Detecting Greenwashing Signals Through Computational Language Technologies

Arian Contessotto[1], Tim Giger[1] and Levin Reichmuth[1]

[1]*Lucerne University of Applied Sciences and Arts, Zentralstrasse 9, Lucerne, 6002, Switzerland*

1. Introduction

This project report describes the findings from a practical multi-step computational language technologies task including methodological approaches.

1.1. Background

Corporate greenwashing has emerged as a major issue of serious concern over the past two years. This phenomenon involves companies highlighting the positive elements of their sustainability initiatives while minimizing the negative impacts of their operations. Although public companies are required to disclose their sustainability efforts, it is postulated that environmental, social, and governance (ESG) performance may be inflated in self-reported data, such as ESG reports. This situation potentially misleads sustainable investors due to information asymmetry. To mitigate this problem, incorporating external media data could provide a more balanced assessment of a company's sustainable practices. By examining different media sources, a holistic understanding of a company's sustainability could be achieved. [1]

1.2. Objectives

This project uses state-of-the-art natural language processing (NLP) techniques, such as large language models (LLMs), and text embedding, to elucidate salient aspects of greenwashing detection. The project has three main objectives: First, to deepen the understanding of greenwashing through comprehensive text analysis; second, to investigate potential insights derived from sentiment analysis regarding greenwashing risks; and third, to analyze the differential susceptibility of different ESG issues to greenwashing. By accomplishing these objectives, the project aims to improve the understanding of greenwashing practices and foster the formulation of more robust methodologies to detect and counter misleading representations of sustainability efforts in corporate disclosures. [1]

1.3. Dataset

The dataset, originally sourced from Kaggle, consists of 11,188 ESG documents on DAX companies, structured into eleven columns. It combines both company reports and third-party data, which can be differentiated by the field "internal". Further details are available on the Kaggle platform or in the task description. [1][2]

1.4. Proceeding

The research is divided into four distinct stages. Stage one includes data and text preprocessing and exploratory data analysis. Stage two involves sentiment annotation of the data for feature generation. In stage three, a sentiment classifier is trained and applied to the data, leading to an analysis that detects the presence of greenwashing in internal documents. The final stage four highlights the importance of SDG themes across different companies and industries and explores potential differences between internal and external documents. This four-stage methodology, supported by the task provider, SwissText, is described in Figure 1. [1]

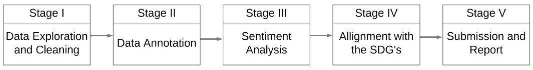

Figure 1: Proceeding

The implemented algorithms and methodologies for this research project have been developed and stored in a GitHub repository. This repository, accessible via https://github.com/syX113/hslu-nlp, contains all the code used in the various stages of this research, from data pre-processing and sentiment annotation to classifier training and subsequent greenwashing analysis. The repository thus serves as a comprehensive archive of the computational techniques used in this research to investigate corporate greenwashing practices and their impact on sustainability disclosures. It provides a resource for future related studies, facilitating both replication and further development of the methods employed.

[†] These authors contributed equally.

arian.contessotto@stud.hslu.ch (A. Contessotto); tim.giger@stud.hslu.ch (T. Giger); levin.reichmuth@hslu.ch (L. Reichmuth)

2. Stage I

This chapter describes the findings and procedure in stage one. Since stage one was developed individually, the different approaches as well as similarities and deviations are explained punctually.

2.1. General Data Preprocessing

The basic procedure for general data preprocessing was similar for all three authors. The data set was checked for missing values, duplicates and incorrect data types. One row had no entry in the ”content” column and was therefore removed since it provides no meaningful content. There were 6 duplicates in the data set, which were removed. In contrast to Tim and Levin, Arian additionally removed all rows that had duplicates in the ”content” column. The reason for this was that these documents summarized information about multiple companies. Assuming that these entries could not be clearly assigned to one company but still affect the analysis at the company level.

All three authors detected erroneous data types in the ”date” column. While Levin and Arian removed the rows with erroneous date data types, Tim kept them in the data set to avoid losing information and gave them a default date.

Arian and Levin also examined the number of documents (internal and external) per company during the general data preprocessing and decided to remove 8 documents concerning the two companies Fresenius and Hannover Re, since only a very small number of documents or no external documents were present for these two companies. Arian and Levin were of the opinion that no representative analysis could be made for these two companies.

Tim and Arian also used a language detector to detect and remove non-English documents, since the analysis is based on English. In total, this affected 106 documents. Both further performed a data enrichment. While Arian collected the data manually and added the column ”industry” to the data set with an additional file import and a subsequent merge, Tim chose a more sophisticated approach and built a scraper into his code to automatically download additional company information such as sector, industry, market capitalization, etc. and add it to the data set. This data enrichment then enabled the two of them to perform analyses at the industry level, for example, which led to interesting results and represented added value.

2.2. Text Preprocessing

Text preprocessing was considered one of the most important and difficult steps in this task. The text preprocessing of the three authors shows similarities but also significant differences. For these reasons, and to account for individual considerations and decisions, each text preprocessing is described in detail in the following subsections. Note that the below descriptions do not evaluate or judge the different preprocessing approaches. It is simply an objective explanation of what has been implemented by each author.

2.2.1. Text Preprocessing of Arian

Arian considered the various preprocessing steps. In the following list, his considerations, thoughts and decisions are described:

- **Lowercase:** Lowercase is applied because it reduces vocabulary, normalizes text and contributes to better accuracy in text classification.
- **Expand Contractions:** Expanding contractions in natural language processing can improve consistency, remove ambiguity and increase vocabulary and is therefore applied.
- **URL and E-Mail Removal:** URLs and E-Mail addresses are undesirable objects for the analysis and thus removed.
- **Punctuation Removal:** Removal of punctuation is only applied in the context of tokenizing words. Because punctuation can have an important meaning and structures a text. Furthermore, it is necessary for tokenizing sentences.
- **Numbers Removal:** Removal of numbers is only done in the context of tokenizing words. Numbers may have important meaning in a sentence and should remain in the processed content.
- **HTML Tag Removal:** Not required, as an analysis showed that the data does not contain HTML tags.
- **Line Break, Spaces, Tabs Removal:** Not required, as an analysis showed that no such elements are present in the data.
- **Emojis Removal:** Removal of emojis is only done as part of tokenizing words because emojis can have an important meaning in a sentence and are therefore retained in the processed content.
- **Accented Character Substitution:** Accented character substitution is applied to reduce noise in the text, prevent sparseness in the data and ensure consistency in the text presentation.
- **Spelling Correction:** Spelling correction is not applied because conducted tests did not lead to better data quality.

- **Word Tokenization:** Tokenization of words is an important preprocessing step in NLP and therefore applied.
- **Sentence Tokenization:** Sentence tokenization is applied in order to perform sentiment analysis on the basis of sentences.
- **Words Removal:** Word removal is only applied to tokenized words in order to preserve sentences and the processed content as a whole. The word removal included stop words, one-letter words, elements from company names and selected non-meaningful frequent words.
- **Lemmatization and Stemming:** Both lemmatization and stemming are applied to the word tokens.

2.2.2. Text Preprocessing of Levin

Levin decided to apply the following text preprocessing:

- **Lowercase:** Lowercasing because it reduces vocabulary, normalizes text and contributes to better accuracy in text classification.
- **Tag Removal:** Tags are unwanted elements in this task and are thus removed from the content.
- **Punctuation Removal:** Punctuation can disturb text analysis and is therefore removed.
- **Multiple Whitespace Removal:** Multiple whitespaces are unwanted elements in a text and thus caught and removed.
- **Stopwords Removal:** Stopwords are frequent words without significant meaning and are removed for better quality of the analysis.
- **Short-Words Removal:** Words shorter than three characters do presumably not held significant meaning and are dropped.
- **Word Tokenization:** As an important NLP preprocessing element, word tokenization is applied.

