2018), they discover the rise of social media texts a source for sentiment analysis.

I will review developments in finance-related sentiment analysis since approximately 2014 as this is the point in time of the most comprehensive reviews in this field. Non-finance applied general evolvements in the field of sentiment analysis are covered in the last part of this literature review and include only major developments that could be useful in the field of finance.

Following the framework of Kearney and Liu (2014), research is organized as the process of a sentiment analysis: starting with the text sources, being processed and analyzed and then be tested and evaluated to achieve a certain goal. This process as well as important contributions can be seen

[1] `https://www.ft.com/content/c1f29e0c-6012-3ac5-9a05-13444b89c5ec`

[2] In the following text, CSR report and Sustainability report is used interchangeable.

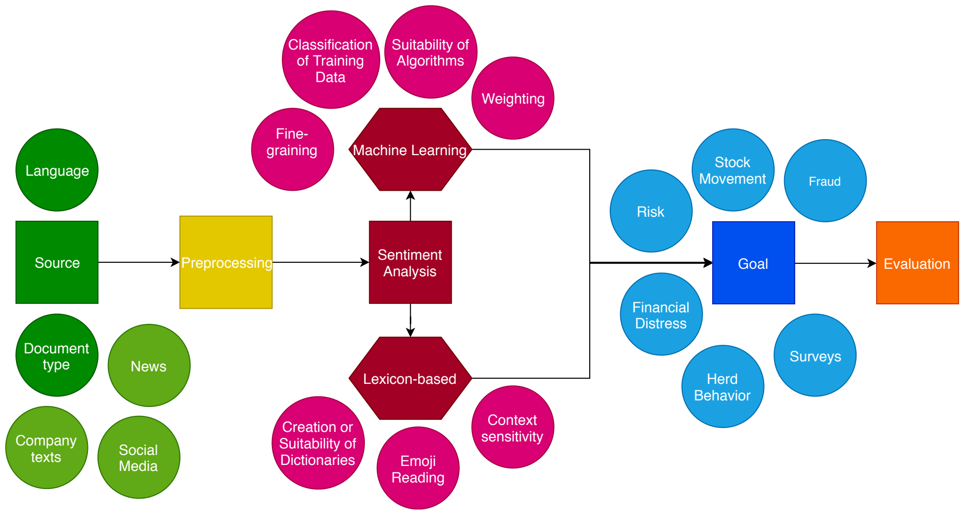


Figure 1: Areas of research

in Figure 1. Focus in research is on the general approach of how to run the analysis, either by a lexicon-based approach or a machine learning one. Furthermore, the goals of the sentiment analysis have been broadened. Also new sources in terms of documents and languages are being developed. Lesser research is done on the preprocessing and the evaluation of the sentiment scores. Research on these steps in the conduction of a sentiment analysis will therefore not be elaborated in detail.

2.2 Sources

There are two main sources used very often for sentiment analysis in finance: One are texts published by companies, such as reports, adhoc announcements, press releases and financial disclosures. They give insights into a company's situation beyond financial statements. Among these disclosures are 10-K-forms ([5], [6], [7], [8], [9], [10]), earnings conference calls ([11]), annual reports ([12], [13]) and press releases ([14], [15]).

Hummel, Mittelbach-Hörmanseder, Cho and Matten (2017) [16] are the first to look beyond the financial disclosures published by companies. They study the styles and subjects of voluntary Corporate Social Responsibility publications and find differences concerning subjects and tone of the disclosures. Liberal market economies like the US seem to offer more explicit and more positive disclosures on CSR topics than Coordinated Market Economies like the United Kingdom.

Another main source for sentiment analysis are microblogs and social media. They are used as indicators for investor sentiment or public opinion that was established by behavioral finance research. Popular are short messages directly related to stocks ([17], [18], [19], [20], [21], [22], [23]). This is advantageous as the analyzed text will most likely have a finance focus. On the other hand, messages with finance focus on a general message board like Twitter are under research. Here, a selection has to be made. Eliaçik and Erdoğan (2018) [24] formed a social community interested in finance in Turkish whose tweets were then analyzed. Users who are included use more than 50% Turkish words, have mutual relationships with important users and are member of the network for more than four weeks. Seed users were two business media channels. Whether or not analyzed tweets were on finance topics or not was not further examined.

Besides these two main approaches, another important research stream focuses on news articles ([25], [26], [27], [28], [29]). The relationship between press releases and stock movements is of special interest as these information are publicly available. Day and Lee (2016) [30] compare four different news providers and find that those more business focused are more useful to build a trading strategy on.

Less used than in the early years of sentiment analysis ([4]), are online product reviews. Teng, Vo and Zhang (2016) [31] use them for comparison to other datasets, Zhou, Xia and Zhang (2016) [32] to extract answers to questionnaires and Aganthangelou, Katakis, Kokkoras and Ntonas (2014) [33] to create a domain-specific lexicon.

Sources as structured by the origin of the material are mostly focused on US-American companies. Creamer, Sakamoto and Nickerson (2016) [34] move away from this US-American bias and build their research on a European news provider to predict the movement of a European stock index, the Stoxx50.

Besides the input data that are used also languages used for sentiment analysis are evolving. Al-Kabi, Gigieh, Alsmadi and Wahsheh (2014) [35] extend the research to the Arabic language with its own challenges: difference between colloquial and Standard Arabic is quite fundamental and various dialects increase the number of different spellings. The authors develop a tool and several lexicons to use for sentiment analysis. Aganthangelou et al. (2014) [33] develop a similar tool to deal with other languages than English. In their work they provide it for Greek. Besides these efforts, the main focus of research is still on US-American English and Chinese text sources from social media or pure business related company disclosures. Less common languages and documents remain to further research as in this thesis.

2.3 Sentiment Analysis

To conduct a sentiment analysis, two main approaches are still prevalent: One is using a pre-defined word list to calculate the polarity of a text, either negative or positive, that is the sentiment[3] of a sentence or document. The

[3]In the further text I will refer to sentiment as the opinion or polarity expressed in a text.

more negative or positive words from this list occur in the given text, the more extreme is the polarity of a statement. This is called the 'lexicon-based'-approach. It is often used in 'Bag of Words'-manner (BOW). This means, the structure of a text is ignored. All words are parsed into a word-frequency-matrix and their occurrence is treated as independent of each other. Sentiment scores are added if a word out of this bag is found on a predefined list. One of the most common collection of word lists used is the one provided by Loughran and McDonald (2011) [36], referred to as LM. It is adjusted to the domain of finance which the authors proved to be necessary. Alongside negative and positive lists, they also created lists with words that express uncertainty[4]. This is especially useful to not only derive positive or negative attitudes but also degrees to which texts express feelings like uncertainty.

Different from this lexicon-based approach is the use of machine learning algorithms. In order to train these, a pre-classified dataset is needed as well as a test dataset to evaluate the performance of the predictor (for metric values) or classifier (overall sentiment polarity to negative or positive). Common quality metrics for classification tasks like accuracy, precision, recall and F1 are used to evaluate the performance.

2.3.1 Lexicon based Approaches

Lexicon-based approaches are easy to conduct as the only resource needed is a predefined list of sentiment words. But the approach has several shortcomings that are discussed in the literature. Exemplarily, Meyer, Bikdash and Dai (2017) list the "lack of domain specificity, the independence assumption [for the occurence of every single word without context], the laborious nature of building a lexicon, the absence of context and their non-robust nature due to missing words"[27, p. 3].

Consequently, research with regard to lexicon-based sentiment analysis is aiming at these drawbacks. Main research areas are the generation of domain-specific lexicons and the inclusion of context in terms of sentence structure, negation and intensification by moving away from a Bag-of-Words-approach that treats all words as if they were independent from each other. An overview of important evolvements in this field are given in Table 1. Included are only articles that are related to finance.

[4]Also word lists that are used in context of litigation risk, business possibilities and necessities are created but are less used in the research.

Cite	Source	Goal	Measurement	Testing	Relevant findings
[5]	Financial disclosures	Sentiment classification	% negative and % positive	Correlation	Commonly used 'Diction' not useful
[6]	Financial disclosures	Financial risk as stock volatility	Frequency based metrics	Ranking and Regression	Strong correlation between soft information and risk
[7]	Financial disclosures	Financial risk as stock volatility	Frequency based metrics	SVM, SVR and Ranking	Risk can be measures from textual analysis
[9]	10-K files (financial reports)	Detect aggressive tax planning (avoiding)	Proportion of negative words	Regression	Replace quantitative with qualitative measure of constraint, distressed firms use aggressive tax planning strategies
[10]	10-K files (financial reports)	Measure financial constraint	Proportion of constraining words	Correlation, Regression	With and without control variables, metric predicts liquidity events, flipping points in financial history detectable
[11]	Earnings Conference Calls	Measuring tone dispersion	Number of sentiment words	Multivariate Regression	Different strategies associated with dispersion
[12]	Banks' annual reports	Detect financial distress	% negative words Document level	Regression	Relationship exists, forward-looking and less prone to window-dressing
[13]	CEO-letters, outlook sections in annual reports	Financial distress for banks	Weighted score for document level	Correlation, Regression	Macro level useful for prediction, for individual banks not reliable: tool for supervisory
[14]	Adhoc announcements	Sentiment classification	Net Optimism	Correlation	Reinforcement outperform in learning to detect automatic negation scope
[15]∗	Adhoc announcements	Stock prices	Sentence: weighted score, Document: classification to neg/ pos	Mean return	Rhetoric Structure Theory to decompose meaning, weighted as in RST tree, Random forest improve prediction
[16]	Voluntary CSR disclosures	Differences in style and topics	% positive words	Regression	Differences in regard to Economy type (LME vs. CME) and CSR policies

[18]	Chinese stock forums	Detect herd behavior	Sentiment index between -1 and 1	Support Vector Regression	sentiment good predictor for stock price, spurious and true herding detectable
[22]	StockTwits	Lexicon for stock market sentiment	Pre-defined and self-developed scores	Comparison via correlation	Negation not necessarily shift of meaning
[25]	Finance related news	Financial risk as stock volatility	Net Sentiment	SVM, ANN	Useful in times of high volume of news
[26]	News articles	Adjust given lexicon to domain	Net frequency	Support Vector Classifier	Relationship to stock movement not strong
[28]	Finance articles	Increase domain accuracy	Manual inspection	Comparison to general lexicon	Efficient algorithm to detect candidates, very small dataset
[29]	Bitcoin news media	Predict Bitcoin price	(P-N)/(P+N) Document level	Trading strategy	Prediction possible, does not outperform other measures
[31]∗	Tweets, movie reviews and mixed	Sentiment classification	Sum of weighted scores plus bias	Comparison to traditional approaches	Context-sensitivity by Recurrent Neural Networks increase accuracy
[33]	Greek product reviews	Domain specific lexicon	Net sentiment	Correct sentiment classification	Small seed sufficient for lexicon-creation
[35]	Arabic social media reviews	Sentiment classification	Net-Sentiment	Comparison with test data	Develop lexicons for Arabic language
[37]	Financial reports	Comparison of sentiment metrics	Several	Association with market reaction after earnings announcement	Proportional metrics competitive to more complex including algorithms, important: domain-specific lexicon
[38]	Financial reports (Management Discussion & Analysis section)	Detect fraudulent behavior in reports written by group	Proportional word counts	Correlation and regression	Word choices transmits to a certain degree (depending on word list), Innocents can transmit fraudulent language without realizing it

Table 1: Lexicon-based studies, ∗Hybrid model

The first research path to improve existing and generate new lexicons is followed by Loughran and McDonald (2015) [5]. They extend their earlier research by showing that another wildly used word list 'Diction' is not suitable when working with texts from a financial domain. To extract more fine-grained sentiment from texts, Bodnaruk, Loughran and McDonald (2015) [10] create a new word list additional to exsting ones that contains words that signal financial constraint. Oliveira, Cortez and Areal (2016) [22] on the other hand suggest a new method to generate domain-specific lexicons from a seed word list. One technique to do this is Semantic Orientation Pointwise Mutual Information (SOPMI). SOPMI is based on the co-occurence of given sentiment words and other words that then are tagged as candidates for new sentiment words. Aganthangelou et al. (2014) [33] develop their own sentiment lexicon to derive sentiment of product reviews in a similar manner from a given seed word list. Moreno-Ortiz and Fernandez-Cruz (2015) [28] try to derive domain-specific lexicons that can be used when needed, just as a plug-in to the analysis. As a caveat of their work, it is used on a very small dataset of only two news articles. The importance of domain-specific lexicons is emphasized by research done Henry and Leone (2016) [37]. They improve their earlier research (2009, [39][5]) by comparing several metrics for sentiment classification. They find that proportion based metrics can compete to more complex metrics and machine learning algorithms in their suitability for finance - under the condition that a domain-specific lexicon is used. More general dictionaries in contrast can not.

Purpose-specific word lists are also in focus in the research provided by Moore, Rayson and Young (2016) [26]. They use a machine learning approach to adjust the lexicon in use. In the training phase of their model, the authors adjust the given word list to the most frequently used words in positive, negative and neutral sentiment in press releases of three companies. Support Vector Classifiers are then used to evaluate the performance of these shortened word lists. As reference classification, they use the movement of stocks of the particular companies which is an inversion of the assumption that stock movements and release sentiments correlate.

Teng et al. (2016) [31] present on the other hand a new approach to take into account more context of sentiment word by using Recurrent Neural Networks[6]. Sentiment scores derived from a given word list are adjusted by weights that are context-sensitive. The overall sentence level sentiment score is a sum of these weighted word sentiment scores plus a bias term. This leads to satisfying results in classification. Märkle-Huß, Feuerriegel and Prendinger (2017) [15] take a similar approach but decompose the sentence

[5]Not included in Table 1 as it is out of scope of this thesis.

[6]Other than neural networks with a feed-forward mechanism, recurrent neural networks do have some kind of memory as they use cycles to learn out of the current instance. This and the general concept of neural networks is explained in Witten et al. (2011) [40, p. 235ff.].

according to Rhetoric Structure Theory[7]. This gives a tree structure for the sentence that explains the main subject and the depth of the elaboration. Sentiment words are weighted according to their position in that tree, that is according to importance to the overall subject of a sentence. In a second experiment, performance is further improved by applying a machine learning approach (Random Forests) with these information given as features.

Another way of incorporating more context in the analysis is the handling of negations. Nopp and Hanbury (2015) [13] apply a very simple approach by shifting the sentiment value of a word if one of the three predecessor is a negation word. This is a rule-based approach. Pröllochs, Feuerriegel and Neumann (2016) [14] improve this simplistic method by comparing different ways of detecting negation scopes in which an inversion of sentiment is necessary. Using a dataset of English language ad hoc announcements by German companies, they achieve good results for classification with reinforcement learning. Oliveira et al. (2016) [22] suggest not to invert the sentiment of a negated word completely, but to use an adjusted sentiment score. In their findings, negated sentiment words are often less intense than the affirmated ones.

To avoid the problem of missing words, Ren and Wu (2018) [18] use several word lists for their analysis. These were partially readily available, partially self-generated by using the SOPMI-method created by Oliveira et al. (2016) [22], to have a problem-specific list of negative and positive words as well as modifiers and shifters.

2.3.2 Machine Learning Approaches

In the recent literature, the use of machine learning applications gains more popularity. Lexicon-based approaches are often used as a baseline model to compare the results of the algorithm based ones to ([8], [17], [27], [30]). The latter are often found to be superior. Sometimes studies employ a hybrid model[8] ([8], [15], [20], [30], [31], [41]) or the two approaches for different subtasks ([13]). Important findings are listed in Table 2.

[7]Rhetoric Structure Theory, as the authors explain it, decomposes sentences or full texts. It focuses on hierarchical relationships between clauses and subclauses. Märkle-Huß, Feuerriegel and Prendinger (2017) use an automatic parser to derive a tree of units and their relations.

[8]Studies named here will be listed in Table 1 or 2 according to the main approach of how the sentiment value is derived.