2.2.3. Text Preprocessing of Tim

Tim decided to apply the following text preprocessing:

- **String Conversion:** Converting the input to a string format ensures consistency and compatibility during subsequent processing tasks.
- **Lowercase:** Transforming all text to lowercase serves as a simple normalization step, reducing the complexity and variability of the input data.
- **Unicode Decoding:** Removing diacritics and normalizing the text encoding mitigates potential discrepancies arising from different encoding formats.
- **URL and E-Mail Removal:** Eliminating URLs and E-Mail addresses reduces noise in the data set, as these elements do not contribute valuable information for the analysis.
- **Extra Whitespace Removal:** Eradicating extra whitespaces improves text analysis and tokenization by ensuring that only meaningful spaces are retained.
- **Contact Detail Removal:** Excluding phone numbers, contact person strings, and social media references further minimizes noise in the data set, honing the focus on relevant text.
- **Table of Contents Removal:** Discarding the table of contents enhances the data quality by eliminating repetitive and non-essential information.
- **Named Entity Removal:** Employing the spaCy model to remove human names and other named entities optimizes the text for analysis and modeling by concentrating on pertinent content.
- **Abbreviation Expansion:** Common and uncommon abbreviations are expanded to improve text interpretation.
- **Special Character Elimination:** Excluding all special characters, except punctuation, refines the input data. Retaining punctuation is necessary for accurate sentence tokenization and removed after sentence tokenization.
- **Tokenization and Lemmatization:** Tokenizing words and sentences, and subsequently lemmatizing words, streamlines the text and reduces morphological variations.
- **Stopwords Removal:** Customizing the nltk stopwords list by adding or removing specific stopwords enables more precise and tailored text analysis.
- **Part-of-speech Tagging:** Assigning POS tags to words and sentences enhances the text representation by providing additional linguistic information, which may be beneficial for subsequent analysis and modeling tasks.
- **Spelling Correction:** Spellchecking was tested with different libraries but delivered not useful results.

2.3. Exploratory Data Analysis

Exploratory data analysis is divided into general data analysis and specific text analysis. All three authors have calculated the length of the reports in connection with the exploratory data analysis. Tim and Arian additionally calculated the polarity and generated a polarity value for each document using nltk sentiment intensity analyzer and vader lexicon. Both got more or less identical results.

This enrichment allows for more in-depth data analysis in stage one. In this chapter, only the most relevant insights are presented. More detailed information can be found in the notebooks handed in.

2.3.1. General Exploratory Data Analysis

The data set consists of internal and external documents. An internal document is a report written and published by the company. External documents are reports produced and published by external media. The ratio of external to internal documents in the data set is strongly imbalanced. More than 99 percent of all documents are external. This is an important fact that must be taken into account in the subsequent sentiment analysis. On the other hand, the few internal documents are significantly longer than the external documents. On average, an internal document consists of around 70,000 words, while an external document is only around 1,000 words long. Figure 2 illustrates the distribution of documents.

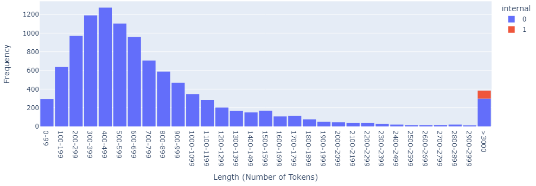

Figure 2: Distribution of Document Length by Internal/External

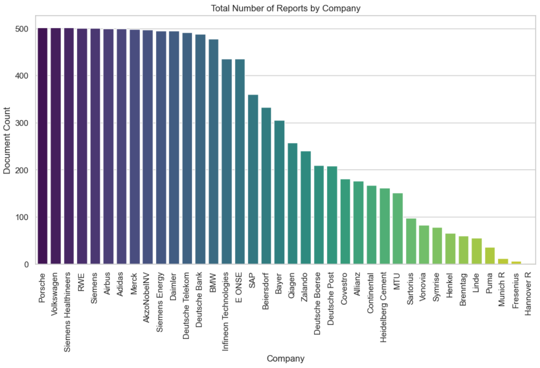

Figure 3: Documents per Company

The number of documents per company also varies considerably. For 15 companies, more than 400 documents are available, for 10 companies, less than 100 (including Fresenius and Hannover Re). The distribution of documents at company level is shown in Figure 3, based on the data from Tim.

Each document is assigned one or more ESG topics. In total, there are 356 different ESG topics in the data set. The most frequent topics are social and environment. After that, the number of ESG topics drops and continuously decreases with each additional topic. Figure 4 illustrates the frequency of the top 20 ESG topics.

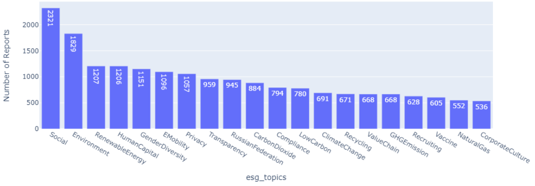

Figure 4: Number of Document Assignments per ESG Topic

In terms of document type (datatype), business, general and technical reports account for around 93 percent of all reports. These are the three most important datatypes in the data set. All other datatypes have subordinate relevance.

2.3.2. Exploratory Text Analysis

Regarding the exploratory text analysis, it is started with a playful approach, which brings out the relevance of individual words grouped by company very well. Tim has computed wordclouds for each company using a TF-IDF analysis. In Figure 5 the results of six randomly selected companies are visible.

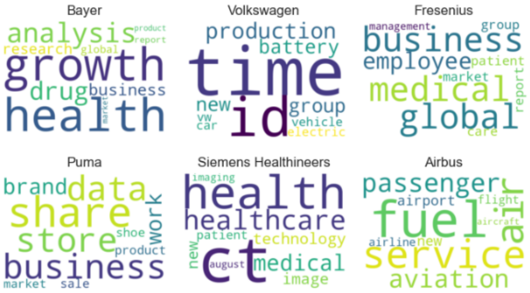

Figure 5: Number of Document Assignments per ESG Topic

From the wordclouds, it can be seen for example that for Bayer the words health, drug, and research seem to have great meaning, which makes sense for a pharmaceutical company. At Volkswagen, interestingly, the words battery and electric appear. This underlines the change that the automotive industry is currently undergoing from gasoline-powered cars to electrically powered ones. The word fuel is interesting at Airbus. Saving and optimizing fuel has been a major topic in the aviation industry for many years, both from a cost and environmental perspective, and this is well reflected in this wordcloud.

Based on the polarity calculations of Tim and Arian, a first impression can be gained about how the sentiment distribution could look like. Of course, these results are still to be taken with caution at the time of stage one and to be verified with the results of stage two and three. In the following, polarity comparisons are made with regard to company, sector and internal (internal vs. external documents). Figure 6 displays the distribution of polarity.

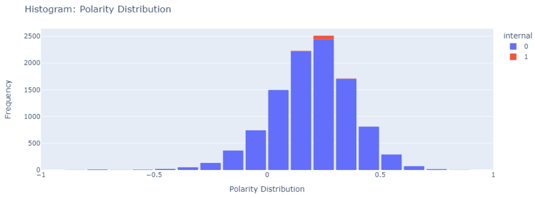

Figure 6: Polarity Distribution

The polarity distribution shows that between 0.2 and 0.3 most reports occur. It is a normal distribution centered at about 0.2, which means that most reports have a slightly positive sentiment. According to this distribution, internal reports do not have a significantly more positive sentiment than external reports. Figure 7 shows polarity box plots on company level.