Cite	Source	Goal	Measurement	Testing	Relevant findings
[8]∗	Annual and quarterly reports	Predict fraudulent reports	Classification into fraudulent or truthful	Comparison to lexicon-based predictions	Strong evidence for own metric to detect fraud in 1-year-samples and time series, outperforms lexicon-based methods
[17]	Message board	Sentiment Classification	Classification in pos/ neg	Comparison to lexicon-based	SVM[9] more accurate than NB[10] and LM[11]
[20]∗	Finance Message Board	Stock movement prediction	Weighted scores, Classification in pos/ neg	Classification quality	Topic modeling with sentiment analysis improves the prediction but needs further development
[21]	Tweets, Twit-Stocks, News headlines	Sentiment Classification	Classification in pos/ neg	Multivariate Regression	Fine-graining as a good way to increase quality, lexicon-based features important
[24], [42]	Turkish Finance Tweets	Stock movement prediction	Classification in pos/ neg	Correlation	Determine features of importance of a Twitter user, weight sentiment according to these features, very useful
[27]	Finance news headlines	Sentiment classification	Classification in pos/ neg	Comparison to lexicon-based	SVM delivers good results
[30]∗	Finance news (4 providers)	Stock Movement	Net sentiment/ classification in pos or neg	Trading Strategy	Sentiment lexicon for Chinese/ English created
[34]	News articles	Stock movement	Classification into pos/ neg	Trading strategy	Sentiment evaluation by crowdsourcing less useful than expert classification

[9] SVM=Support Vector Machines, SVR=Support Vector Regression

[10] NB=Naïve Bayes

[11] LM revers to the commonly used dictionary provided by Loughran and McDonald (2011) [36]

[32]	Online product reviews	Generating Questionnaires	Classification in positive or negative according to algorithm or weighted sum	Test dataset	Differences in attitudes among Chinese and American customers
[43]	Financial disclosures from German and European companies	Classification	Classification accuracy compared to other studies	Trading strategy	Feature engineering useful to improve classification, removes noise from data by proceeding only relevant parts
[44]	Tweets	Sentiment classification	Average score (LB), classification to sentiment class (ML)	Comparison to traditional approaches	Emoticons, slang and abbreviation increase accuracy, Lexicon-feature in ML outperform
[45]	NASDAQ news feed	Prediction of stock price and volatility simultaneously with sentiment	Net sentiment on document level, Classification	Non-parametric simulated Maximum-Likelihood Estimation	News exaggeration quickly corrected, but exaggeration sustains market turbulence
[46]	News and social media	Association/ Prediction of Exchange Rates	Sentiment indicators externally given	Multivariate linear regression and Multilayer Perceptron Neural Networks[12]	Non-linear model predict continuous returns better than regression, after market movements association increases between sentiment and exchange rate
[47]	Chinese financial reports	Prediction of CSR-rating	Classification to good or bad	Multivariate Regression	Dividing text into subcategories and extract aspect-related sentiment useful, CSR rating as non-financial performance indicator predictable

Table 2: Machine Learning based studies, ∗Hybrid model

[12] For a general introduction to the mechanisms of Neural Networks, see Witten, Frank and Hall (2011) [40].

Just as in lexicon-based approaches, sentence structures are taken more into account when training the algorithm. Feature selection[13] is an area under extensive research. Hagenau, Liebmann and Neumann (2013) [43] were among the first to systematically approach the advantages of feature engineering when analyzing financial statements. Novel is the idea to use a market feedback mechanism additionally to frequency-based features. By this, features that in fact made investors make a trading decisions were considered. By this, they were able to improve classification results substantially and followed a profitable trading strategy. Fine-graining the sentiment analysis is a promising research stream to improve sentiment classification further. Eliaçik and Erdoğan employ in their work from 2015 [42] and more in detail 2018 [24] a measurement to weight the sentiment of a tweet according to the influence and interest in finance of a user. They are the first to not only take the message itself into account but also who sends it and thereby acknowledging that it matters who writes about stocks. Others merely focus on what is written.

Fine-graining the features used also works on a semantic base. Semantic patterns are used by Meyer et al. (2017) [27] improve their sentiment classification. When deriving sentiment from news headlines, they pass the grammatical structure of the headlines as features to the algorithms. On the dataset of US-American finance news headlines, they achieve very good results for the classification task. For example, they find that stemmed headlines like 'Company name profit beat' most likely express positive sentiment. As explained earlier, Märkle-Hußet al. (2017) [15] combine lexicon-based feature engineering as well with a machine learning approach just as Kolchyna, Souza, Treleaven and Aste (2016) [44] do.

Fine-graining was also a subtask of the competition SemEval2017: Cortis, Freitas, Daudert, Hürlimann, Zarrouk, Handschuh and Davis (2017)[14] [41] list in their description of participants' solutions that hybrid models using both a lexicon-based and machine-learning/ deep-learning approach yielded best results. One of the highest ranking solutions, Jiang, Lan and Wu (2017) [21] for example, extract linguistic features like weighted and unweighted n-grams (word groups of length n), part-of-speech-tags (indicating the grammatical type of a word) and word clusters but also very detailed keyword-number-combinations and meta-data. Sentiment lexicon scores are included as one feature and rank among the most important features. The very good evaluation results come with one major downside: The trained

[13]'Feature selection' refers to the process of reducing the complexity of high-dimensional data as text data are often presented. It is used not only to 'clean' data but also to improve computing performance. As a caveat, extensive feature selection is prone to overfitting. Li, Cheng, Wang, Morstatter, Trevino, Tang and Liu (2018) [48] provide an overview of the goals of feature selection and commonly used algorithms.

[14]This text does not appear in the tabularized overview as it is seen as a literature overview.

model is, depending on the features chosen, very subjective to a certain type of text. Trained on tweets, the model is not useful in predicting the sentiment of a company's financial disclosure. This is true for all machine-learning-approaches that use a certain type of training data but is even more severe in this granular pre-processing.

The most important algorithms used in machine-learning sentiment analysis are still Support Vector Machines (SVM) and Naïve Bayes[15]. They are used by Hu and Tripathi (2015) [17], Creamer et al. (2016) [34], Kolchyna et al. (2016) [44] and Eliaçik and Erdoğan (2015) [42]. Nguyen, Shirai and Velcin (2015) [20] used only SVM for sentiment classification with lexicon-based features. Besides traditional algorithms, Day and Lee (2016) [30] test Deep Learning (Deep Neural Networks) for suitability and achieve good result compared to a lexicon-based baseline in order to exploit abnormal returns in a trading strategy.

All machine learning algorithms rely on a pre-labeled training dataset to learn from how to classify new input. Creamer et al. (2016) [34] use an in finance uncommon way to produce this dataset: They use clickworkers[16] to classify their training and test data. They compare these to other methods that are commonly used such as domain experts and pre-defined sentiments. In terms of the returns, crowdsourced sentiment as training input turns out not to be as successful as the sentiment given by finance experts.

Sentiment analysis is also combined with other Natural Language Process techniques to improve the classification and usefulness of the derived sentiment. Nguyen et al. (2015) [20] for example include topic modeling via Latent Dirichlet Allocation (LDA) or another method called JST. The idea behind this is that each text covers a number of hidden topics which are characterized by the use of different words. Similar to a factor analysis, these latent topics are extracted. Sentiments are only evaluated for certain topics. This approach was proposed by Dermouche, Kouas, Velcin and Loudcher (2015) [50] but without a focus on finance. Results for Nguyen, Shirai and Velcin (2015) are mediocre and need further research as the authors recognize.

2.4 Goals

In the finance domain, one of the major task is still the prediction of stock prices. Different approaches have been tried to predict the aggregate market in terms of indices. Special attention is given to newer investment categories,

[15] Naïve Bayes is a probabilistic model used for classification based one Bayes theorem of conditional propability. Support Vector Machines uses the idea to classify data by dividing them by a hyperplane. Both algorithms are described in Hastie, Tibshirani and Friedman (2009) [49, p.210ff, 417ff].

[16] Clickwork means outsourcing small tasks that typically can be done online, for example writing SEO texts, web research or as in this case tagging of given texts and pictures. The principle of crowdsourcing is used to avoid a personal bias.

namely the prediction of Bitcoin prices. Karalevicius, Degrande and De Weerdt (2018) [29] research on how sentiment in Bitcoin related news can be used to predict the price of the cryptocurrency. With a basic lexicon-based approach, they were able to yield abnormal returns but not in a magnitude that could outperform other methods of prediction.

Barunik, Chen and Vecer (2019) [45] implement an innovative approach by developing a high-frequency tool for sentiment analysis.The intuition is to relate the sentiment of a continuous news stream to a continuous stream of stock prices and volatility. For this, they develop a sentiment-driven stochastic volatility model. They find interesting patterns in news sentiment, for example the over- or understating in news is quickly reverted, but also relationships to price and volatility: The severeness of exaggeration in the news can sustain market turbulence by prolonging phases of volatility.

Besides the stock market prediction as target, new targets arise: Risk as a category expressed more in words than numbers is one major target to be predicted. This can be on microeconomic level for stocks or more on a macroeconomic level for the banking system. The authors take advantage of the fact that certain parts of companies' disclosures are describing the future and the sentiment expressed in these texts is therefore forward looking. This gives the opportunity to parse the information into a forward-looking variable for predicting future distress. This research stream has been provoked by work done before, for example by Wang, Tsai and Liu (2013) [7] who researched risk in terms of volatility of stocks. Tsai and Wang (2017) [6] develop this approach further as well as Wu and Olson (2015) [25]. Song, Wang and Zhu (2018) [47] look at CSR topics as a non-financial performance indicator and therefore risk metric for investors. For this, they relate the sentiment of a financial report of Chinese companies to their CSR rating. They found a strong connection by dividing the given text into subcategories according to management models.

On a macro-prudential level, risk prediction has been applied to banks' disclosures. Nopp and Hanbury (2015) [13] for example are able to predict financial distress for banks in a forward-looking manner. As their findings are robust on a macro level but lesser on individual banks, they suggest sentiment analysis as one tool to be used for banking authorities. Additional to a negative and positive word list, they use an uncertainty-word-list to derive risk perception from texts.

A similar approach was taken by Ghandi, Loughran and McDonald (2019) [12]. They are also able to find indicators for financial distress in banks' official disclosures. Besides the advantage of a forward-looking horizon, they emphasize that these disclosures are less prone to window-dressing given the risk of litigation. Other than Nopp and Hanbury (2015), they find a correlation even on an individual level. Both studies used a BOW-approach. A similar procedure is used by Law and Mills (2015) [9] on companies in financial distress: They find that firms in financial constraint tend to use more negative words as they engage in more aggressive tax planning activities. Law and Mills (2015) use a method introduced as a working paper version and later published by Bodnaruk et al. (2015) [10]. They develop a separate list of constraining words that indicate financial distress additional to their already existing word lists. The intuition is similar to the use of negative words: The more constraining words used, the more in financial distress a company is. Another approach was suggested by Purda and Skillicorn (2015) [8][17]: They develop their own textual analysis tool and metric in a hybrid fashion and compare its suitability to detect fraudulent 10-K reports with other metrics. They find that word list approaches, either using negative, fraudulent and litigious words, cannot compete. The developed metric was used in further research by Murphy, Purda and Skillicorn (2018) [38]: They find that language cues that give fraudulent behavior away can be transmitted by innocent co-writers without them realizing that fraud is happening. This is an important finding as the analyzed sections of annual reports as well as other disclosures are not written by one single author.

Apart from risk or stock prediction, foreign exchange markets are also under investigation. Crone and Koeppel (2014) [46] investigate the relationship of various sentiment indicators with the exchange rate return of Australian Dollar to US-Dollar. Their time series research suggests that prediction is possible to a good degree and that relationship is closest in times of market movement.

Different from other goals is the use-case researched by Ren and Wu (2018) [18]. They are the first to use sentiment analysis to detect herd behavior in financial markets. This means instead of making investment decisions based on fundamental research, investors tend to rather follow the market which leads to excess volatility and lower liquidity. They find this behavior by analyzing sentiments expressed in online forums talking about stocks. Just like in the measurement of risk, Ren and Wu use the forward-looking nature of the sentiment approach which is a big advantage compared to traditional numerical analysis.

Zhou et al. (2016) [32] use a machine-learning-approach for their research: They extend the existing literature on product reviews but their task goes beyond the classification. The authors generate questions and answers for online surveys. Sentiment analysis is used alongside other Natural Language Processing methods, here topic modeling. While topic analysis delivers questions and the corpus for answers, sentiment analysis is used to fill out these questions. By this, they produce insights in differences between Chinese and American customers without having to conduct a costly survey.

[17] As Purda and Skillicorn use a mixture of lexicon-based and machine learning approach in their own method and use the lexicon-based mainly for comparison, the study appears in the list for machine learning approaches.

As discussed, disclosures from companies are a commonly used source. Allee and DeAngelis (2015) [11] go beyond measuring the sentiment of their analyzed earnings conference calls. Given the sentiment words, they calculate the dispersion of positive and negative sentiment and are able to identify different strategies that managers are following when communicating good or bad news. The authors find that analysts and investors react more heavily to sentiment loaded news if they are spread over the whole call - for the good and the bad.

2.5 Missing and underrepresented research streams

Several topics in sentiment analysis have not or not in full advantage been applied to the domain of finance but could be useful to improve the analysis.

In the context of public health research, Shah, Martin, Coiera, Mandl and Dunn (2019) [51] present a method to distinguish baseline sentiments for time and location: They discover differences in the general sentiment for regions, cities, countries and times of day when analyzing tweets. This could be useful especially when predicting short term stock prices for specific companies as it is likely that baseline sentiment differs among them and that overall sentiment varies as well in the finance context according to the time of day. The value of derived tweet sentiment for prediction might be different for example if a tweet was posted in the home market like Germany for the DAX or somewhere else.

As social media are a commonly used source, the more pronounced work with medium specific words should be worked on: Emoticons and emojis often give a clearer indication of mood than words. A special sentiment lists for emoticons could extend a traditional word list. This is especially important when analyzing tweets and posts written by non-professional users. Kolchyna et al. (2016) [44] follow this approach. Including emojis, slang and and abbreviations in their sentiment word list increased accuracy and could be useful when sentiment is used for stock prediction[18]. Priya (2019) [52] follows a similar approach. Despite these efforts, authors working with social media posts mostly delete emojis as punctuations from their texts while preprocessing.

With earnings conference calls, non-written textual sources have been used. Other multimedia resources and social media channels have not been touched in the domain of finance. Corporate content published on Youtube or influential users on platforms like Instagram have been ignored so far.

Especially promising are approaches like Zhou et al. (2016) [32] to use other textual analysis methods alongside sentiment analysis. Social media posts might cover different topics regarding one stock/ company and many

[18] The authors used a dataset from a competition to predict stock movements but worked only on the sentiment classification. Nevertheless, the text is included in the literature overview.

publications offered by companies contain different subjects. Aspect related sentiments might differ even across one text. The method by Dermouche et al. (2015) [50] - extracting topics and related sentiment - needs further investigation.

Besides that the main focus of research is still on English language sources and US-American stocks. Even when German companies are in focus as in Pröllochs et al. (2016) [14], sentiment analysis has been conducted on English language texts. Using German language news or reports like in Remus, Ahmad and Heyer (2009) [53][19] is uncommon. Sentiment analysis on Emerging Markets has not been explored systematically so far. Also approaches to automatically translate either texts or resources like word lists might be worth developing.

3 Empirical

Sentiment analysis has shown to be a powerful tool in order to derive attitudes towards certain topics but also to proxy for risk perception. Both aspects are important in the context of challenges that firms and therefore investors are facing in the light of climate change. Risks in terms of stranded assets are often in the discussion as well as business opportunities that might arise from problem solving and establishing less emitting production processes. Furthermore, social aspects have become an important topic for companies.

In the last years, more and more firms have engaged in CSR policies but also in reporting about these. This has become an important source for textual analysis. As discussed above, it has not gained track in sentiment analysis, but in qualitative analysis. In a study that featured several years of environmental disclosures of French firms, Albertini (2014) [54] was able to derive four categories of disclosing companies. They are categorized by a company's willingness and strategy to adopt CSR policies. She showed that CSR reports and sections are therefore a viable source to detect the attitude towards these issues. Also the research conducted by Song et al. (2018) [47] showed the strength of financial report sentiment to be related to CSR rating scores.

In the following analysis, I am interested in a company's attitude towards CSR. Do they perceive upcoming environmental, social and ethical challenges as risk and burden (negative sentiment) or as an opportunity (positive sentiment)? For this, in my first experiment I will analyze not only CSR reports and CSR sections of annual reports but also CEO letters that are mostly included in the beginning of the annual report. CEO letters were chosen as they are carefully crafted and give special insights in impor-

[19] The article was published in 2009 and is therefore out of scope of this thesis. It is therefore not included in the overview table.

tant topics in the management of a company as Nopp and Hanbury (2015) [13] explain.

Additional to the sentiment analysis, a second experiment will be run: The resulting sentiment scores will be compared against rating scores that companies are given by independent agencies in order to assess their activities regarding Environmental, Social and Governance issues (ESG). These ratings are similar to credit risk ratings and are offered by several agencies. This comparison will give value to the sentiment analysis in terms of hints about the predictive power of this type of analysis.

The contribution of this empirical analysis will be threefold:

- Investigating CSR reports taps a seldom used source in finance. This yields insights in the content and style of these reports. Also German as a language is not often used in the finance domain sentiment analysis. This is a weakness, given that differences to English-language publications will be important given Germany's economic strength.
- By exploring companies' attitude towards CSR subjects, companies can be characterized and ranked in an aspect of their business activity that is not short term finance related but reflects companies' attitudes and willingness towards long-term uncertainty and risks. This can be included in overall risk management and long-term investment strategies.
- This baseline investigation can be used for further research: Variation over time could be analyzed as well as the predictive power of the sentiment of CSR reports. German companies have extensive obligations to report on non-financial aspects since 2017[20]. The reporting tone before and after this obligation was established can be investigated.