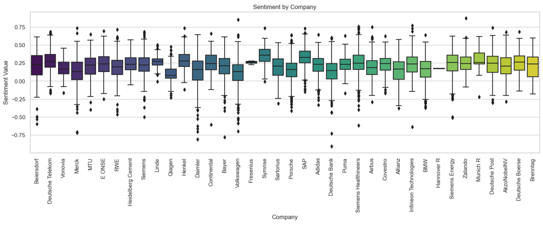

Figure 7: Polarity Box Plots by Company

The box plots definitely display some differences between companies. As expected from the histogram, median sentiment values are all positive. Symrise is the company with the highest median sentiment value, while Qiagen has the lowest median value. Furthermore, strong sentiment outliers can be observed for Deutsche Bank. Daimler and Volkswagen are showing quite a lot of negative outliers. Figure 8 shows polarity box plots on sector level.

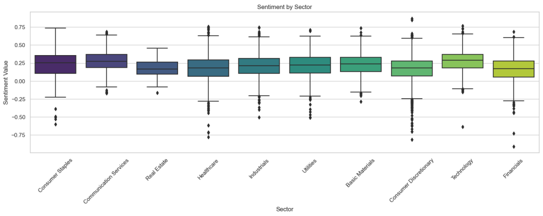

Figure 8: Polarity Box Plots by Sector

Sentiments by sector show a more homogeneous picture. The sectors with the highest median sentiment values are Communication Services and Technology, while the Financial sector displays the lowest median value. The sectors Healthcare and Consumer Discretionary (also includes Auto Manufacturers) are showing most of the negative outliers. Figure 9 visualizes polarity box plots regarding internal vs. external documents.

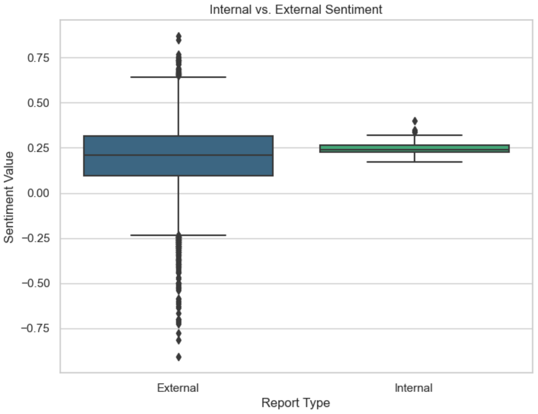

Figure 9: Polarity Box Plots by Internal / External Documents

The comparison of the sentiment of external vs. internal documents is showing an expected pattern: Internal documents are more standardized and therefore have a lower variance and interquartile range, while external documents are more heterogeneous and have a higher variance and interquartile range. According to the box plots, the median sentiment value of internal documents is slightly higher than for external documents. The difference does not look significant, but it still may be an indicator that some extent of greenwashing is present in internal documents.

2.3.3. Time Series

A time series analysis of the ESG topics represents the end of the exploratory data analysis. The time series plot in Figure 10 shows how the occurrence of ESG topics has developed over a period of around three years. For better interpretability, only the top 15 ESG topics have been included.

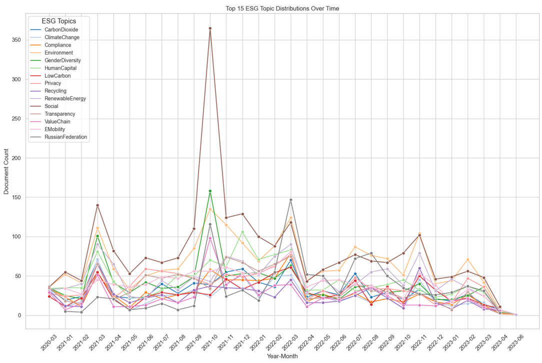

Figure 10: Time Series of ESG Topics

The time series of ESG topics clearly shows some peaks. However, the results have to be interpreted with caution. Even though it is an indicator, peaks do not necessarily mean that topics were more important at these peaks or significant events happened. E.g., in October 2021, a lot of documents mentioned the topic social. However, in a Google search, it could not be found any explainable reason for this peak. On the other hand, one medium published an unusually large number of reports at this time. Meaning, one of the reasons for peaks and drops is simply that more reports are published in spring (March; annual report season) and autumn (October) than in summer and winter. The Russian Federation theme is interesting, however. It appeared for the first time in January 2021 and was at times one of the most frequent topics, especially in February 2022 when the war started.

2.4. Summary Stage I

The key findings from stage one are that the data set and especially the content were raw and required some preprocessing. Especially, the text preprocessing was a challenging task. A comparison shows that the preprocessing was generally quite similar among all authors. With regard to the text preprocessing, however, differences were also evident, especially in the scope. Tim chose the most extensive approach, while Levin decided to apply a leaner proceeding, assuming that LLM might not require heavily preprocessed text. With regard to the exploratory data analysis, several aspects were interesting. The wordclouds impressively showed which words have the greatest characteristics in relation to the individual companies. There are relatively few internal reports in the data set, but these are much longer than the external documents. The polarity comparisons showed sentiment medians and averages of around 0.2 to 0.3, which would mean that the majority of the documents have a positive sentiment. An internal vs. external comparison shows that internal documents have a slightly higher median sentiment value. These findings will be deepened in the subsequent project stages.

3. Stage II

In stage two, data annotation was performed as feature engineering in preparation for stage three. First, an annotation data set was generated. Then, a manual sentiment annotation was performed. Three categories were used: 0 for negative, 0.5 for neutral and 1 for positive. Each sentence had to be read and annotated manually. Subsequently, different LLM's were tested, and the results were compared with the manual sentiment annotation. Finally, the most suitable LLM was selected for annotation of the whole data set and the LLM annotation was performed. In this chapter, the individual steps are described.

3.1. Annotation data set Generation

All three authors opted for sentence-level sentiment annotation, as this is the recommended approach of the task owner, SwissText, and provides more granular insight. While Levin worked with a sample of 200 sentences, Arian prepared a data set of 500 sentences, as this is considered the gold standard. Tim chose 1,000 sentences to get even more test samples than the gold standard. Levin created his data set by randomly selecting 100 documents and then choosing two sentences from each document. He specifically selected sentences longer than 25 characters and containing at least one verb to capture more meaningful sentences. Furthermore, the annotation process was repeated once, as the output contained only few positive examples. To ensure that both external and internal records are present in the sample, Arian and Tim chose an oversampling approach. Since less than 1 percent of the documents are internal, they generated their sample in such a way that 20 percent of the sentences came from internal documents.

3.2. Manual Sentence Annotation

Interestingly, different approaches were taken for manual sentiment annotation by Arian, Tim, and Levin. The results are also considerably different.

3.2.1. Approaches

Arian and Tim used a similar classical sentiment classification approach. That is, sentences were primarily classified according to negative (0), neutral (0.5) and positive (1) meaning. For example:

- The strategy of the last years was very successful. = Positive.
- The strategy of the last years was alright. = Neutral.
- The strategy of the last years turned out to be disadvantageous for the company. = Negative.

Where possible and useful, both made an attempt to classify sentiment in relation to ESG topics. However, this was often not possible at sentence level outside context, so both considered a classical sentiment classification approach as best.

Levin, on the other hand, tried a more task-related approach. He classified the sentences as greenwashing (1), sustainable (0) and neutral (0.5). The different annotation strategies now allow an interesting comparison of the results.

3.2.2. Results

The results of the manual sentiment annotation are presented numerically and graphically below.

Table 1
Results of Manual Sentiment Annotation

Name	0	0.5	1
Arian	101	174	225
Tim	100	706	194
Levin	34	130	36

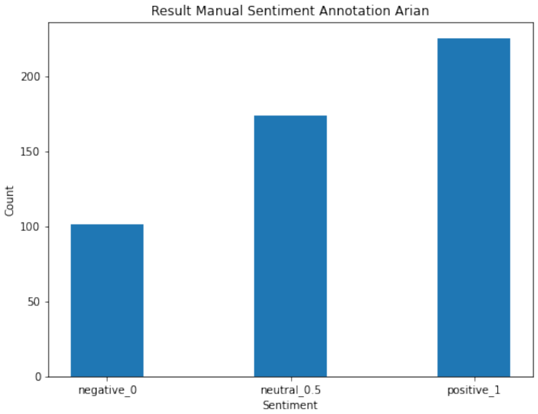

Figure 11: Results Manual Sentiment Annotation of Arian

Although Tim and Arian used a similar annotation approach, their results are very different. While Tim annotated about 70 percent of the sentences with sentiment neutral, Arian annotated the most sentences (45 percent) with sentiment positive. Interestingly, Levin and Tim have a virtually identical distribution of sentiments, even though a different annotation strategy was used.

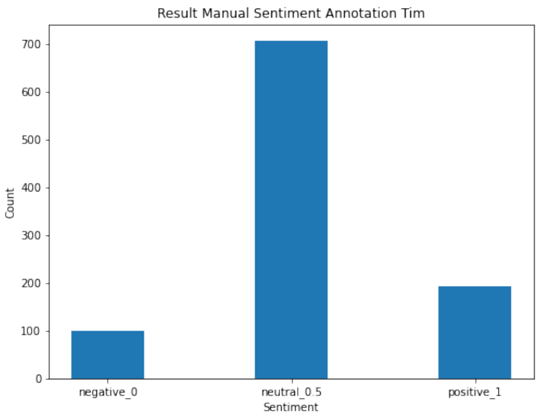

Figure 12: Results Manual Sentiment Annotation of Tim

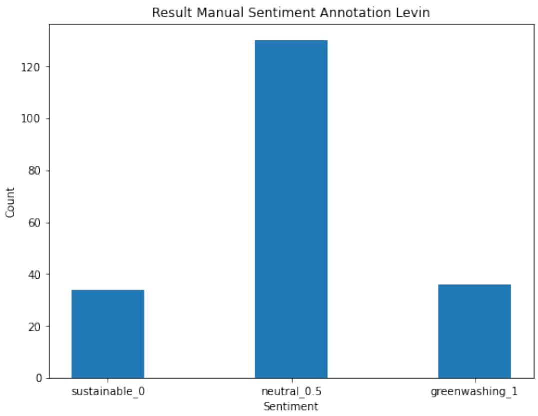

Figure 13: Results Manual Sentiment Annotation of Levin

3.3. LLM Testing for Annotation

After the manual sentiment annotation was done, different LLM's were tested, and the results were compared with the manual sentiments. Zero-shot classifications as well as Few-shot classifications with different prompting strategies were used. The selection of strategy largely depends on the complexity of the task and the nuances of the text involved. While zero-shot strategies are quicker and more straightforward, few-shot strategies can offer a more nuanced understanding of the task and potentially improve the model's performance. They help the model generalize the sentiment classification pattern from the examples provided. Therefore, a balance of both strategies, depending on the specific use case and requirements, can result in more accurate and reliable sentiment analysis outcomes.

3.3.1. Zero-Shot-Strategy

Zero-shot strategy means that only the sentence to be predicted is passed to an LLM as input, without any further information. All three authors have worked with the Hugging Face library in relation to the zero-shot strategies. Interestingly, there are some similarities with respect to the LLM selection. Since Arian and Tim performed a classical manual sentiment annotation, they primarily used fine-tuned models specifically for sentiment classification. Levin, on the other hand, used a model fine-tuned for classification, which is more flexible with respect to classification. This allowed him to test his labels (greenwashing, sustainable and neutral). The following models were tested:

- **distilbert-base-uncased-fine-tuned-sst-2-english:** Arian, Tim
- **siebert/sentiment-roberta-large-english:** Arian, Tim
- **ahmedrachid/FinancialBERT-Sentiment-Analysis:** Arian
- **facebook/bart-large-mnli:** Arian, Levin
- **cardiffnlp/twitter-roberta-base-sentiment:** Tim
- **Seethal/sentiment-analysis-generic-data set:** Tim
- **nlptown/bert-base-multilingual-uncased-sentiment:** Tim

3.3.2. Few-Shot-Strategy

Few-shot strategy means that, in addition to the sentence to be classified, further contextual information is provided as input to an LLM. This can be, for example, some sentiment classified example sentences or task specific information. Various prompting strategies can thus be used for few-shot classification.

Arian and Tim both tried GPT language model because it is a powerful question-answer model that allows the input of contextual information and can be used for sentiment classification, among other things. While Arian used the text-davinci-002 model, Tim decided to work with text-davinci-003. Arian performed two annotations with GPT, meaning two prompting strategies. Once three positive, three negative and three neutral sentences from the manually annotated test data set were passed as context, once only one sentence each was passed as sentiment examples. While Arian decided to provide some sentiment classified examples, Tim used the task specific information approach. He asked the LLM to determine the sentiment regarding sustainability practices. In addition to the GPT model, Tim applied the same prompting strategy to a FLAN-T5 model as well.

Levin applied another Hugging Face model for the few-shot strategy with deepset/roberta-base-squad2, which is suitable as a question answer model for various prompting strategies. In his prompting strategy, he gave the model both, example sentences and task-specific contextual information as input. Specifically, he asked the LLM whether the input sentence implied greenwashing, sustainable, or neutral behavior. The sentence and one example sentence each for greenwashing, sustainable and neutral were then passed.

3.3.3. Results

The individual results of the LLM annotation testing can be summarized as follows:

Levin has evaluated his LLM results with a confusion matrix and the metrics accuracy, precision and F1 score, which are judged as good evaluation methods. The results, however, are judged rather critically. Neither of the models applied classified any sentence as greenwashing. The Bart prediction classifier also classified most sentences as sustainable, whereas in its manual annotation most sentences are neutral. Overall, the annotation strategy used is probably less suitable for this task. Nevertheless, it is an exciting approach and definitely worth a try.

Tim evaluated his results using methods of descriptive statistics and visualization of box plots and histograms. The outputs can be seen in Figure 14 and Figure 15.

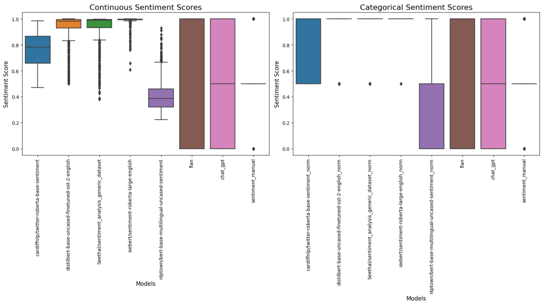

Figure 14: LLM Annotation Results of Tim - Box Plots

His investigation of various models for sentiment analysis led to several noteworthy observations. Foremost among these is the superior performance of the GPT model in terms of agreement with its manual sentiment annotation. This is confirmed by quantitative and visual analyses, with both box plot averages of GPT and manual sentiment and histogram comparisons demonstrating the superior fit of GPT outputs with manual sentiment. However, an important consideration in this context is the cost and computational efficiency associated with each model. Despite its promising performance, the GPT model requires an API call for each execution. This introduces latency that significantly slows down the overall processing speed, making it less suitable for annotating the entire data set under the given constraints. Among the other models evaluated, the nlptown/bert-base-multilingual-uncased-sentiment model shows the most satisfactory performance. It has the smallest difference in mean value to manual sentiment compared to the other models. Furthermore, its processing speed is commendably efficient, placing it in the midfield of the tested models. Consequently, this model offers a good balance between performance and computational efficiency, making it a suitable choice for annotating the whole data set.