3.1 Data

For the experiments, annual and CSR reports from German MDAX companies have been collected. MDAX contains the 60 biggest companies in terms of market capitalization and exchange turnover that follow the DAX companies. The index was chosen as it includes less finance institutes than the DAX. It is said to be more representative for Germany's '*Mittelstand*' (medium sized businesses) economy given that it comprises "classic and technology sectors" [55, p. 6]. Furthermore, as these companies represent the manufacturing sector better, they are more directly exposed to issues related to climate change and migration than others. On the other hand, by adapting their supply chain and production, they have more influence on

[20] *Gesetz zur Stärkung der nichtfinanziellen Berichterstattung der Unternehmen in ihren Lage- und Konzernlageberichten (CSR-Richtlinie-Umsetzungsgesetz)* as in BGBl. I, p. 802ff, 20, 18.4.2017. It implemented the European directive 014/95/EU.

	AR-CEO	**AR-Sustain**	**SR-total**	**SR-CEO**
Texts	52	33	35	21
Total	117820	3485827	501644	11101
Average	2266	105631	14333	529
Median	1688	70528	12959	582
Minimum	168	2300	2298	160
Maximum	11604	576501	59840	1160

Table 3: Descriptive information on the sample: AR=Annual Report, Sustain= sustainability section, SR=Sustainability report, CEO=Letter of the CEO.

Germany's emissions. In terms of governance and diversity, given that they all have a crucial size, the companies in the sample have to deal with social and governance issues.

All the companies are publicly listed and have a significant size. They are therefore legally obliged to offer at least an annual report and a non-financial report which avoids problems with availability of data. In order to be included, companies must have been listed in the MDAX at the 28th December 2018[21]. The sample therefore reflects the composition of the index after the reformation in September 2018[22].

For the analysis, annual reports for the year 2018 (or 2017/ 2018) were included if they fulfilled at least one of two criteria: They had to have a CEO letter to the investors or interview and/ or a dedicated sustainability section. The section *'Nicht-finanzieller Bericht'* (Non-financial report) counted as such. CSR reports (*'Nachhaltigkeitsberichte'*) were included in the sample if they existed and were online available. If they contained an additional letter by the CEO to the investors, these are listed additionally. The reports had to be online available during the time of collection (March until End of June 2019). Only German language reports were taken into account which excluded some companies that only published in English. A list of all companies and reports that are included can be found in the Appendix in Table A1. Descriptive statistics on the sample can be seen in Table 3. 'Texts' refers to the number of texts that have been included in the sample. All other metrics are given as the number of 1-word-pieces (tokenized unigrams), short 'words'. The number of words is given after preprocessing (see next section).

[21] The composition can be inspected at the website of DAX indices: `https://www.dax-indices.com/zusammensetzung`

[22] `https://www.teleboerse.de/aktien/Boerse-sortiert-MDax-SDax-und-TecDax\-neu-article20633463.html`

	Positive list		Negative list	
	score	word	score	word
max	1	*gelungen* (nicely done)	-0.0042	*nervös* (nervous)
min	0.004	f.e. *Diskretion, Schuldlosigkeit* (confidentiality, guiltlessness)	-1	*Gefahr* (thread)

Table 4: Minimum and maximum scores given in the positive and negative word list used.

For the lexicon-based analysis, I use the dictionary provided by Remus, Quasthoff and Heyer (2010) [56] for German language sentiment analysis. It has been derived in a financial context and should therefore be suitable to be used in this domain. They have a positive and a negative list. A list of words of uncertainty like in the popular English-language LM-word list is not included.

A slight adjustment to the positive word list will be done to adapt the list to the context, as Allee and DeAngelis (2015) [11] suggest it: All words related to 'sustainability' ('*Nachhaltigkeit*') will be excluded. Especially when looking at sustainability reports and sections, these terms will be neutral. Counting the score into the overall score would be misleading. The final stemmed positive list includes 3.402 words, the stemmed negative list 3.968. In Table 4, minimum and maximum values for possible scores are shown. The minimum positive score is achieved by 1.161 words.

To account for negations, the lexicon provided by Schulder, Wiegand and Ruppenhofer (2018) [57] has been used. The authors provide a list of more than 400 verbs in German that indicate a polarity shift. If a word exists in both the sentiment list and the polarity shift list, it is kept in the sentiment list and excluded from the polarity shift list. The words *nicht* (not), *kein* (no) and *niemand* (no one) have been added as non-verb shifters.

The stopword list for German provided by the 'lsa'-package ([58]) in R is used. It contains 370 words. Stopwords are words that are important to make a text readable, but do not carry the sentence meaning.

ESG data are provided by Thomson Reuters database Eikon ([59]). For the analysis a broader spectrum of variables were retrieved to find possible associations[23]: The ESG Score, ESG Controversies Score and ESG Combined Score are interrelated as the combined score is a mixture of the first two: Overall ESG score is given according to reported activities in environmental, social and governance issues, the controversies score is given according to appearance in negative press articles about public controversies

[23]The following explanations can be found more in detail in the Thomson Reuters ESC scores methodology (2019) [60].

in these issues. The ESG score is derived from three pillar scores: the Environmental, Social and Governance pillar score. They are given according to the performance in several subcategories.

Additionally, the CSR Strategy score is included in the analysis. It is one of the subcategory scores of the Governance pillar and reflects a company's approach to report on their Corporate Social Responsibility activities in everyday business life. It has been singled out as the investigated reports are a crucial part of this communication strategy that is in focus of the score.

As one last factor, the emissions score is added. This reflects a company's ability to reduce emissions in the process of manufacturing or delivering. It is used as one of the key factors to mitigate climate change. It is one of the subcategories to the Environmental pillar score.

To get more context one the companies included, the operating sector according to North American Industry Classification System (NAICS) are added[24]. These are available from Thomson Reuters Eikon [59] as well. in the sample are companies from the following sectors:

- Arts, Entertainment, and Recreation
- Construction
- Finance and Insurance
- Information
- Manufacturing
- Mining, Quarrying, and Oil and Gas Extraction
- Professional, Scientific, and Technical Services
- Real Estate and Rental and Leasing
- Retail Trade
- Transportation and Warehousing
- Utilities
- Wholesale Trade

The exact mapping can be seen in Table A2 in the Appendix.

[24]The classification and ESG scores have not been available for all companies in the sample. They have been added only as available.

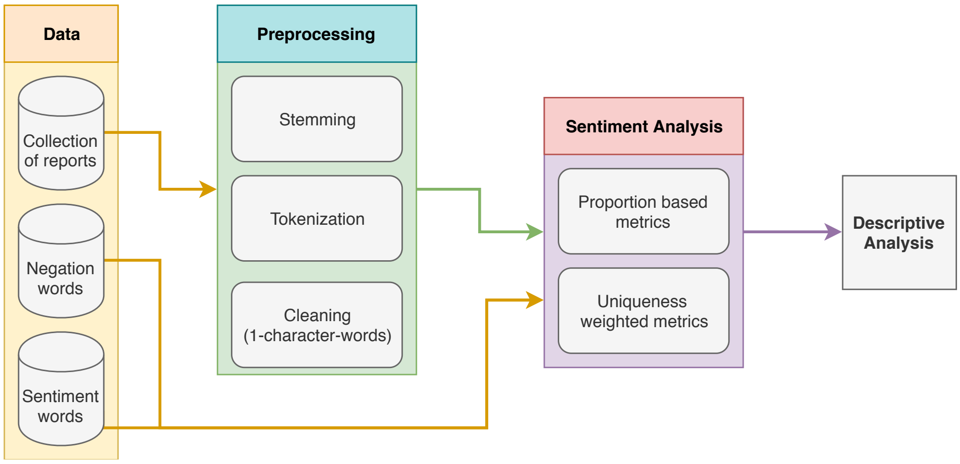


Figure 2: Research model using lexicon-based sentiment analysis.

3.2 Methodology

To conduct the sentiment analysis, I will use a lexicon-based approach. This uses the advantage of the method to be an unsupervised method without the need to extensively label texts manually as it would be necessary to conduct a machine learning sentiment analysis. This is not feasible: Using another text sort, like product reviews, as training data is not useful as the linguistic features are very different. To label texts manually is not viable as personal bias will be too big and it would be out of scope of this paper. No pre-labelled dataset of German CSR-reports and annual reports is available. For the lexicon-based approach, several improvement strategies similar to those explained earlier in section 2.3.1 are used to mitigate the disadvantages of the method against a machine learning approach.

The research model is shown in steps in Figure 2. First, the collected data will be preprocessed by using the R package tabulizer [61].The text will then be tokenized to unigrams (1-word-pieces) with the tidytext package ([62], [63]). Words that consist of one character only will be deleted. Stopwords like the articles *der, die, das* are excluded. In order to reduce complexity, a stemming algorithm is used: The 'tm'-package for text mining [64] performs Porter's stemming algorithm [65] which has been improved and adapted to the German language since its original publication in 1980[25]. Lastly, sentiment scores as given in the word list and negation indicators are added.

To account for negations and possible shifts in meaning, a lexical proximity approach is used. It is based upon the aspect level sentiment metric that was used by Zhou et al. (2016) [32]. The authors determined whether a sentiment word is referring to a certain topic by calculating the distance as

[25]Negation and sentiment words will be stemmed, too, to keep them coherent with the dataset. For reasons of clarity, this is not shown in Figure 2.

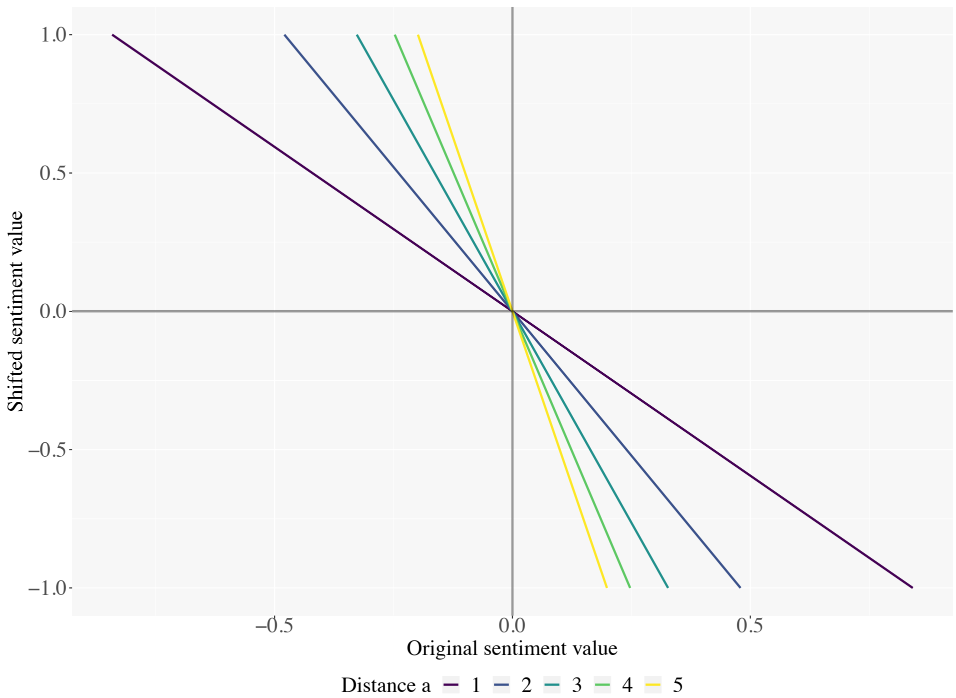


Figure 3: Negation handling

the number of words between the two. The further it is away, the lesser it is weighted. Translated to negation this means: The bigger the word distance between a negation and a sentiment word is, the less likely a shift of the meaning of a sentiment word is. Taking the comparison of negation rules done by Pröllochs et al. (2016) [14], the threshold will be set to 5 words to decide on shift or not. 5 words is also the window used by Hagenau et al. (2013) [43] in which a 2-gram must be found and it was argued to be a good threshold to avoid spilling over into different subclauses. In the given context of negation this means, if a negation word appears in a 11-word-window around the sentiment word (5-sentiment word-5), then the sentiment will be shifted by multiplying the value by $-sin(1/distance)$. This approach is taken from Oliveira et al. (2016) [22] who argued that a shifted sentiment word has not the same strength in the negated sense as in the affirmated. The negation handling is illustrated in Figure 3. It shows the transformed as well as the original value: a high positive original sentiment with a negation word directly before or after (a=1) will become a little less high negative score. The same original score will be transformed to a small negative score with a negation word being five words apart.

After this preprocessing, several sentiment metrics will be calculated:

- Negativity in percent of a document as the proportion of the count of negative words w_{neg} relative to the count of total words in the document (without stop words) w_d will be computed alongside the positivity as can be seen in equation 1.

$$negativity_d(positivity_d) = \frac{\sum_{i=1}^{n} w_{neg}(\sum_{i=1}^{n} w_{pos})}{w_d} * 100 \tag{1}$$

$$netsentiment_d = positivity_d - negativity_d \tag{2}$$

- The sum of negative and positive words relatively to the total word count will be calculated as net sentiment. Positive words count as 1, negative words as -1 or, negativity will be subtracted from positivity as in equation 2. If the calculated value is above 0, net sentiment is positive, if below 0 net sentiment is negative.

- Further on, to account for uniqueness of words used in certain reports only, the sentiment scores of each word-report pair will be weighted by the TF-IDF score. This score is given by the frequency of a term in a given text (tf) and the inverse of the frequency of the term in the corpus of all analyzed document (idf) [63]. In equation 3, D denotes the number of all documents in the corpus, D_i the number of documents containing the specific word w_i. w_d is the number of all words in the document.

$$tfidf_i = ln(\frac{D}{D_i}) * \frac{\sum_{m=1}^{n} w_i}{w_d} \tag{3}$$

TF-IDF reflects the relative importance of a word in one document contrasted with the importance in the overall dataset. Based on this, equation 4 shows the calculation of the weighted sentiment scores:

$$sentiment_i = sentiment_g * negation_i * tfidf_i \tag{4}$$

$sentiment_i$ is the individual sentiment score for a word in word-report pair, $sentiment_g$ the general sentiment score as given in the word list. $negation_i$ is the negation value to be used: it is 1 if a negation word is more than 5 words apart or $-sin(\frac{1}{a})$ with a being the distance value between 1 and 5. $tfidf_i$ denotes the individual TF-IDF score of the given word in the given document. These values will be added for the sum of positive, negative and total sentiment with $totalsentiment_d$ being the sum of all k weighted sentiment scores, may they be positive or negative, in a document as in equation 5.

$$totalsentiment_d = \sum_{n=1}^{k} sentiment_i \tag{5}$$

By this, six metrics on a document level will be calculated for each company-report pair: negativity, positivity, proportional net sentiment in percent, weighted positive sentiment score, weighted negative sentiment score and weighted net sentiment score.

4 Results

4.1 Sentiment analysis

4.1.1 Complete sample

Figure 4 shows the proportional net sentiment (positive minus negative words in relation to total word count). Color is given according to the weighted sentiment score. The inversed graph, showing the sentiment scores colored according to the proportion scores are shown in Figure B1 in the Appendix.

Proportional net sentiment is always positive: All analyzed reports contain more positive than negative words. CEO letters contain more sentiment words than sustainability reports and sections. Sustainability reports are more emotional than sustainability sections. Sustainability sections in annual reports achieve the lowest percentage of sentiment words compared to the other text sorts.

Weighted sentiment scores show the same picture: Highest values are achieved in the CEO letter of the annual report, second most in the CEO letter of the sustainability section. Weighted sentiment scores in sustainability texts are mostly negative which means that individual negative sentiment words dominate.

Weighted scores do not perfectly correlate with the proportional net sentiment. This would have shown in the color gradient of the bar plot. But the graph shows that CEO letters also have higher sentiment scores according to this metric than sustainability texts. From the proportion and the score alike, overall sentiment for the sustainability texts is homogenous on a low to negative level.

A scatterplot like in Figure 5 shows that there is a relation between the two net sentiment metrics but it depends on the type of report. While the correlation is positive and mediocre to strong for both CEO letters, no association can be found for the two other text types. Bravais-Pearson-correlation coefficient emphasizes this. As this correlation coefficient is sensitive to outliers, the Spearman Rank Correlation coefficient ρ is computed as well which is based on ranks not absolute distances between observations.