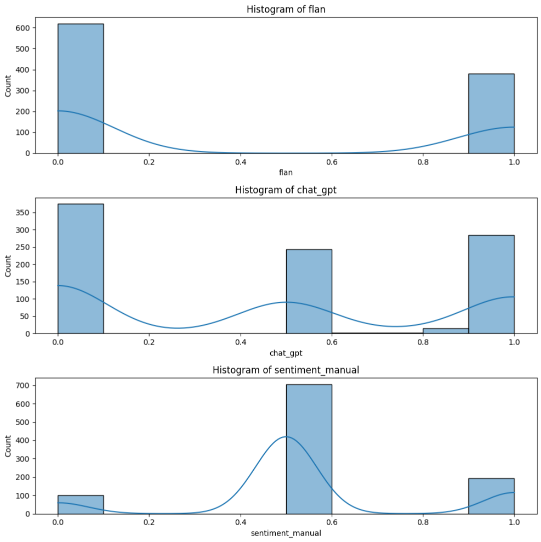

Figure 15: LLM Annotation Results of Tim - Histograms

Arian compared his results using graphical histograms as well as calculating the deviation of each model to the manual sentiments. Figure 16 and Figure 17 illustrate his results.

His results can be summarized as follows. When manually annotated, the majority of sentences were marked as positive. The BERT, RoBERTa and BartLarge models

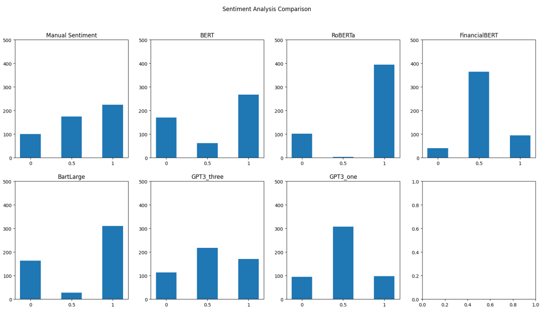

Figure 16: LLM Annotation Results of Arian - Histograms

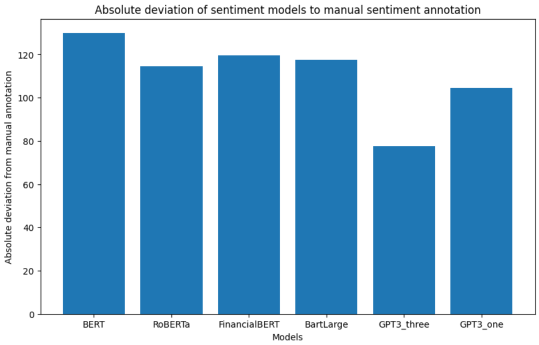

Figure 17: LLM Annotation Results of Arian - Deviation

also have the most values for positive sentiment. The FinancialBERT and GPT models scored the majority of sentences as neutral. Regarding the distribution, the BERT and GPT models (with 3 examples each) have the most reasonable results, even if the distributions differ from the manual annotation. For the BERT model, the proportion between 0 and 1 is similar to the manual annotation, and for GPT the distribution is the most balanced. The GPT model has the smallest absolute deviation from manual annotation. The other four models have a similar deviation. Overall, the GPT model with three example sentences each (few-shot-classification-strategy) is evaluated as the best performing LLM in the present case and should be used for annotation of the whole data set. Unfortunately, the GPT model is not free of charge and annotating the entire data set would result in significant costs. Therefore, the BERT model (DistilBERT base uncased fine-tuned SST-2) is used for the annotation of the data set, as it is considered to be the model with the second-best performance regarding this task.

3.4. LLM Annotation

Since the LLM tests of Tim and Arian yielded better results, only the LLM annotations of Tim and Arian will be presented here.

For the dataset annotation, Tim utilized the nlptown/bert-base-multilingual-uncased-sentiment model by Hugging Face, generating a score on a continuous scale ranging from 0 to 1. A secondary categorical sentiment score was consequently derived. This discretized sentiment, based on the continuous sentiment score, was categorized into 0 (negative), 0.5 (neutral), and 1 (positive), using 0.33 and 0.66 as threshold values.The distribution of the sentiments in the whole data set is shown in figures 18 and 19.

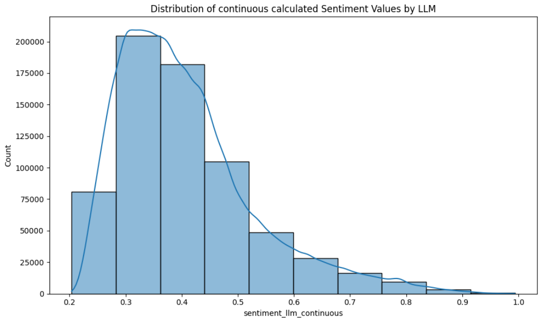

Figure 18: LLM data set Annotation of Tim - Continuous Sentiment

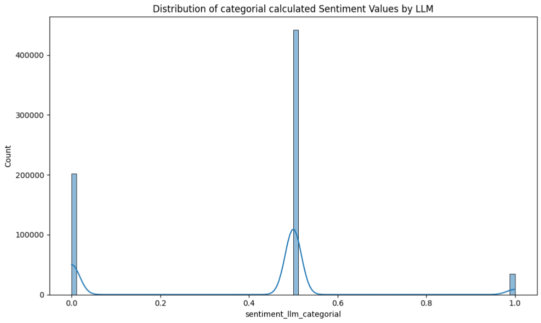

Figure 19: LLM data set Annotation of Tim - Categorical Sentiment

Arian used the Hugging Face model distilbert-base-uncased-fine-tuned-sst-2-english for his data set annotation. He has only done a categorical annotation. In addition, he calculated the average sentiment value at the document level and presented the distribution of the sentiments in a histogram and in box plots. In the box plots, internal and external documents were compared with each other in order to obtain an indicator of whether the average sentiments for internal documents are higher than for external documents. The results are shown in Figure 20 and Figure 21.

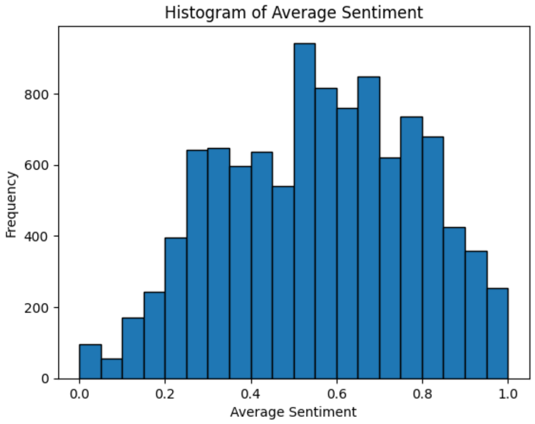

Figure 20: LLM data set Annotation of Arian - Histogram

The box plot comparison is very similar to the polarity calculation and comparison from stage one by Arian and Tim, and shows only a slightly higher median sentiment value for internal documents compared to external documents. In addition, the median sentiment value is around 0.6, which again indicates that a majority of the documents have a positive sentiment.

3.5. Summary Stage II

Stage two showed that it is very important to find a suitable strategy for manual annotation. An unsuitable strategy can lead to the fact that subsequently no suitable LLM can be found for the complete annotation. Another exciting finding is that even comparable annotation strategies can produce very different manual sentiments, as seen in the example of Tim and Arian. This is probably due to individual judgement of sentences. When choosing LLMs, it is important to use appropriate, already fine-tuned models. Models not fine-tuned for sentiment classification or base LLMs proved to be unsuitable for the task. Both Tim and Arian, considered to have achieved the best results with GPT. This is interpreted to mean that the GPT model is the most sophisticated model and, as a question-answer model, allows the input of contextual information and sentiment examples (few-shot strategy). Unfortunately, for reasons of cost and computational efficiency, it was not possible to use this model for annotation of the entire data set, so the best performing Hugging Face model had to be used in that respect.

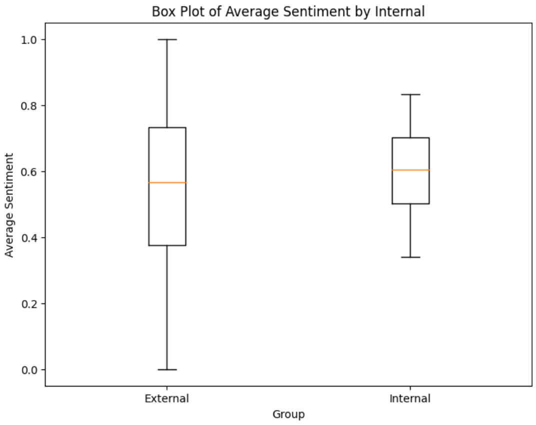

Figure 21: LLM data set Annotation of Arian - Box Plots

4. Stage III

Stage three consists of training a sentiment classifier, performing sentiment analysis and finally compare outputs for internal and external documents. In this section, the methodology and key findings of stage three are presented.

4.1. Methodology

According to the task description, a sentiment classifier was trained which produces as output scores on a continuous scale between 0 and 1 (0 for negative, 1 for positive). Two possible approaches were considered for training a sentiment classifier. The first was to train an ordinary machine learning model, such as a neural network or a support vector machine. The second option was to take a pre-trained Hugging Face LLM as a base and fine-tune it for sentiment classification. The second approach was judged to be the more appropriate because it is a widely used and valid approach in the field of NLP. In order to use a suitable LLM for the task, several LLMs were selected in a first step and trained on a reduced subset of 96,450 sentences with adjusted class balance (48,225 sentences from internal documents and 48,225 sentences from external documents). The training was done at sentence level, as this was assessed to be a more appropriate approach. fine-tuning at document level was also considered, but rejected. Due to the truncation, only the first 512 word tokens are used in the Hugging Face LLMs. This would lead to wrong or misleading training data, since most documents include more than 512 words. A train/dev/test split of 70/15/15 was applied. This means that 70 percent of the data was used as training data, 15 percent as development data for model evaluation during training and 15 percent as final test set for model evaluation. The trained LLMs were compared using the metrics mean squared error (MSE), mean absolute error (MAE) and R2 because the objective is a continuous classifier which represents a regression problem. A comparison of the metrics allows determining the best performing model. Subsequently, the selected LLM was trained with the full data set, again with a train/dev/test split of 70/15/15, and evaluated. The trained sentiment classifier was then used to predict sentiment for each sentence in the full data set. Finally, a document-level aggregation was performed with the average of all sentences in a document and the sentiments of internal vs. external documents were compared. The methodology of stage three is displayed in Figure 22.

4.2. LLM Selection

As sentiment classifier, four different LLMs were considered based on the pre-trained language family: [3]

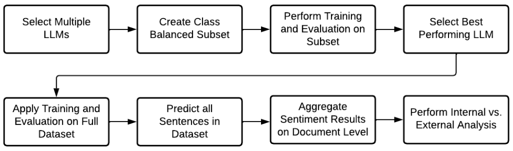
Select Multiple LLMs
Create Class Balanced Subset
Perform Training and Evaluation on Subset
Select Best Performing LLM
Apply Training and Evaluation on Full Dataset
Predict all Sentences in Dataset
Aggregate Sentiment Results on Document Level
Perform Internal vs. External Analysis

Figure 22: Methodology - Stage III

- **distilbert-base-uncased:** The distilbert-base-uncased model was selected because distilbert models demonstrated good results for Tim and Arian in stage two. Furthermore, it was regarded as a good choice to have a lightweight model in the selection.
- **roberta-base:** The RoBERTa model was chosen because it is a further development of BERT and thus expected to perform better. This model benefits substantially from extended training duration, larger data batches, and an increase in data set size. Its performance further increases by eliminating the next sentence prediction objective and integrating longer sequences during training. Lastly, the model's optimization is boosted by dynamically altering the masking pattern applied to the training data.[4]
- **xlnet-base-cased:** XLNet is a pre-trained model for natural language processing tasks that combines the advantages of both auto-regressive language models and denoising auto-encoding models like BERT. Unlike BERT, XLNet mitigates dependency issues between masked positions and avoids a pre-train-fine-tune discrepancy by employing a generalized auto-regressive method. This approach allows for learning bidirectional contexts by maximizing expected likelihood over all possible factorization orders. Furthermore, it integrates the strengths of Transformer-XL, a leading auto-regressive model, into its pretraining procedure. Empirical evidence suggests that XLNet surpasses BERT in performance across a range of tasks including question answering, natural language inference, sentiment analysis, and document ranking.[5]
- **flan-t5-base:** T5 models are normally used for text-to-text tasks. Therefore, fine-tuning T5 for a classification task with a continuous prediction between 0 and 1 is a bit of a diversion. Still, the model was tested, and the training could be started. However, the user-defined loss function did not work properly (did not drop, but showed 0

throughout the whole training), and there is also uncertainty about the correct encoding. Therefore, the model was dropped at an early stage and no more compared to the other three models.

After the initial LLM selection, several fine-tuning runs have been executed and the results inspected by using Tensorboard. Note that the evaluation is performed on the development data. Figure 23 (Fine-Tuning Loss per Model) and Figure 24 (Evaluation Metrics on Development Data) display the results of the best considered runs per model (Legend: distilbert = pink, roberta = orange, xlnet = blue).

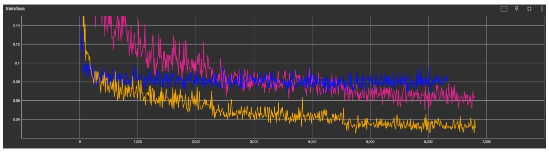

Figure 23: Fine-Tuning Loss per Model

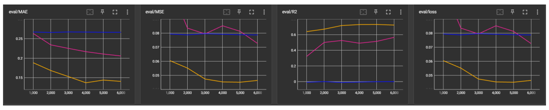

Figure 24: Fine-Tuning Metrics per Model

After the fine-tuning with 3 epochs per 6,000 steps on 96,450 single sentences, the following metrics can be observed:

Table 2
Comparison of Evaluation Metrics on Development Data

Model	MAE	MSE	R2
distilbert	0.2056	0.07265	0.5666
roberta	0.1405	0.04641	0.7233
xlnet	0.2662	0.07874	-8.03e-5

The results can be summarized as follows: XLnet performed worst according to MSE, MAE and R2. All metrics are quite flat and not increasing (R2) respectively decreasing (MSE, MAE) as it was expected. Perhaps this is due to data quality problems. E.g., too less training data, wrong preparation or tokenization etc. RoBERTa outperforms distilbert (lower MAE, MSE and higher R2) as expected, since it is a further evolution of BERT and distilbert a lightweight version of BERT.

Finally, after training the models and evaluating them against development data, the models were compared by evaluating them against test data. Figure 25 visualizes the inference runs on the test data set (Legend: distilbert = pink, roberta = orange, xlnet = blue).

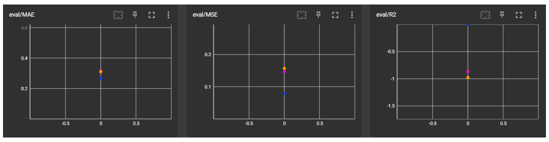

Figure 25: Inference Metrics per Model

Table 3 summarizes the metrics of the inference runs, including runtime metrics.

Table 3
Training and Evaluation Results

Model	MSE	MAE	R2
distilbert	0.147	0.316	-0.870
roberta	0.157	0.311	-0.977
xlnet	0.079	0.266	0.000
Model	**Runtime (s)**	**Samples (/s)**	**Steps (/s)**
distilbert	72.85	198.61	24.83
roberta	70.47	205.32	25.67
xlnet	836.21	57.67	7.21

In general, all models demonstrated bad performance according to the MSE, MAE and R2 on completely unseen, new data. Surprisingly, XLNet performed the best out of the three models on test data. According to the metrics from the fine-tuning, the best results are expected from a RoBERTa model even if the model did not show a good inference performance. Therefore, it was decided to use RoBERTa for the fine-tuning on the full data set.