The correlation is strongest for the CEO letters in the sustainability reports as can be seen in Table 5. Both the correlation and the rank coefficient are highest. In general, sentiment metrics in CEO letters are related. For the sustainability texts, relationship is not existing which also shows in the

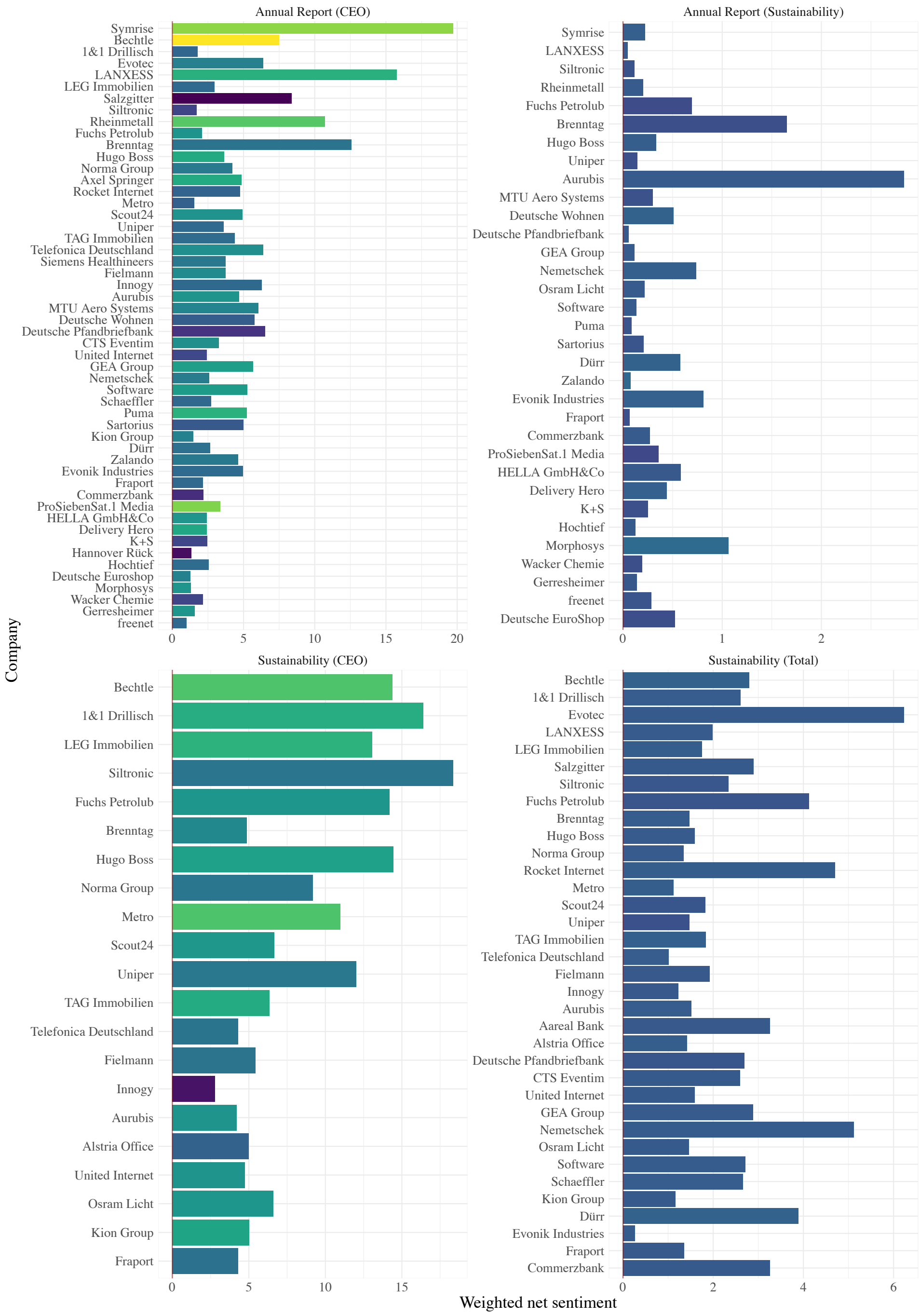


Figure 4: Net Sentiment in percent per report type and company, colored by the weighted sum sent.

flat line of the graph in Figure 5. Values differ in negative and positive direction according to the two metrics. Absolute values are close to 0.

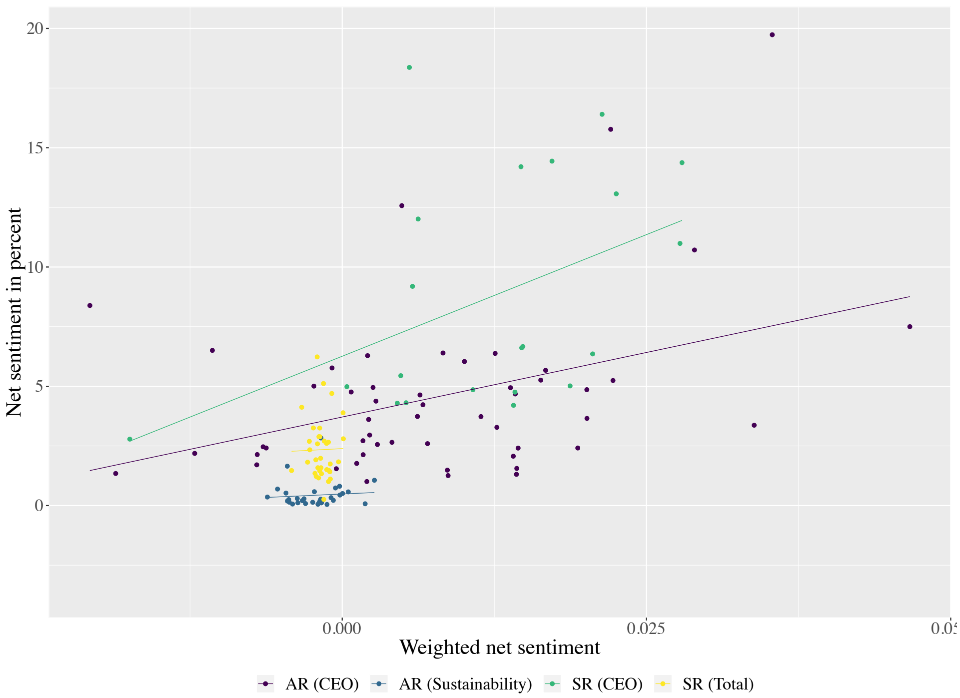


Figure 5: Relationship between proportion of sentiment words and weighted net sentiment.

The Rank Correlation shows that the strong Bravais-Pearson-coefficient for annual report's CEO letters is driven likely by outliers: The Spearman coefficient is only moderately positive in this case compared to a mediocre relationship according to the correlation coefficient.

Spearman's ρ correlation is significant for both CEO letters. Pearson correlation is significant on a 95%- confidence interval for the Sustainability CEO letter and on a 99% confidence interval for the annual report CEO letter.

Report	Type	Correlation	Spearman's ρ
Annual report	CEO	0.3890∗∗	0.2800∗
Annual report	Sustainability	0.0859	0.0949
Sustainability report	CEO	0.4410∗	0.5520∗
Sustainability report	Total	0.0183	−0.0585

Table 5: Association between the sentiment metrics Net sum weight and Net sentiment (proportional) per report type. ∗p-value < 0.05, ∗∗p-value < 0.01

4.1.2 Annual reports

Looking at annual reports more in detail, the overall picture is confirmed: CEO letters are far more emotional and sentiment driven than the sustainability sections of the same reports. This can be seen in the first two rows of Table 6. Net sentiment as proportion of sentiment words diverges by factor 10 for the CEO letters from the sustainability section.

In Figure 6, the relationship is shown on a company level. The scores diverge to a large degree. Whereas CEO letters are mostly positive, the average score for the sustainability sections are negative. This is mostly due to low positive sentiment scores: Negative scores determine for most companies the overall sentiment score. Positive values are nearly steady on a low level. This split between positive (Figure B2) and negative (Figure B3) sentiment scores is shown in the Appendix.

The company with the highest weighted sentiment score in the CEO letter is Symrise. This is driven by comparatively high weighted positive sentiment (see Figure B2 in the Appendix). Symrise uses quite unique words in high frequency as can be seen in Figure 7. The barplot shows the most important sentiment words in the CEO letter. These are adjectives like: *stolz* (proud) and *spannend* (exciting). The adjective *talentiert* (talented) refers to employees. Other than in aggregated values, a correlation between the original sentiment and the weighted sentiment score seems to hold as the color gradient fits. In the list of the most important sentiment words for Symrise are also the words *Duft* (fragrance) and *bunt* (colorful). These two are problematic: Symrise is a company that produces flavors for perfumes as well as foods. The words are most likely in the company's texts to describe their business model and are therefore neutral terms. Counting them as sentiment words is misleading and illustrates a downside of the lexicon-based approach with a fixed list of words.

Nearly as high as Symrise ranks ProSiebenSat.1 Media. Words that are causing this ranking are among others *begeistert* (enthusiastic), *perfekt* (perfect), *gelung(en)* (stemmed for successful). Once again, a potentially problematic term appears: As a media company, *kreativ* (creative) might

Report	Type	Positivity	Negativity	Net sentiment	Weighted pos	Weighted neg	Weighted Net
AR	CEO	5.4776	0.9246	4.5530	0.0177	−0.0099	0.0078
AR	Sustain	0.6040	0.1691	0.4349	0.0035	−0.0057	−0.0022
SR	CEO	9.8059	1.0756	8.7303	0.0201	−0.0080	0.0121
SR	Total	3.3617	1.0191	2.3426	0.0023	−0.0040	−0.0017

Table 6: Mean Values for sentiment metrics per report type.

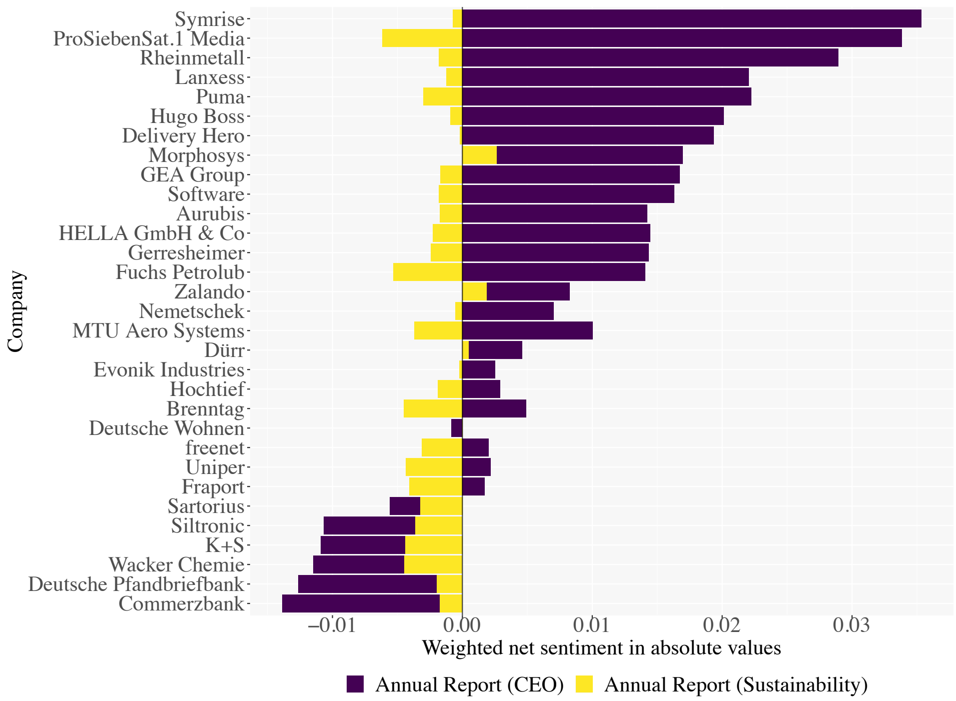


Figure 6: The Weighted Net Sentiment of two types of texts given in the annual report.

not be as positive as it is in the common use. At least partially, it is part of the business model of ProSiebenSat.1 Media to be creative.

The worst sentiment in annual report CEO letters is observed for Commerzbank. This result is driven by positive weighted sentiment words that have very low sentiment scores. Words used like *erfreulich, gut, wichtig* (pleasant, good, important) have low sentiment value from the beginning. The low weighted scores tell that the uniqueness is also very low in the sample of documents.

In 9 out of 30 examples, the net weighted sentiment does not diverge for the sustainability section and the CEO letter. Six companies get negative sentiment scores for both reports, only 3 are all positive (Morphosys in Professional, Scientific, and Technical Services sector, Zalando in Retail Trade and Dürr in Manufacturing). The overall negative companies work in different sectors as indicated by the NAICS classification: Finance and Insurance (Commerzbank, Deutsche Pfandbriefbank), Manufacturing (Wacker Chemie, Siltronic, Sartorius) and Mining, Quarrying, and Oil and Gas Extraction (K+S).

In the majority of cases, the CEO letter is positive and the CSR section negative. A rough segmentation can be given as finance-related compa-

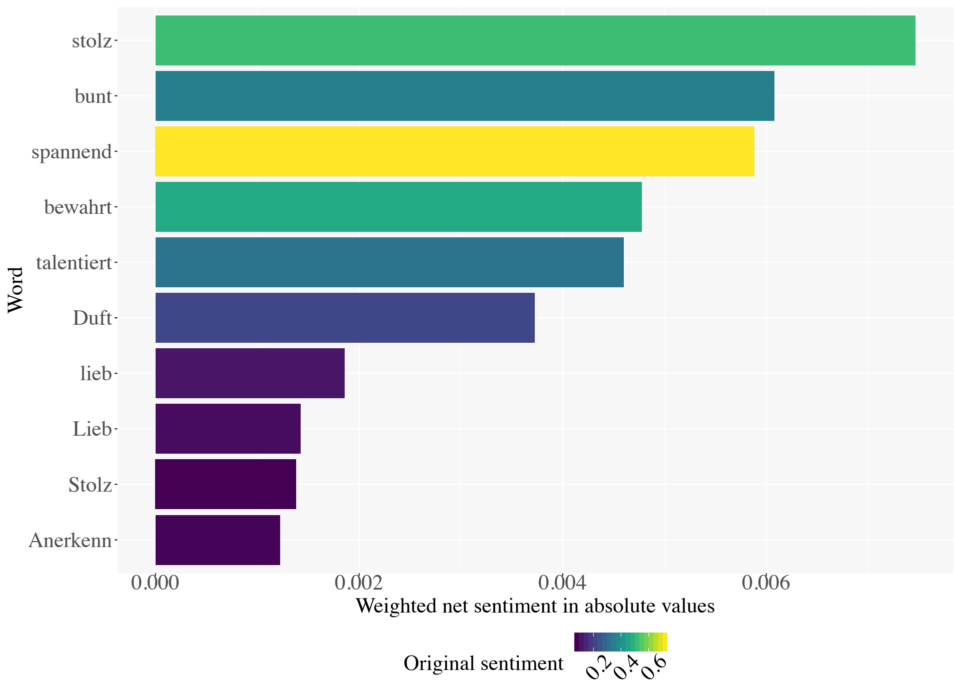


Figure 7: Most important sentiment words for Symrise

nies (Commerzbank, Deutsche Pfandbriefbank) achieve negative sentiment scores in their CEO letters rather than other firms.

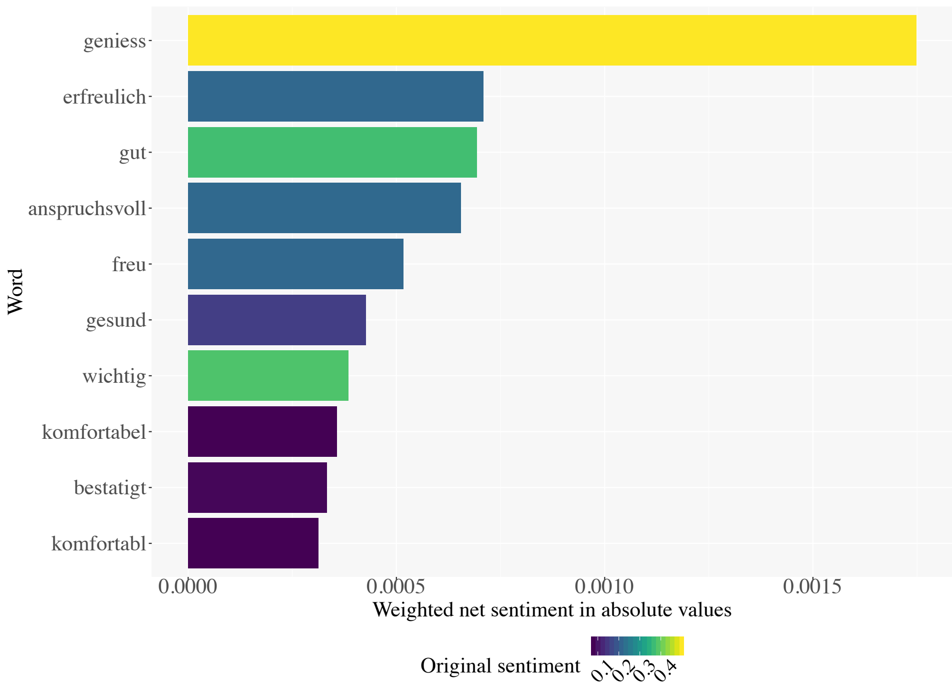


Figure 8: Most important sentiment words for Commerzbank.