4.3. Full Training on Selected Model and Evaluation

After deciding to continue with the RoBERTa (roberta-base) model, the model training on the full data set took place. To speed up the full training, a state of the art Parameter-Efficient Fine-Tuning method (PEFT config with sequence classification task) was used. RoBERTa supports LoRa, Prefix Tuning, P-Tuning and Prompt Tuning.[6] Three full training runs with a LoRa were applied.

- **Training Run I:** Data = New Stage II Sentiment Predictions (red line, mostly overplotted by run II).
- **Training Run II:** Data = New Stage II Sentiment Predictions (orange Line); Second run on the same data to validate results.
- **Training Run III:** Data = Older Stage II Sentiment Predictions with more equally distributed sentiment classes (green line)

The results of the evaluation metrics (based on development data) are displayed in Figure 26.

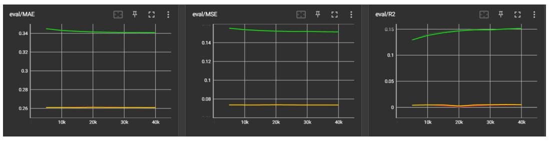

Figure 26: Evaluation Metrics per Training Run

Training run one and two produced basically the same results. Training run three where the older stage two sentiment predictions with more equally distributed sentiment classes were used as input resulted in higher R2, MSE and MAE. The three different training runs on the full dataset on sentences level are displayed in table 4.

Table 4
Different RoBERTa Fine-Tuning Runs and Evaluation Metrics

Training Run	MSE	MAE	R2	Runtime (h)
Run I	0.074	0.261	0.005	2.218
Run II	0.074	0.261	0.005	3.056
Run III	0.151	0.341	0.152	2.977

The model evaluation of training run three on completely new data, respectively the test data (101,784 sentences), resulted in the metrics MSE = 0.152, MAE = 0.341, and R2 = 0.151. The evaluation metrics based on the test data after the full training still show unsatisfactory values. There are two possible explanations for this. Either, something could have been set incorrectly during model training, resulting in an insufficient learning effect of the model. Or it is in fact a data quality issue (e.g., inconsistency in manual sentence annotation, choice of an unsuitable LLM for sentiment annotation in stage two, problems regarding text preprocessing etc.) which leads to the model not being able to learn as necessary with the according training data. However, these are the best possible results which could be produced within this short period of time, and the trained model will still be used for the following sentiment analysis. The next step should of course be to tune the hyperparameters. In addition, it would be important to discuss the results with NLP experts to identify possible sources of error. However, due to time constraints, this cannot be included in the present project.

4.4. Sentiment Analysis

After training the RoBERTa model on two different stage two outputs, the results can be checked and the final sentiment analysis performed. Both data sets were slightly adjusted in discrete class distributions before training. Both fine-tuned RoBERTa models were used to perform a sentiment analysis on all sentences. The results can be observed in Figure 27, Figure 28,Figure 29 and Figure 30.

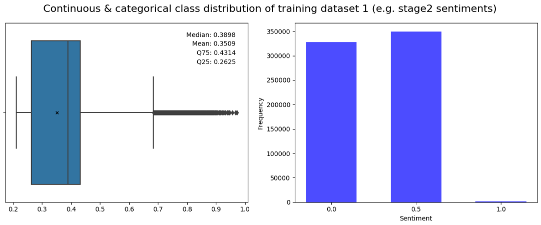

Figure 27: Continuous and Categorical Class Distributions on Training data set I

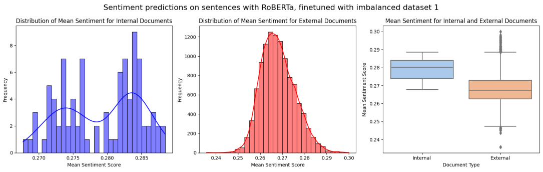

Figure 28: Sentiment Predictions on Training Data Set I and Comparison

It can be observed, the imbalance in the training data set directly affects the sentiment predictions with the fine-tuned RoBERTa model. The class imbalance in training data set I leads to a much more narrow prediction distribution compared to the predictions from the model, which was fine-tuned with a more balanced data set. The mean sentiment scores on document level are quite different. With the first classifier, the median values are

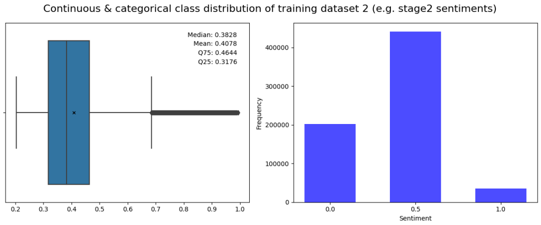

Figure 29: Continuous and Categorical Class Distributions on Training Data Set II

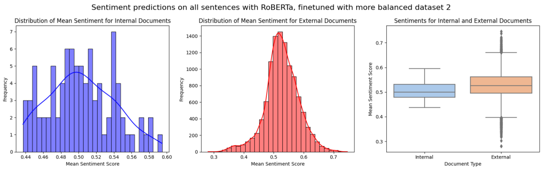

Figure 30: Sentiment Predictions on Training Data Set II and Comparison

at about 0.27 to 0.28. With the second classifier, the median values are close to 0.5. Regarding the distributions, the results with the second classifier seem much more reasonable. The comparisons between internal and external documents do not display significant deviance for both classifier. To conclude on the sentiment analysis, the sentiment analysis did not detect any greenwashing patterns. The sentiment analysis was also carried out at company level. However, no significant systematic greenwashing patterns were identified there either. Perhaps with the exception of Qiagen, for which the second classifier classified significantly higher mean sentiment values for internal than for external documents. This could possibly indicate greenwashing of Qiagen. However, due to the poor evaluation metrics of the classifier, this statement should be taken with caution and no judgment should be made.

4.5. Summary Stage III

In stage three, a sentiment classifier was trained using the data annotation from stage two. According to the task instructions, the model was trained in a way that the output is a continuous prediction between 0 and 1. An LLM approach was chosen. That is, a pre-trained LLM was fine-tuned for the present sentiment analysis. The model was trained and applied at sentence level, as this is considered more appropriate than applying it at document level. Of the selected LLMs, the RoBERTa model showed the most promising results, which is why it was chosen as the classifier. Unfortunately, the evaluation of the metrics showed unsatisfactory results in all training and evaluation phases. The full training was performed on two data sets to compare how the class balance of the data set affects the model. The final sentiment analysis results showed clear and significant differences with respect to the predictions. This means, that class balance or imbalance certainly influences the performance of a predictor. However, both trained classifiers could not detect any significant sentiment difference between internal and external documents, which is why the result of the sentiment analysis is that no greenwashing could be detected in the data.

5. Stage IV

Within this final stage, the technique of sentence embedding is applied to analyze the SDG alignment of German DAX companies, as reflected by the internal and external documents. Furthermore, differences between internal and external perception are pointed out not only for each company but also at sector and industry level. Please note that in contrary to the provided notebook by the task owner, SwissText, it was decided to use the cleaned document contents from stage one instead of the initial raw content. Since text cleaning is very important in NLP, this replacement was regarded as a sensible adjustment.

5.1. Sentence Embedding

First, string representations of all text elements are converted into corresponding Python objects. Second, texts are formed by concatenating columns title and cleaned content. In addition, the company Hannover RE is excluded for stage four, as only internal documents are available for it. Third, unique companies, sectors and industries are assigned to lists. The text forming and company, sector and industry storing are necessary preprocessing steps for sentence embedding. Fourth, the retriever model flax-sentence-embeddings/all_datasets_v3_mpnet-base is loaded from Hugging Face to encode company text embeddings and SDG embeddings.