4.1.3 Sustainability reports

Sustainability reports give a similar picture to the overall annual reports. As can be seen in Figure 9, most companies create a positive sentiment in their sustainability CEO letters. Overall sentiment of the sustainability reports is mostly negative and diverges therefore from the sentiment of CEO's remarks. The most positive one is provided by the IT provider Bechtle. The by far most negative is the energy provider Innogy, a child company of German energy supplier RWE. It is also the only that reaches negative scores in both weighted net sentiment and proportional net sentiment.

As the CEO letter is part of the overall sustainability report, a diverge in the sentiment of these two means a gap between the introducing word to the rest of the report. In order to create a sentiment for the total report opposite to the CEO letter, the net sentiment of the CEO letter has to be counterbalanced. Mostly, this is observed: Only the most positive CEO letter published by Bechtle, is accompanied with a positive sustainability report in total. The value is very close to 0 with 0.0000945. At the other side of the spectrum, Innogy has both a negative CEO letter and total report.

Looking at the sector distribution does not give a clear indication of who is getting negative sentiment: The most negative total sustainability report

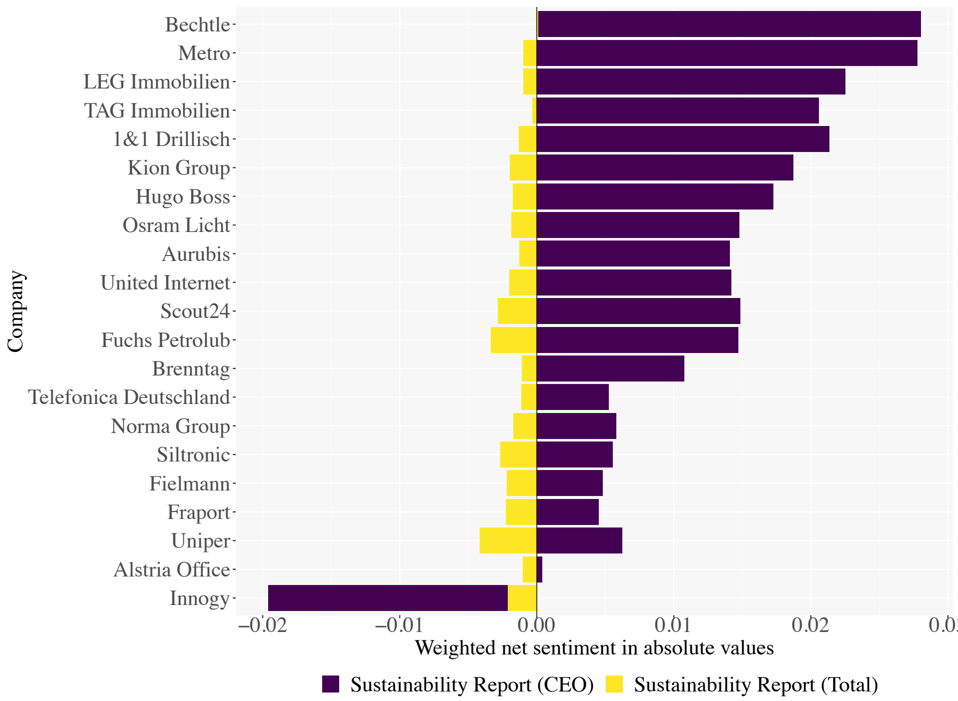
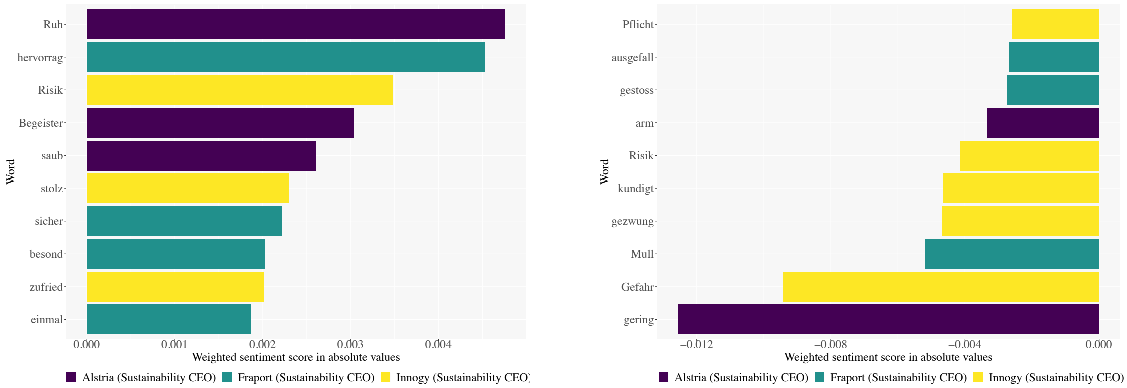


Figure 9: Weighted net sentiment for sustainability reports and CEO letters inside of them.

is offered by Uniper in Utilities, followed by Fuchs Petrolub and Siltronic (both Manufacturing). As these sectors are broad in their classification this is not a clear indicator: Other firms in Manufacturing in the sample achieve higher values. For the most positive (or rather least negative) sentiment value, the picture is clearer: It is lead by Bechtle, operating in the sector of Professional, Scientific, and Technical Services. Companies in the sector Real Estate & Rental & Leasing are following: TAG Immobilien and LEG Immobilien. These are also among those with the most positive CEO letters in the CSR report.

Looking at aggregated values, differences between the overall sentiment of sustainability reports to those in the annual report section, as can be seen in Table 6, become clear: In sustainability sections, the proportion of negative and positive sentiment words is more equal, which lowers the proportional net sentiment. Sustainability reports have a higher proportion of positive words compared to negative ones: They differ by more than factor 5. Negativity differs by factor 6. Weighted sentiment scores at the same time are nearly equal. The more used positive words in sustainability reports therefore seem to be shared by the most reports in the sample and not very unique.



(a) Words with highest weighted sentiment score.

(b) Words with lowest weighted sentiment score.

Figure 10: Sentiment words

Differences appear also in the introducing section of the CSR report: Figure 10 shows the most important sentiment words for the three least positive CEO letters in a sustainability report. Among the most positive is one inverted (*Risik(o)* (risk)) used by Innogy, furthermore mostly adjectives like *hervorrag(end), saub(er), stolz, sicher, besond(ers), zufrieden* (outstandig, clean, proud, secure, special, satisfied). Absolute values are higher for negative sentiment scores than for positive ones, which explains the overall negative sentiment of the reports - despite more positive than negative words in total.

The negative sentiment words on the other hand contain more nouns: *Pflicht, Risik(o), Müll, Gefahr* (obligation, risk, trash, thread). Second most are adjective, followed by verbs and participles which is a difference to the mostly adjectives in the positive values.

One short paragraph illustrates what makes Innogy's CEO letter so negative:

> *Trotz aller Erfolge und dem Wirken unserer über 46.000 Mitarbeitenden wissen wir: Der Weg in eine noch nachhaltigere Welt verläuft nicht immer geradlinig. Wir können nicht zufrieden sein mit einem steigenden Energieverbrauch oder einer wachsenden Menge an Abfall in unserem Konzern, und wir sind es auch nicht. Hier sind wir in der Pflicht.*[26]
>
> (In spite of all the successes and the work of our 46,000 employees, we know: The path to an even more sustainable world is not always straightforward. We cannot and are not satisfied

[26] CSR report Innogy, 2018: https://www.innogy.com/web/cms/mediablob/de/3947110/data/0/5/nachhaltigkeitsbericht-2018.pdf

with rising energy consumption or a growing amount of waste in our Group. We have a duty here.)

Innogy's CEO is makes it clear that the company has a difficult challenge before them. At the same time, before this paragraph, he describes an already difficult time that lies behind with protests in Germany against the company, a cancelled acquisition and a general tough energy market.

On the opposite, Bechtle describes a comfortable situation for the company:

> *Wir sind, wie ich finde, auf einem sehr guten Weg. Vieles ist bereits erreicht - weitere, ambitionierte Ziele sind gesetzt.*[27]
>
> (I think we're on a very good path. Much has already been achieved - further ambitious goals have been set.)

In general, it is striking that Bechtle's CEO describes their path in very general terms whereas Innogy's CEO is very specific on the company's challenge and difficulties in the past as well as in the future.

4.1.4 CEO letters

Proportional net sentiment is positive for all CEO letters, no matter the position of the texts (Figure B4 in the Appendix). Differences appear in the weighted sentiment scores for these letters as can be seen in Figure 11.

Sentiment scores for annual report's as well as CSR report's CEO letters have been described above. Observed patterns are similar nad sentiment for the CEO letters is mostly positive.

Four companies show a diverging pattern: Innogy (energy supplier) has the only negative CEO letter for sustainability letters while the annual report letter is slightly positive. In three other cases the picture is the other way round: SR's CEO letter is positive; AR's letter is negative. This diverge is given for Siltronic (Manufacturing), United Internet (Information) and Metro (Retail Trade).

Bechtle achieved the highest score for annual report's CEO letters and is among the highest for the SR's. This position is divided with CEO letter by Metro - which had a negative sentiment for the annual report. For comparison, both sentiment values are added up for those companies in the sample that provide both. The result can be seen in Figure 12.

This shows the outstanding position that Bechtle takes in the sample: The company was already in the lead with the positive sentiment for the annual report letter.The high positive value adds up to a sum that is double as high as the following value for Hugo Boss.

[27] CSR report Bechtle, 2018: `https://www.bechtle.com/dam/jcr:7155f829-820e-4fe3-8e3c-2abab596c671/Nachhaltigkeitsbericht_2018_de.pdf`

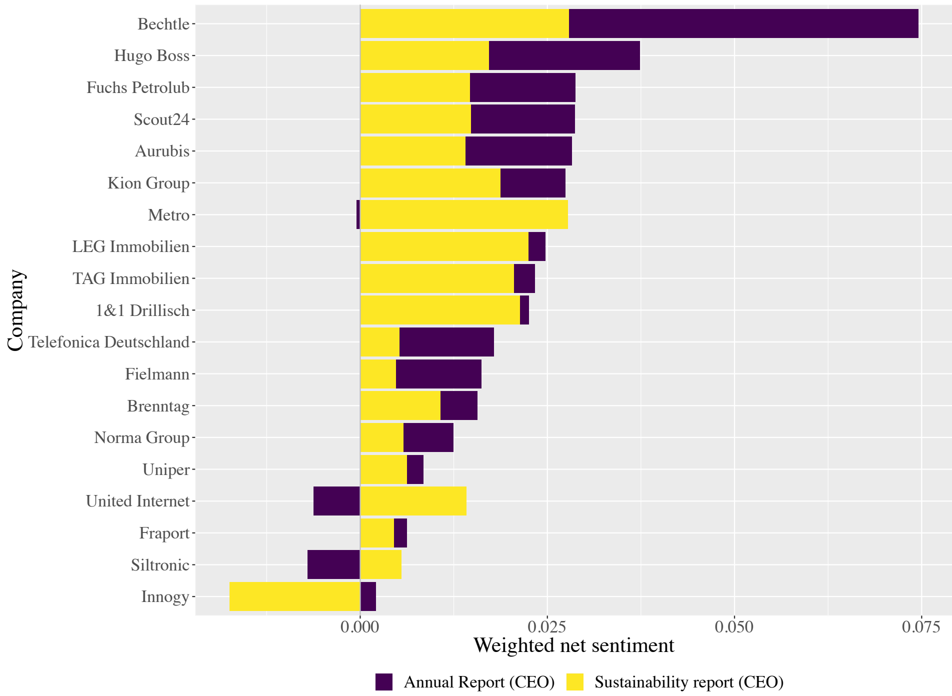


Figure 11: Weighted net sentiment in comparison for CEO letters.

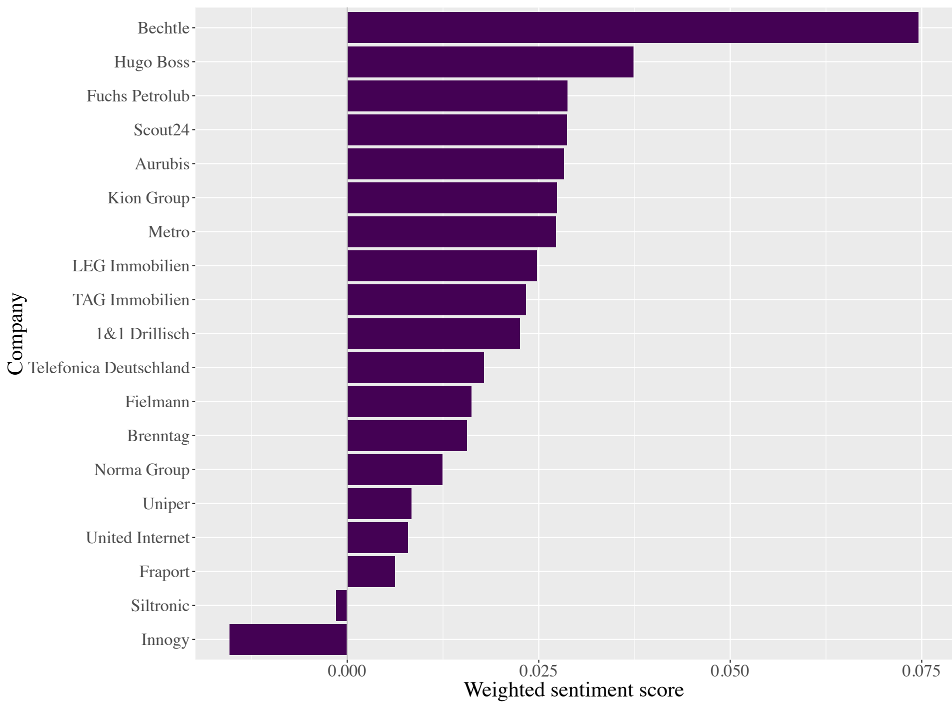


Figure 12: Sum of weighted sentiment scores for CEO letters per company

4.2 Correlation to ESG data

Correlation between sentiment scores and ESG related scores is shown in the following section. The Bravais-Pearson and Spearman coefficients are calculated per report type. As they are similar but Spearman's ρ is more robust to outliers, the ρ-values are given in the Appendix, in Table A3 to A6. Values below 0.25 are ignored and not tabulated as no association can be found.

Highest score in absolute values that is overall achieved in the sample is a Spearman rank correlation coefficient of -0.6472 (Pearson: -0.6992) between net sentiment in percent in Sustainability Reports as a total and the CSR Strategy score from the Thomson Reuters Eikon data base. The association is shown in Figure 13. This means, the lower the net sentiment the higher is the CSR strategy score. Nearly as high is a positive correlation of 0.6241 to the number of total words which shows in the color of the points in the scatterplot in Figure 13: The more words a company takes in the sustainability report, the higher the score it achieves for CSR strategy.

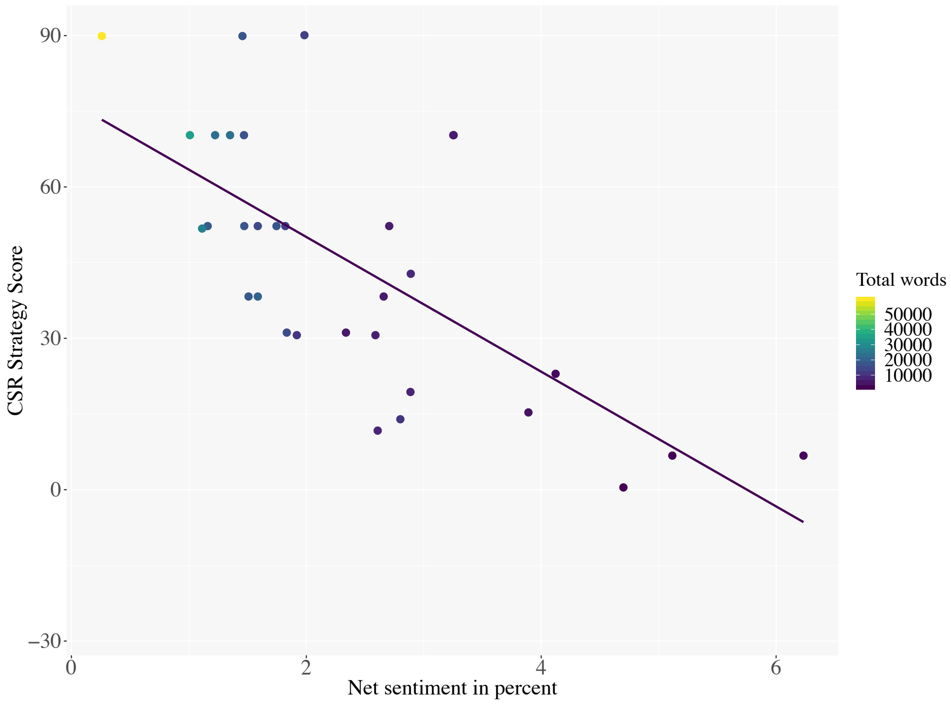


Figure 13: Correlation between Net Sentiment in Percent and CSR Strategy Score

Based on correlation coefficients in absolute terms, the highest average association is achieved in sustainability report CEO letters[28]. Highest ab-

[28]Values below 0.25 have been ignored.

solute values as described above in sustainability reports. The strongest correlation for the base ESG score is found with the CEO letters in the sustainability section for sum of the weighted negative scores: the Spearman coefficient ρ is -0.4525. The found association is mediocre and negative. A graphical representation of the association can be seen in Figure 14. The darker the color the lower the absolute value of ρ.

Figure 14: Heatmap for correlation coefficients ρ between ESG ratings and sentiment metrics for the CEO letter in the Sustainability report.

For the combined ESG score the strongest correlation found is a positive one to the net sentiment in percentage of the CEO letters in the sustainability reports: With ρ =0.3838 it is mediocre as well.

The emissions score correlates stronger with two values: the total word count (0.5236) and negativity of the sustainability reports (-0.5484). Both are mediocre to strong. But for both annual report texts, no correlation to the emissions score was found.

Pillar scores correlate rather weak with the sentiment scores for any given text type. The strongest associations can be found with the sustainability reports and the contained CEO letters. Correlations are stronger for the environmental pillar scores than for the social pillar scores. Weakest to non-existent is the association with the governance pillar.

Particular interesting is the relationship to the ESG Controversies score as a main indicator for risk related to ESG issues: Being in the news on controversies around ESG issues can have financial (decrease of demand for products) and non-financial (loss of reputation) consequences. Sustainability report sentiment metrics as well as sustainability sections metrics have no correlation to this important measure. For the CEO letters of annual reports, the only correlation found is to negativity in percent (ρ =0.2725) but mediocre. CEO letters of the sustainability report correlate to a certain degree. But all four metrics are rather weakly to mediocre associated: sum of negative words (-0.3374), negativity in percent (-0.2601), net sentiment in percent (0.3096), and sum of weighted negative sentiment (0.3498).

Correlation analysis does not show a clear pattern as to how sentiment translates into ESG rating scores. To gain more insights on this, a cluster analysis is run on a subset of CEO letters in sustainability reports. These are the text types with the highest absolute correlation coefficients. At the same time, Hopkins statistic that determines cluster tendencies in a dataset is highest for this subset[29]. With H=0.43, it is close to 0.5 which is the threshold used to characterize randomly distributed data.

Two different cluster algorithms, hierarchical Ward and k-means[30], are used. The first is used to derive the number and centers of clusters used as starting point for the second.

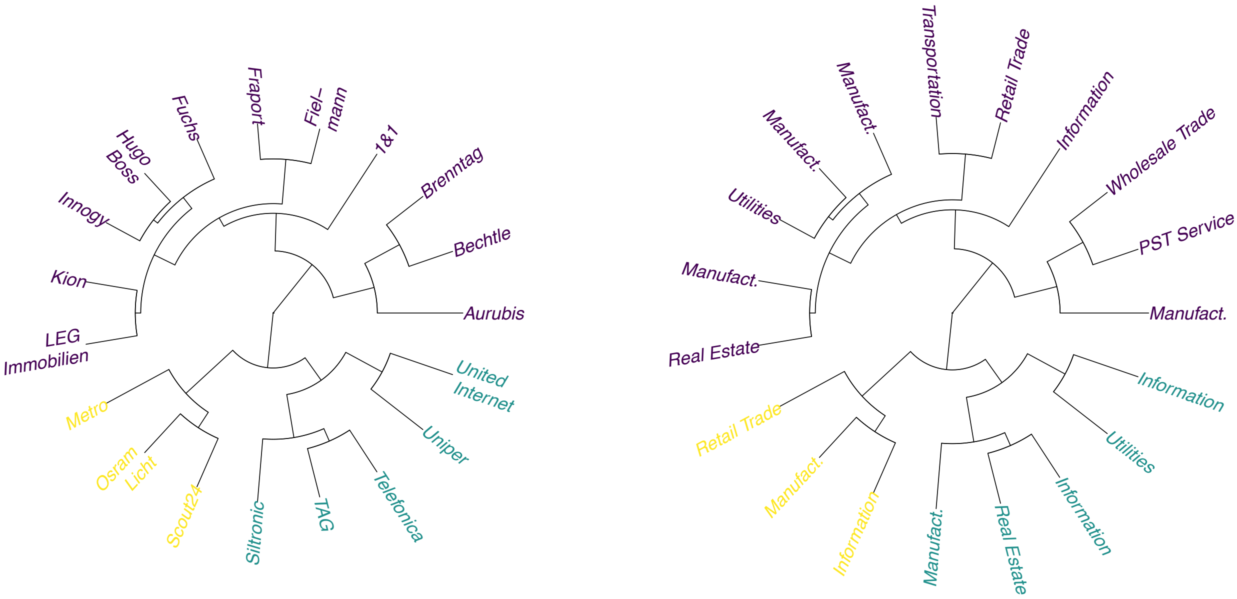
According to a dendrogram, three to four clusters are a good solution. As the fourth cluster would consist of only one observation, the three-cluster-solution is used. The outcome is shown in Figure 15. In this case, both cluster algorithms produce the same outcome.

As to what the clusters mean, the same picture as before is given: From Figure 16, a clear division in cluster membership is made by the net sentiment given in percent. Companies in cluster 1 have a outbalanced proportion of negative and positive words whereas in cluster 2 and 3, companies have a higher proportion of positive words in their CEO letter of the sustainability report. According to the ESG Controversies score, no clear indication can be given. It is clear that cluster 3 companies achieve a high score (all above 50) but are mixed together with cluster 3 companies regarding their net sentiment.

This is emphasized by looking at correlation for cluster membership: As can be seen in Table 7, cluster membership is associated with sentiment related metrics. Spearman's correlation coefficient ρ is highest for the total word count, the positivity as well as the net sentiment in percent. The

[29]Hopkins statistics uses a hypothesis test to check whether a dataset is generated by a uniform distribution. The test statistic H is between 0 and 1, 0 meaning a dataset is highly skewed and 1 a dataset is highly clustered. The test is derived in Han, Kamber and Pei (2012) [66, p.484f]

[30]The first starts with each single observation as cluster and agglomerates the nearest two to the next level cluster. This is performed until all observations are member of one cluster that is the whole sample. The second one has a fixed number of cluster that is to derive. Starting from a selected or randomly chosen point, all observations are attributed to one cluster given the mean distance. The algorithm basics are described more in detail in Han et al. (2012) [66, p.394ff and 451ff]



(a) Clustered companies

(b) Clustered sectors with abbreviated sector names

Figure 15: Clustered Sustainability CEO letters, color according to cluster

relationship is positive. For the ESG related scores, the correlation to the emissions score is highest in absolute values but mediocre in comparison to the sentiment correlation. None of the ESG related metrics has a significant correlation on the 95%-confidence interval whereas four sentiment metrics are significant on a 99%-confidence interval.

The clustering of the data is also not related to the business fields companies are operating in. This can be seen in Figure 15, Panel b. 'Manufacturing' and 'Information' is found in all three clusters. The two companies in the sector 'Real Estate & Rental & Leasing' are found in two different clusters.

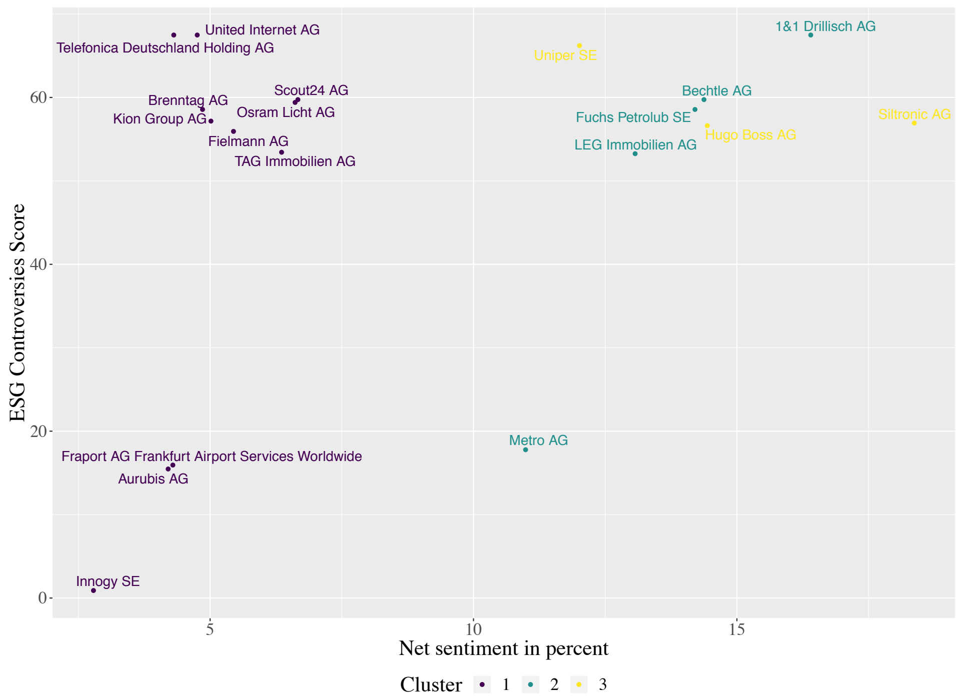


Figure 16: ESG related scores given a cluster.

Sentiment metric	ρ	**ESG metric**	ρ
Total word count	$-0.7919**$	ESG score	-0.0396
Sum positive words	$-0.4841**$	ESG Controversies	0.1190
Positivity	$0.8628**$	ESG Combined	0.1900
Sum negative words	-0.2523	CSR Strategy	-0.2730
Negativity	$0.4910*$	Emissions	-0.3958
Net sentiment	$0.8470**$	Environment Pillar	-0.3087
Weighted negative score	0.0317	Social Pillar	-0.2929
Weighted positive score	$0.5145*$	Governance Pillar	0.3008
Weighted net score	0.3325		

Table 7: Spearman's correlation coefficient for metrics and cluster membership, $*$p-value < 0.05, $**$p-value < 0.01

4.3 Discussion

Several findings of this exploratory study can enhance the research in the field of sentiment analysis:

- Sustainability sections tend to contain less sentiment words than sustainability reports but carry equal weighted sentiment scores.

Companies in the MDAX meet the criteria to be obliged to include non-financial reporting in their service to the investors. Stricter rules are in place since 2017. This has increased formal requirements in the reporting, both in terms of content and formalities. This leaves less room for companies to design the sustainability sections to their own pleasure. The shortness of the section in one overall report at the same time, forces companies to be more concise and use less commonly shared words and reach similar weighted sentiment scores as longer texts.

On the other hand, for companies that use both text forms, sustainability sections as a legal requirement and sustainability reports as voluntary disclosure, tend to be less precise in the latter. This is shown by the increase of the sentiment words but the homogenous weighted scores.

This could be seen as an indicator for green washing which refers to the policy of firms to exaggerate their own environmental actions or advertise neutral activities in a misleading manner. Concern for ESG-issues is then reduced to a marketing topic. Using overall positive language without adding value to the investor seems to be a hint for this behavior.

- CEO letters contain far more sentiment than other text type, no matter the position of the letter.

As other researchers have shown ([13]), CEO letters are a useful source for sentiment analysis. This is verified in my analysis. The extracted sentiment is higher as well as correlation to important ESG metrics. At the same time, this hand crafted piece of writing is not subject to any regulation. Companies, their CEOs and their marketing/ legal departments are completely free in their focus and word choice. This gives the opportunity for meaningful textual analysis.

Other text types, especially the sustainability sections of annual reports are far more regulated. Their value for textual analysis will increase with time as regulation gives also more comparability and makes time series analysis feasible.

- Different companies are associated with a different sentiment level.

Sentiment among the companies differs to a huge degree. Aggregated values disguise this. When trying to predict stock prices from sentiment scores, this has to be taken into account. Further research has to be done on the consistency of these differences.

- Positive sentiment is related to adjectives, negative to nouns.

In the sample analyzed, the most positive words for all text types were adjectives or participles used as attributes. This is an important finding for feature engineering in machine learning. The indication of adjectives would be important information for a learning algorithm. This supports the research done by Meyer et al. (2017) [27] who found certain word combination to be improving the classification accuracy.

- Risk perception can only to a small degree be derived from sustainability related texts via sentiment analysis.

Negative sentiment has been expressed by words that are related with risk: such as *Risik(o)* (risk) and *Gefahr* (thread). At the same time, correlation to risk-indicating scores like the ESG Controversies score is weak to mediocre as is the association with ESG scores. Governance pillar score for example has no association to most sentiment metrics - even though the association to one subcategory (CSR Strategy Score) is particular strong. More than risk perception, sentiment seems therefore to proxy reporting quality. Here, proportional metrics have shown to be more associated than weighted scores. CEO letters in annual report - so texts without clear sustainability focus - have the least association and therefore most likely the least predictive power for ESG related ratings.

Positive wordings are expressed by adjectives mostly, none of them being especially related to opportunities or business cases. Regarding the attitude towards CSR, no conclusion can therefore be drawn.

- Sustainability reports are most associated with high CSR scores if they are well-balanced.

Less is more when it comes to sentiment: CSR strategy is valuated most effective and informative if the emotional content of a text is outbalanced, that means net sentiment is close to 0. A high proportion of positive words are not effective to disguise a missing. or otherwise non-productive CSR strategy.

For future analysis, focus on the proportion of negative words (as done by Gandhi, Loughran and McDonald (2019) [12] or net sentiment is therefore a viable approach. This finding is also in line with findings by Hagenau et al. (2013) [43].

- No sector-sentiment association seems to exist.

In the experiments, not only different sort of texts have been analyzed. The companies in the sample are operating in different sectors. Neither for the sustainability texts nor the CEO letters could offer a clear picture of segmentation as to which sector is associated with which polarity of sentiment. This finding is backed also by the cluster analysis: The most correlated subset of data are clustered - but no clear lines as to what makes companies features belong to which cluster can be derived from the analysis. Only sentiment metric divide the sample.

This work has several limitations. Some of the analyzed texts, especially the letters of the CEOs, are quite short with only a few hundred words. Gandhi and Loughran (2019) [12] suggest to only include texts in the sample that have more than 2000 words. Given the characteristics of the text source, this critique is ignored as length is one key factor of these letters. Furthermore, the sample in use is quite small due to the scope of this thesis compared to several thousands tweets or pages of reports other research analyze.

Despite the regulation as to what kind of reporting is needed, companies are still free to decide on the scope and format of their CSR disclosure. This makes the subset of companies differing for each report type: More annual reports CEO letters are included in the sample for example than sustainability report CEO letters.

With regards to the second experiment, other ESG rating metrics are available on the market. No universal approach as to how to compute these ratings exist. They are subjectively calculated by the provider. Choosing another provider than Thomson Reuters could yield different results. The same is true for the choice of word lists - even though they might be less subjective. But still there are debatable word choices: Words that are perceived as positive but are neutral in the specific context might occur: I excluded *nachhaltig* (sustainable) and its related inflections. More adjustments might be necessary as they were done by Loughran and McDonald (2011) [36]. At the same time, this problem matters most for the proportional sentiment metrics. Taking weirdness into account in the given sample will equalize the impact of these words as it is reasonable to assume that all texts contain them. Nevertheless, problematic terms that are related to the business model of a company are candidates for exclusion.

Not handled is the problem of compounds. German language can build new words out of existing, rigid word lists will not reflect this. For example, before the adjustment, *nachhaltig* (sustainable) was part of the positive word list. A compound like *Nachhaltigkeitsbericht* (sustainability report) was not - even though one could argue that it should carry the same or at least some sentiment as one of its root words. In the scope of this work, adapting the lists to these words is not possible.

Last but not least: The used approach for negations is improved compared to others but still rule-based. With more resources, a more sophisticated approach to handle negation and additionally intensifications like *sehr* (very much) or *gar* (at all or even) could be used.

Further research on CSR reports as risk indicators can be done in the future: Regulation on CSR disclosure is still very recent. Time series analysis in the future can show the true predictive power of sentiment in CSR reports in a forward looking manner. Also taking the structure of the document more into account could yield valuable insights. For this, a tone dispersion analysis ([11]) or the retrieval of topics and the topic-related sentiment ([32]) will be useful. The research done by Song et al. (2018) [47] points in a similar direction.

5 Conclusions

With this thesis, I set out to achieve two major points: One was to give an overview of recent advancements in the field of sentiment analysis in finance since approximately 2014. The second was to conduct such a sentiment analysis in order to give insights on the suitability of the methodology to analyze risks and risk perception towards environmental, social and governance (ESG) issues.

In the literature review, results show that two major approaches to sentiment analysis prevail: Lexicon-based and machine learning. Improving both by fine-graining or taking semantic information into account is in the focus of research on methodology. Pure bag-of-words approaches tend to be less used. Common targets for the sentiment analysis are variations of predicting stock markets, either in terms of abnormal returns to create a profitable trading strategy, or in terms of predicting stock movements and volatility. Risk as a category has been an important research focus on micro and macro level as well as detection of unusual behavior in financial markets like herding or fraud. For this, new resources in languages additional to English have been used.

In the empirical part of the thesis a lexicon-based sentiment analysis has been done on ESG-related texts. CEO letters have been shown to be a major carrier of sentiment in comparison to more technical parts of reports. Association with core ESG rating scores are weak to mediocre in the sample. Sentiment expressed in sustainability reports might work as an indicator of reporting quality rather than risk.

References

[1] Andrei Shleifer. *Inefficient Markets: An Introduction to Behavioral Finance*. Oxford University Press, 2000.

[2] Eugene Fama. Efficient capital markets, a review of theory and empirical work. *Journal of Finance*, 25:383–417, 1970.

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

[4] Mika V. Mäntyläa, Daniel Graziotin, and Miikka Kuutila. The evolution of sentiment analysis - a review of research topics, venues, and top cited papers. *Computer Science Review*, 27:16–32, February 2018.

[5] Tim Loughran and Bill McDonald. The use of word lists in textual analysis. *Journal of Behavioral Finance*, 16(1):1–11, 2015.

[6] Ming-FengTsai and Chuan-Ju Wang. On the risk prediction and analysis of soft information in finance reports. *European Journal of Operational Research*, 257(1):243–250, February 2017.

[7] Chuan-Ju Wang, Ming-Feng Tsai, Tse Liu, and Chin-Ting Chang. Financial sentiment analysis for risk prediction. In *International Joint Conference on Natural Language Processing*, pages 802–808, Nagoya, Japan, October 2013.

[8] Lynnette Purda and David Skillicorn. Accounting variables, deception, and a bag of words: Assessing the tools of fraud detection. *Contemporary Accounting Research*, 32(3):1193–1223, 2015.

[9] Kelvin K. F. Law and Lillian F. Mills. Taxes and financial constraints: Evidence from linguistic cues. *Journal of Accounting Research*, 53(4):777–819, September 2015.

[10] Andriy Bodnaruk, Tim Loughran, and Bill McDonald. Using 10-k text to gauge financial constraints. *Journal of Financial and Quantitative Analysis*, 50(4):623–646, August 2015.

[11] Kristian D. Allee and Matthew D. DeAngelis. The structure of voluntary disclosure narratives: Evidence from tone dispersion. *Journal of Accounting*, 53(2), May 2015.

[12] Priyank Gandhi, Tim Loughran, and Bill McDonald. Using annual report sentiment as a proxy for financial distress in u.s. banks. *Journal of Behavioral Finance*, 0(0):1–13, 2019.

[13] Clemens Nopp and Allan Hanbury. Detecting risks in the banking system by sentiment analysis. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, September 2015.

[14] Nicolas Pröllochs, Stefan Feuerriegel, and Dirk Neumann. Negation scope detection in sentiment analysis: Decision support for news-driven trading. *Decision Support Systems*, 88:67–75, August 2016.

[15] Joscha Märkle-Huß, Stefan Feuerriegel, and Helmut Prendinger. Improving sentiment analysis with document-level semantic relationships from rhetoric discourse structures. In *Proceedings of the 50th Hawaii International Conference on System Sciences*, pages 1142–1151, 2017.

[16] Katrin Hummel, Stéphanie Mittelbach-Hörmanseder, Charles H. Cho, and Dirk Matten. Implicit versus explicit corporate social responsibility disclosure: A textual analysis. *Available at SSRN: https://ssrn.com/abstract=3090976 or http://dx.doi.org/10.2139/ssrn.3090976*, December 2017.

[17] Tianyou Hu and Arvind Tripathi. The performance evaluation of textual analysis tools in financial markets. In *Workshop of Information Technology and Systems*, Dallas, 2015.

[18] Rui Ren and Desheng Wu. An innovative sentiment analysis to measure herd behavior. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, pages 1–11, 2018.

[19] Mattia Atzeni, Amna Dridi, and Diego Reforgiato Recupero. Fine-grained sentiment analysis on financial microblogs and news headlines. In *Semantic Web Evaluation Challenge*, pages 124–128, 2017.

[20] Thien Hai Nguyen, Kiyoaki Shirai, and Julien Velcin. Sentiment analysis on social media for stock movement prediction. *Expert Systems with Applications*, 42(24):9603–9611, December 2015.

[21] Mengxiao Jiang, Man Lan, and Yuanbin Wu. Ecnu at semeval-2017 task 5: An ensemble of regression algorithms with effective features for fine-grained sentiment analysis in financial domain. In *Proceedings of the 11th International Workshop on Semantic Evaluations (SemEval-2017)*, pages 888–893, Vancouver, Canada, August 2017. Association for Computational Linguistics.

[22] Nuno Oliveira, Paulo Cortez, and Nelson Areal. Stock market sentiment lexicon acquisition using microblogging data and statistical measures. *Decision Support Systems*, 85:62–73, May 2016.

[23] Francesco Corea. Can twitter proxy the investors' sentiment? the case for the technology sector. *Big Data Research*, 4:70–74, 2016.

[24] Alpaslan Burak Eliaçik and Nadia Erdoğan. Influential user weighted sentiment analysis on topic based microblogging community. *Expert Systems with Applications*, 92:403–418, 2018.

[25] Desheng Dash Wu and David L. Olson. *Financial Risk Forecast Using Machine Learning and Sentiment Analysis*, pages 32–48. Palgrave Macmillan UK, London, 2015.

[26] Andrew Moore, Paul Edward Rayson, and Steven Eric Young. Domain adaptation using stock market prices to refine sentiment dictionaries. In European Language Resources Association (ELRA), editor, *Proceedings of the 10th edition of Language Resources and Evaluation Conference (LREC2016)*, 2016.

[27] Bradley Meyer, Marwan Bikdash, and Xiangfeng Dai. Fine-grained financial news sentiment analysis. In *SoutheastCon 2017*, 2017.

[28] Antonio Moreno-Ortiza and Javier Fernández-Cruz. Identifying polarity in financial texts for sentiment analysis: a corpus-based approach. In *7th International Conference on Corpus Linguistics: Current Work in Corpus Linguistics: Working with Traditionally-conceived Corpora and Beyond (CILC 2015)*, 2015.

[29] Vytautas Karalevicius, Niels Degrande, and Jochen De Weerdt. Using sentiment analysis to predict interday bitcoin price movements. *The Journal of Risk Finance*, 19(1):56–75, 2018.

[30] Min-Yuh Day and Chia-Chou Lee. Deep learning for financial sentiment analysis on finance news providers. In *016IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 2016.

[31] Zhiyang Teng and Duy-Tin Vo andYue Zhang. Context-sensitive lexicon features for neural sentiment analysis. In Association for Computational Linguistics, editor, *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1629–1638, Austin, Texas, November 2016.

[32] Qingqing Zhou, Rui Xia, and Chengzhi Zhang. Online shopping behavior study based on multi-granularity opinion mining: China versus america. *Cognitive Computation*, 8(4):587–602, August 2016.

[33] Pantelis Agathangelou, Ioannis Katakis, Fotios Kokkoras, and Konstantinos Ntonas. Mining domain-specific dictionaries of opinion words. In *WISE 2014: Web Information Systems Engineering*, pages 47–62, 2014.

[34] German G. Creamer, Yong Ren, Yasuaki Sakamoto, and Jeffrey V. Nickerson. A textual analysis algorithm for the equity market: the european case. *The Journal of Investing*, 25(3):105–116, 2016.

[35] Mohammed N. Al-Kabi, Amal H. Gigieh, Izzat M. Alsmadi, and Heider A. Wahsheh. Opinion mining and analysis for arabic language. *(IJACSA) International Journal of Advanced Computer Science and Applications*, 5(5):181–195, 2014.

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

[37] Elaine Henry and Andrew J. Leone. Measuring qualitative information in capital markets research: Comparison of alternative methodologies to measure disclosure tone. *The Accounting Review*, 91(1):153–178, January 2016.

[38] Pamela R. Murphy, Lynnette Purda, and David Skillicorn. Can fraudulent cues be transmitted by innocent participants? *Journal of Behavioral Finance*, 19(1):1–15, 2018.

[39] Elaine Henry and Andrew J. Leone. Measuring qualitative information in capital markets research. *Available at SSRN: https://ssrn.com/abstract=1470807 or http://dx.doi.org/10.2139/ssrn.1470807*, 2009.

[40] Ian H. Witten, Eibe Frank, and Mark A. Hall. *Data Mining - Practical Machine Learning Tools and Techniques*. Elsevier, 3 edition, 2011.

[41] Keith Cortis, André Freitas, Tobias Daudert, Manuela Hürlimann, Manel Zarrouk, Siegfried Handschuh, and Brian Davis. Semeval-2017 task 5: Fine-grained sentiment analysis on financial microblogs and news. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 519–535, Vancouver, Canada, August 2017. Association for Computational Linguistics.

[42] Alpaslan Burak Eliaçik and Nadia Erdoğan. User-weighted sentiment analysis for financial community on twitter. In *2015 11th International Conference on Innovations in Information Technology (IIT)*, November 2015.

[43] Michael Hagenau, Michael Liebmann, and Dirk Neumann. Automated news reading: Stock price prediction based on financial news using context-capturing features. *Decision Support Systems*, 55:685–697, 2013.

[44] Olga Kolchyna, Tharsis T. P. Souza, Philip Treleaven, and Tomaso Aste. *Handbook of Sentiment Analysis in Finance*, chapter Twitter Sentiment Analysis: Lexicon Method, Machine Learning Method and Their Combination. Albury Books, 2016.

[45] Jozef Barunik, Cathy Yi-Hsuan Chen, and Jan Vecer. Sentiment-driven stochastic volatility model: A high-frequency textual tool for economists. *Available at SSRN: https://ssrn.com/abstract=3397314 or http://dx.doi.org/10.2139/ssrn.3397314*, 2019.

[46] Sven F. Crone and Christian Koeppel. Predicting exchange rates with sentiment indicators: An empirical evaluation using text mining and multilayer perceptrons. In *IEEE/IAFE Conference on Computational Intelligence for Financial Engineering, Proceedings (CIFEr)*, pages 114–121, October 2014.

[47] Yuan Song, Hongwei Wang, and Maoran Zhu. Sustainable strategy for corporate governance based on the sentiment analysis of financial reports with csr. *Financial Innovation*, 4(2), 2018.

[48] Jundong Li, Kewei Cheng, Suhang Wang, Fred Morstatter, Robert P. Trevino, Jiliang Tang, and Huan Liu. Feature selection: A data perspective. *ACM Computing Surveys (CSUR)*, 50(6), January 2018.

[49] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning - Data Mining, Inference, and Prediction.* Springer Series in Statistics. Springer, 2 edition, 2009.

[50] Mohamed Dermouche, Leila Kouas, Julien Velcin, and Sabine Loudcher. A joint model for topic-sentiment modelingfrom text. In *Proceedings of the 30th Annual ACM Symposium on Applied Computing - SAC '15*, 2015.

[51] Zubair Shah, Paige Martin, Enrico Coiera, Kenneth D Mandl, and Adam G Dunn. Modeling spatiotemporal factors associated with sentiment on twitter: Synthesis and suggestions for improving the identification of localized deviations. *Journal of Medical Internet Research*, 21(5), 2019.

[52] B. Gnana Priya. Emoji based sentiment analysis using knn. *International Journal of Scientific Research and Review*, 7(4):859–865, April 2019.

[53] Robert Remus, Khurshid Ahmad, and Gerhard Heyer. Sentiment in german-language news and blogs, and the dax. In *Text Mining Services 2009*, pages 149–158, 2009.

[54] Elisabeth Albertini. A descriptive analysis of environmental disclosure: A longitudinal study of french companies. *Journal of Business Ethics*, 121(2):233–254, May 2014.

[55] Deutsche Börse AG. *The DAX index universe*, October 2018.

[56] Robert Remus, Uwe Quasthoff, and Gerhard Heyer. Sentiws - a publicly available german-language resource for sentiment analysis. In *Proceedings of the 7th International Language Ressources and Evaluation*, pages 1168–1171, 2010.

[57] Marc Schulder, Michael Wiegand, and Josef Ruppenhofer. Automatically creating a lexicon of verbal polarity shifters: Mono- and cross-lingual methods for german. In *Proceedings of the 27th International Conference on Computational Linguistics*, volume 1, pages 2516–2528, Santa Fe, New Mexico, USA, August 2018. International Committee on Computational Linguistics.

[58] Fridolin Wild. *lsa: Latent Semantic Analysis*, 2015.

[59] Thomson Reuters Eikon. *ESG Scores (ESG Score, ESG Controversies Score, ESG Combined Score, CSR Strategy Score, Emissions Score, Environment Pillar Score, Social Pillar Score, Governance Pillar Score)*, July 2019.

[60] Thomson Reuters Eikon. *Thomson Reuters ESG Scores*. Thomson Reuters, February 2019.

[61] tabulizer: Bindings for Tabula PDF Table Extractor Library. *Thomas J. Leeper*, r package version 0.2.2 edition, 2018.

[62] Julia Silge and David Robinson. tidytext: Text mining and analysis using tidy data principles in r. *JOSS*, 1(3), 2016.

[63] Julia Silge and David Robinson. *Text Mining with R: A tidy approach*. O'Reilly Media, 1st edition, June 2017.

[64] Ingo Feinerer, Kurt Hornik, and David Meyer. Text mining infrastructure in r. *Journal of Statistical Software*, 25(5):1–54, March 2008.

[65] Martin F. Porter. An algorithm for suffix stripping. *Program: electronic library and information systems*, 14(3):130–137, 1980.

[66] Jiawei Han, Micheline Kamber, and Jian Pei. *Data Mining - Concepts and Techniques*. Elsevier, 3 edition, 2012.

Appendices

A Tables

Company	AR-Sust	AR-CEO	SR-Total	SR-CEO
1&1 Drillisch AG	No	Yes	Yes	Yes
Aareal Bank AG	No	No	Yes	No
Alstria Office	No	No	Yes	Yes
Aurubis AG	Yes	Yes	Yes	Yes
Axel Springer SE	No	Yes	No	No
Bechtle AG	No	Yes	Yes	Yes
Brenntag AG	Yes	Yes	Yes	Yes
Commerzbank AG	Yes	Yes	Yes	No
CTS Eventim AG & Co KgaA	No	Yes	Yes	No
Delivery Hero SE	Yes	Yes	No	No
Deutsche EuroShop AG	Yes	Yes	No	No
Deutsche Pfandbriefbank AG	Yes	Yes	Yes	No
Deutsche Wohnen SE	Yes	Yes	No	No
Dürr AG	Yes	Yes	Yes	No
Evonik Industries AG	Yes	Yes	Yes	No
Evotec SE	No	Yes	Yes	No
Fielmann AG	No	Yes	Yes	Yes
Fraport AG (shortened)	Yes	Yes	Yes	Yes
freenet AG	Yes	Yes	No	No
Fuchs Petrolub SE	Yes	Yes	Yes	Yes
GEA Group AG	Yes	Yes	Yes	No
Gerresheimer AG	Yes	Yes	No	No
Hannover Rück SE	No	Yes	No	No

HELLA GmbH & Co KgaA	Yes	Yes	No	No
Hochtief AG	Yes	Yes	No	No
Hugo Boss AG	Yes	Yes	Yes	Yes
Innogy SE	No	Yes	Yes	Yes
K+S AG	Yes	Yes	No	No
Kion Group AG	No	Yes	Yes	Yes
Lanxess AG	Yes	Yes	Yes	No
LEG Immobilien AG	No	Yes	Yes	Yes
Metro AG	No	Yes	Yes	Yes
MorphoSys AG	Yes	Yes	No	No
MTU Aero Engines	Yes	Yes	No	No
Nemetschek SE	Yes	Yes	Yes	No
Norma	No	Yes	Yes	Yes
Osram Licht AG	Yes	No	Yes	Yes
ProSiebenSat.1 Media SE	Yes	Yes	No	No
Puma SE	Yes	Yes	No	No
Rheinmetall AG	Yes	Yes	No	No
Rocket Internet SE	No	Yes	Yes	No
Salzgitter AG	No	Yes	Yes	No
Sartorius AG	Yes	Yes	No	No
Schaeffler AG	No	Yes	Yes	No
Scout24 AG	No	Yes	Yes	Yes
Siemens Healthineers AG	No	Yes	No	No
Siltronic AG	Yes	Yes	Yes	Yes

Software AG	Yes	Yes	Yes	No
Symrise AG	Yes	Yes	No	No
TAG Immobilien AG	No	Yes	Yes	Yes
Telefonica Deutschland Holding AG	No	Yes	Yes	Yes
Uniper SE	Yes	Yes	Yes	Yes
United Internet AG	No	Yes	Yes	Yes
Wacker Chemie AG	Yes	Yes	No	No
Zalando SE	Yes	Yes	No	No

Table A1: Companies and reports included in the analysis. A company has to contribute at least one report to appear in the list.

Company	NAICS Sector
1&1 Drillisch AG	Information
Aareal Bank AG	Finance and Insurance
Aurubis AG	Manufacturing
Axel Springer SE	Information
Bechtle AG	Professional, Scientific, and Technical Services
Brenntag AG	Wholesale Trade
Commerzbank AG	Finance and Insurance
CTS Eventim AG & Co KgaA	Arts, Entertainment, and Recreation
Delivery Hero SE	Information
Deutsche EuroShop AG	Real Estate and Rental and Leasing
Deutsche Wohnen SE	Real Estate and Rental and Leasing
Dürr AG	Manufacturing
Evonik Industries AG	Manufacturing
Evotec SE	Professional, Scientific, and Technical Services
Fielmann AG	Retail Trade
Fraport AG Frankfurt Airport Services Worldwide	Transportation and Warehousing
freenet AG	Information
Fuchs Petrolub SE	Manufacturing
GEA Group AG	Manufacturing
Gerresheimer AG	Manufacturing
Hannover Rück SE	Finance and Insurance
HELLA GmbH & Co KgaA	Manufacturing
Hochtief AG	Construction
Hugo Boss AG	Manufacturing
Innogy SE	Utilities
K+S AG	Mining, Quarrying, and Oil and Gas Extraction

Kion Group AG	Manufacturing
Lanxess AG	Manufacturing
LEG Immobilien AG	Real Estate and Rental and Leasing
Metro AG	Retail Trade
MorphoSys AG	Professional, Scientific, and Technical Services
Nemetschek SE	Information
Osram Licht AG	Manufacturing
ProSiebenSat.1 Media SE	Information
Puma SE	Manufacturing
Rheinmetall AG	Manufacturing
Rocket Internet SE	Retail Trade
Salzgitter AG	Manufacturing
Sartorius AG	Manufacturing
Schaeffler AG	Manufacturing
Scout24 AG	Information
Siemens Healthineers AG	Manufacturing
Siltronic AG	Manufacturing
Software AG	Information
Symrise AG	Manufacturing
TAG Immobilien AG	Real Estate and Rental and Leasing
Telefonica Deutschland Holding AG	Information
Uniper SE	Utilities
United Internet AG	Information
Wacker Chemie AG	Manufacturing
Zalando SE	Retail Trade

Table A2: Companies in the sample and NAICS sector

Metric	ESG	CSR Strategy	Emissions	Environmental Pillar	Social Pillar
Total words	0.3224	0.3016	/	0.3330	/
Sum positive words	0.3571	0.2529	0.2587	0.3713∗	/
Positivity	−0.3215	−0.3302	/	−0.3143	/
Sum negative words	/	0.2958	/	/	/
Negativity	−0.4162∗	−0.2685	/	−0.4180∗	/
Net sentiment	−0.2868	−0.3416	/	−0.2828	/
Weighted negative	/	−0.2629	/	/	/
Weighted positive	/	/	0.2672	0.3161	/
Weighted Net	/	/	/	/	0.2716

Table A3: Correlation coefficients ρ for the Annual report's sustainability section (only values above 0.25 are given), $*$p-value < 0.05, $**$p-value < 0.01
ESG Combined Score, ESG Controversies Score and Governance Pillar Score omitted as no values above threshold.

Metric	ESG	ESG Controversies	Strategy	Environmental Pillar	Social Pillar
Total words	$-0.3183*$	/	$-0.3096*$	$-0.3666**$	/
Sum positive words	/	/	/	$-0.2998*$	/
Positivity	$0.3519*$	/	$0.3512*$	$0.3710**$	$0.3102*$
Sum negative words	/	/	/	/	/
Negativity	$0.3140*$	-0.2726	$0.3961**$	$0.3034*$	0.2703
Net sentiment	$0.3178*$	/	$0.2857*$	$0.3370*$	0.2816
Weighted negative	/	/	/	/	/
Weighted positive	/	/	/	/	/
Weighted Net	/	/	/	/	/

Table A4: Correlation coefficients ρ for the Annual report CEO letter (only values above 0.25 are given),
$*$p-value < 0.05, $**$p-value < 0.01
ESG Combined Score, Emissions Score and Governance Pillar Score omitted as no values above threshold.

Metric	ESG	ESG Comb.	Contro.	Strategy	Emissions	E. Pillar	S. Pillar
Total words	/	/	/	0.4166	0.3563	0.3484	/
Sum positive words	/	/	/	0.3725	0.2779	0.2559	/
Positivity	/	0.2912	/	−0.4342	/	−0.3702	−0.3351
Sum negative words	/	/	−0.3374	0.3877	0.3844	0.3409	/
Negativity	/	/	−0.2600	/	/	/	/
Net sentiment	/	0.3035	0.3096	−0.4662∗	−0.3719	−0.3579	−0.3070
Weighted negative	−0.4783∗	/	0.3498	−0.4308	−0.4458	−0.4326	−0.3616
Weighted positive	/	/	/	/	−0.2954	/	/
Weighted Net	/	/	/	−0.4004	−0.2526	/	/

Table A5: Correlation coefficients ρ for the Sustainability report's CEO letter (only values above 0.25 are given), $*$p-value < 0.05, $**$p-value < 0.01
ESG Comb. = ESG Combined Score, Contro. = ESG Controversies Score, E. Pillar = Environmental Pillar, S. Pillar = Social Pillar, Governance Pillar Score omitted as no values above threshold.

Metric	ESG	CSR Strategy	Emissions	Environmental Pillar	Social Pillar	G. Pillar
Total words	0.2628	0.6125∗	0.4157∗∗	0.3783∗	0.2760	/
Sum pos words	/	0.4461	0.2940	0.2742∗	/	/
Positivity	−0.2537	−0.6133	−0.3853∗	−0.3739∗∗	−0.2933∗	/
Sum neg words	/	0.5095	/	/	/	/
Negativity	−0.2570	−0.4414∗	−0.5051∗∗	−0.4967∗∗	/	/
Net sentiment	−0.2537	−0.6465∗∗	−0.3420	−0.2775	−0.2951	/
Weighted negative	/	/	/	/	/	−0.2702
Weighted positive	/	/	/	−0.3207	/	/
Weighted Net	−0.2705	/	/	/	/	−0.3251

Table A6: Correlation coefficients ρ for the Sustainability report in total (only values above 0.25 are given),
∗p-value $<$ 0.05, ∗∗p-value $<$ 0.01
G. Pillar = Governance Pillar Score, ESG Combined Score and ESG Controversies Score omitted as no values above threshold.

B Figures

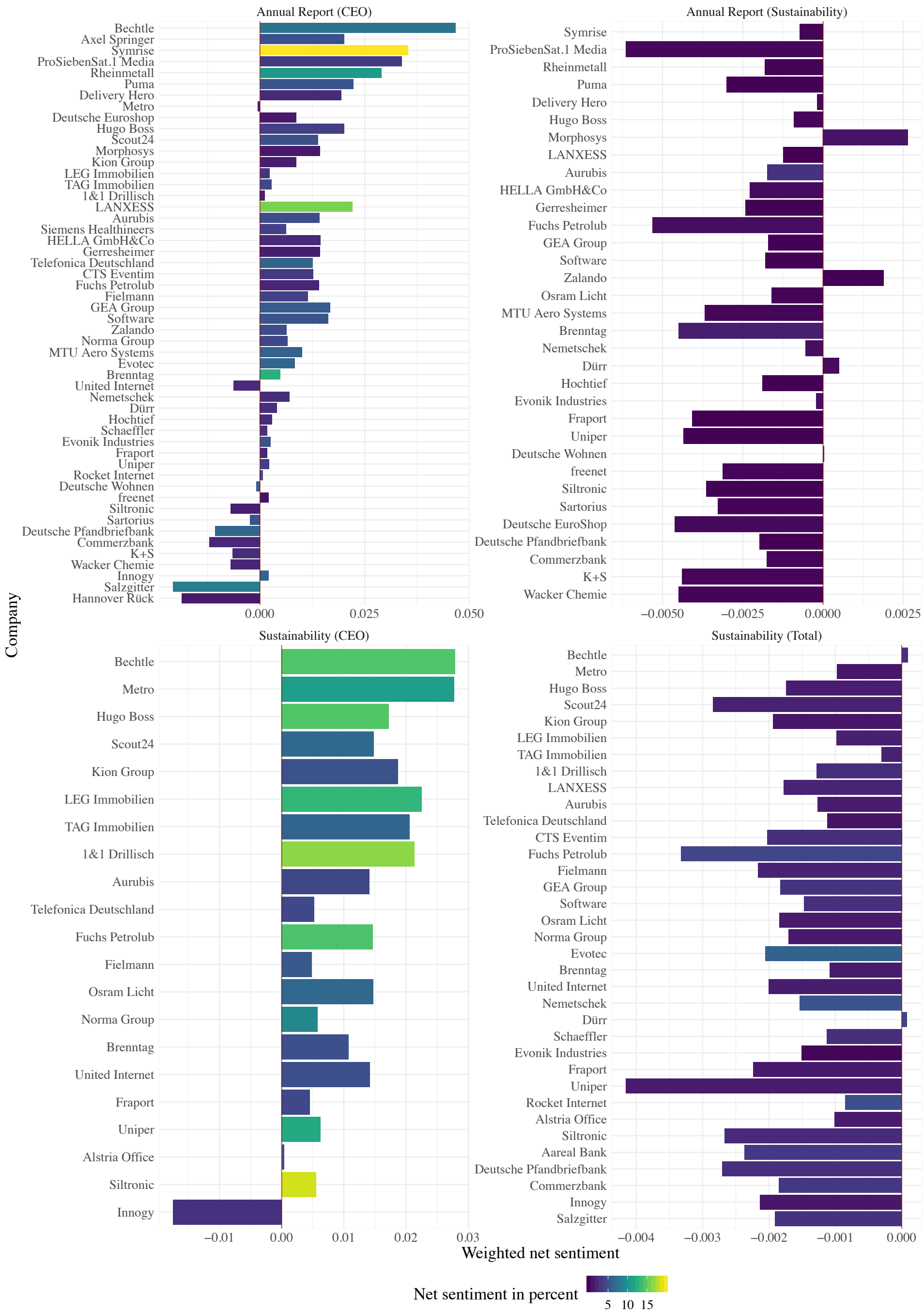


Figure B1: Net Sum Weight per report type and company, colored by the net sentiment.

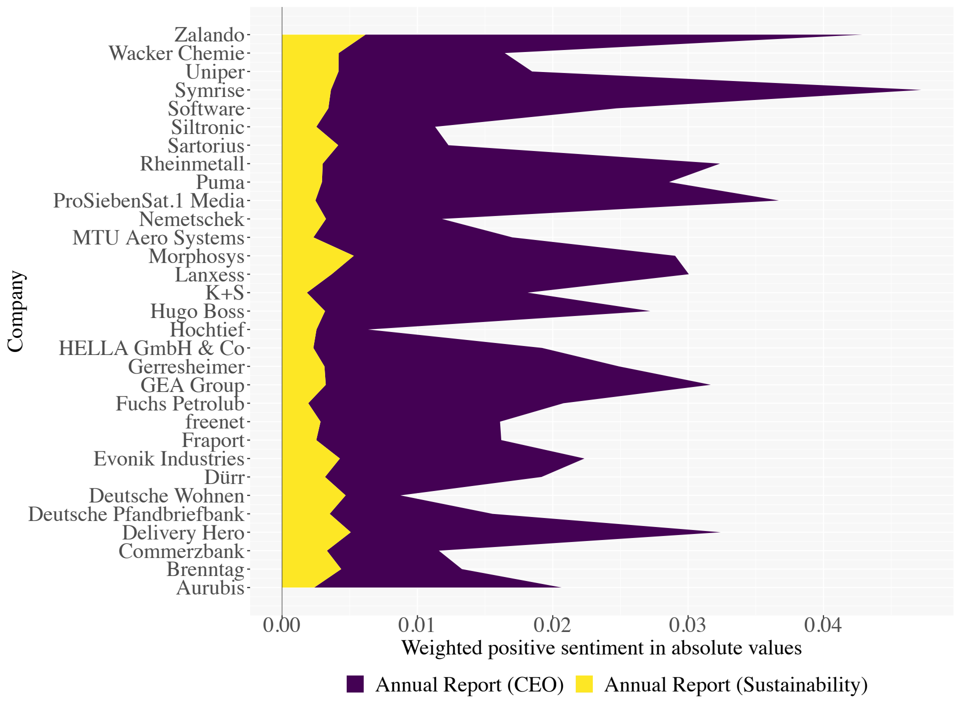


Figure B2: The Weighted Positive Sentiment of two types of texts given in the annual report.

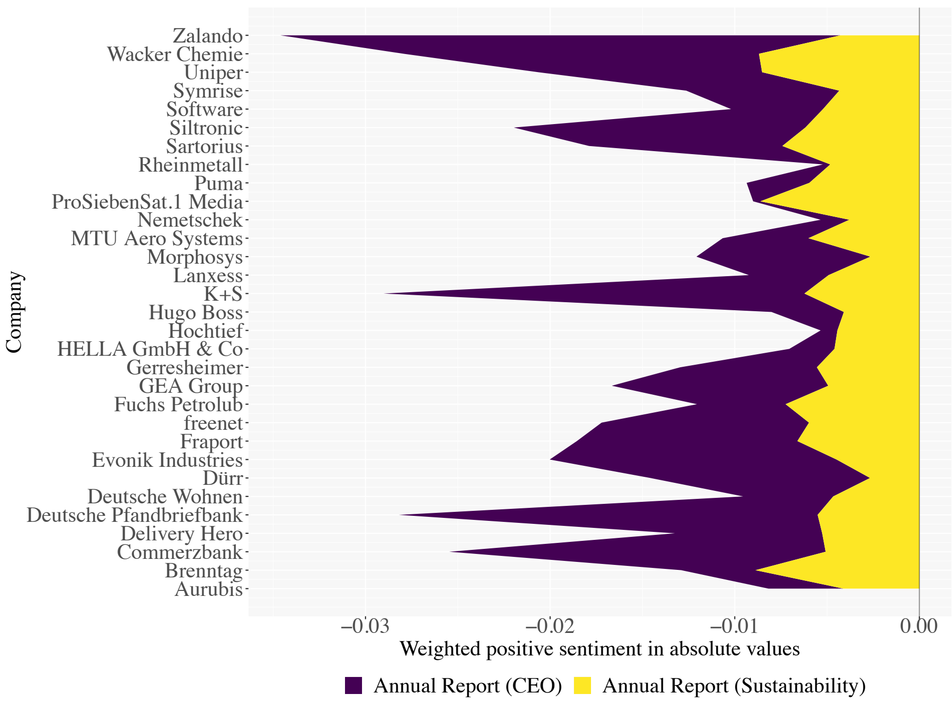


Figure B3: The Weighted Negative Sentiment of two types of texts given in the annual report.

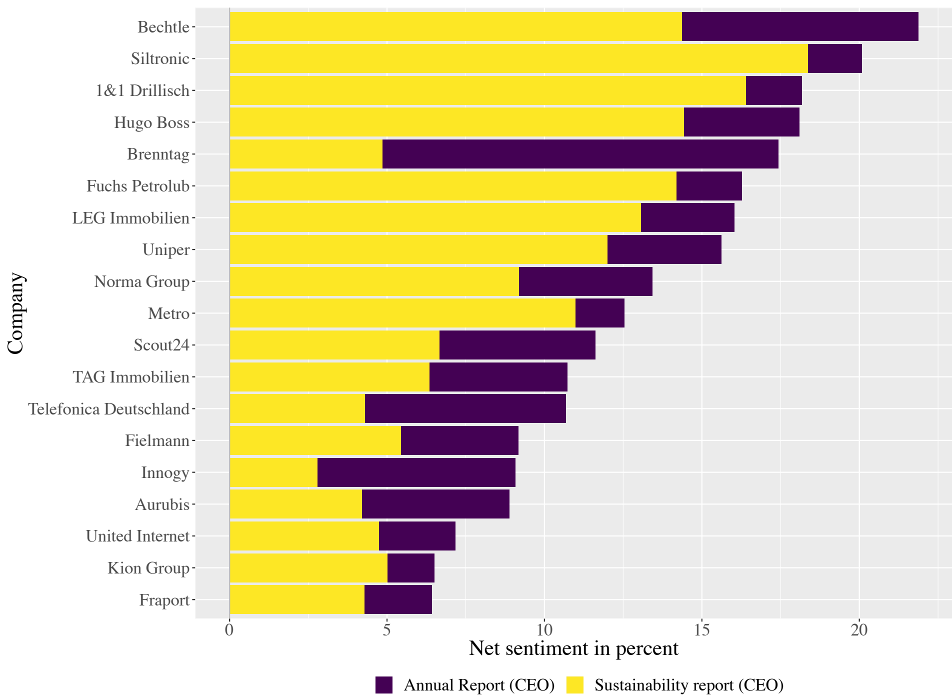


Figure B4: Net sentiment for CEO letters according to their position: in the Annual or the Sustainability Report.