5.2. SDG Alignment of the DAX Companies

In this section, SDG alignment is modeled as the similarity between the company-related texts and the SDG descriptions. The similarity function is defined using standard cosine similarity. Cosine-similarity was already used by the creators of the model 'all-mpnet-base-v1' for its fine-tuning [7]. Then, alignment scores are analyzed, including visualizations and interpretations. All analyzes are executed on a company, sector and industry level.

5.2.1. Most Relevant SDGs for Companies

Figure 31 shows the importance of the SDG goals sorted according to the calculated cosine-similarity between companies- and SDG-embeddings.
The chart shows that 'Affordable and Clean Energy', with a value of 0.48, is the SDG with the greatest relevance for DAX Companies. The SDG 'Industry, Innovation and Infrastructure' is also highly relevant. The SDGs 'Peace, Justice and Strong Institutions' and 'Gender Equality' appear to be of less relevance. These results are considered as to be consistent, as Germany is in the middle of the energy transition and renewable and affordable energy is therefore an important topic for many companies. In addition, Germany is a strong industrial nation, especially with the automotive industry, which is currently very busy with innovations in the field of electric cars. The lesser important SDGs are not particularly surprising, as Germany is a well-functioning, constitutional state with reliable institutions in an international comparison. In terms of gender equality, Germany also has an already progressive equality, and divisive voices receive less attention.

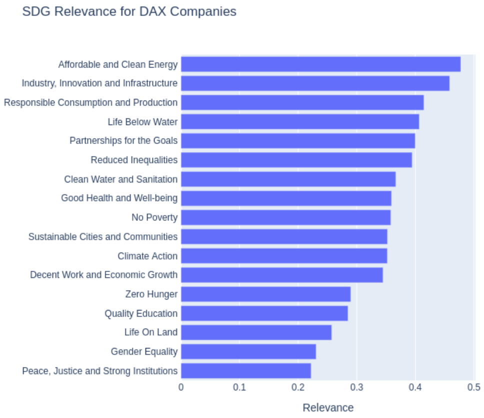

Figure 31: SDG Relevance for DAX Companies

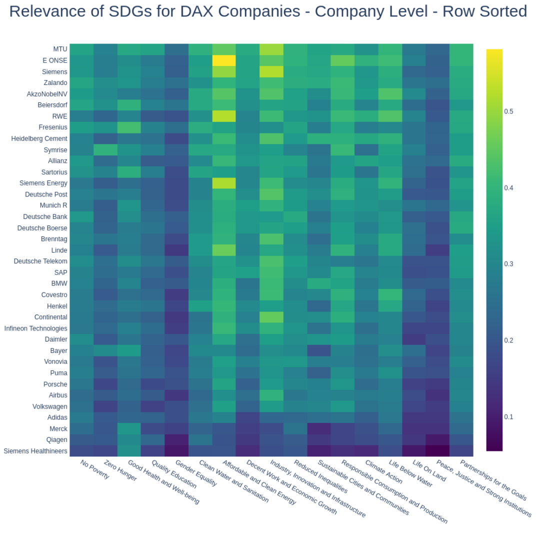

Figure 32: Relevance of SDGs for DAX Companies at Company Level - Rows sorted by Sum of Row

Figure 32 shows the overall importance of SDG goals for DAX companies, using the mean of the SDG embeddings over all documents. As mentioned, SDG alignment is defined as the similarity between the company-related texts and the SDG descriptions. In addition, the plot is sorted by the sum of each row. The heatmap shows that the companies MTU, E ONSE and Siemens have the highest overall sum of SDG alignment while Siemens Healthineers, Qiagen and Merck have the lowest. The overall highest relevance of a SDG goal for one company is ’Affordable and Clean Energy’ for E ONSE (0.58) which makes absolutely sense since E ONSE is an operator in energy networks and energy infrastructure and an energy provider as well.

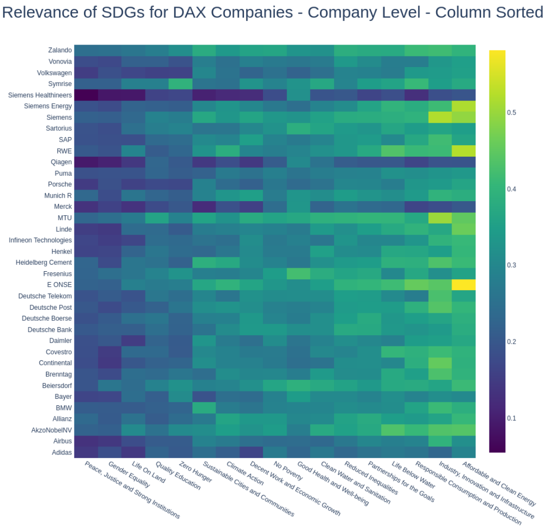

Figure 33: Relevance of SDGs for DAX Companies at Company Level - Columns sorted by Sum of Column

When sorting the overall SDG alignment by the sum of each column, as visualized in Figure 33, the most important SDGs for DAX companies can be identified on the right side of the plot. The three most important SDGs are seemingly ’Affordable and Clean Energy’, ’Industry, Innovation and Infrastructure’ and ’Responsible Consumption and Production’. Figure 34 compares internal and external SDG embedding based on the difference of the internal minus the external SDG scores. The darker an SDG is colored in red, the more relevant it is according to the company, but less relevant according to external sources. Yellow colored boxes indicate more consistent internal and external scores. At the bottom of the chart, in increasing green color, are the companies for which the external sources show greater alignment with the SDG descriptions than the internal documents.

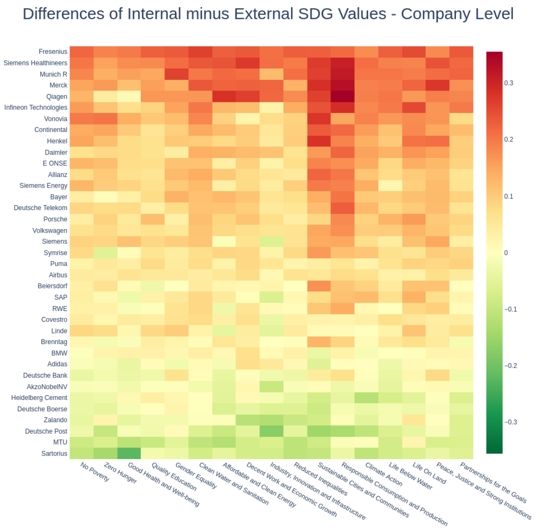

Figure 34: Differences of Internal minus External SDG Values at Company Level

5.2.2. Most Relevant SDGs for a Single Company: BMW

In this analysis, the focus is on the most important SDGs at the company level. In Figure 35, the SDG alignment of BMW is examined, by its ’internal’ and ’external’ embeddings, averaging these values and measuring their

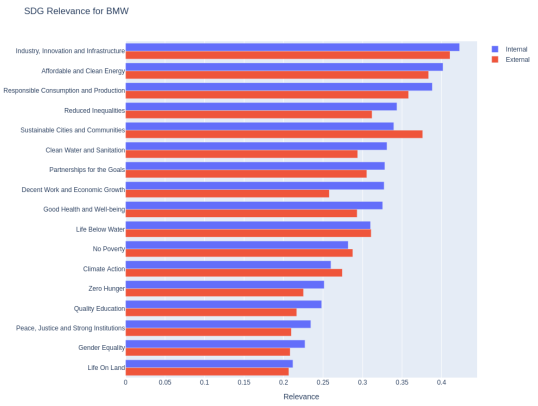

Figure 35: SDG Relevance for BMW

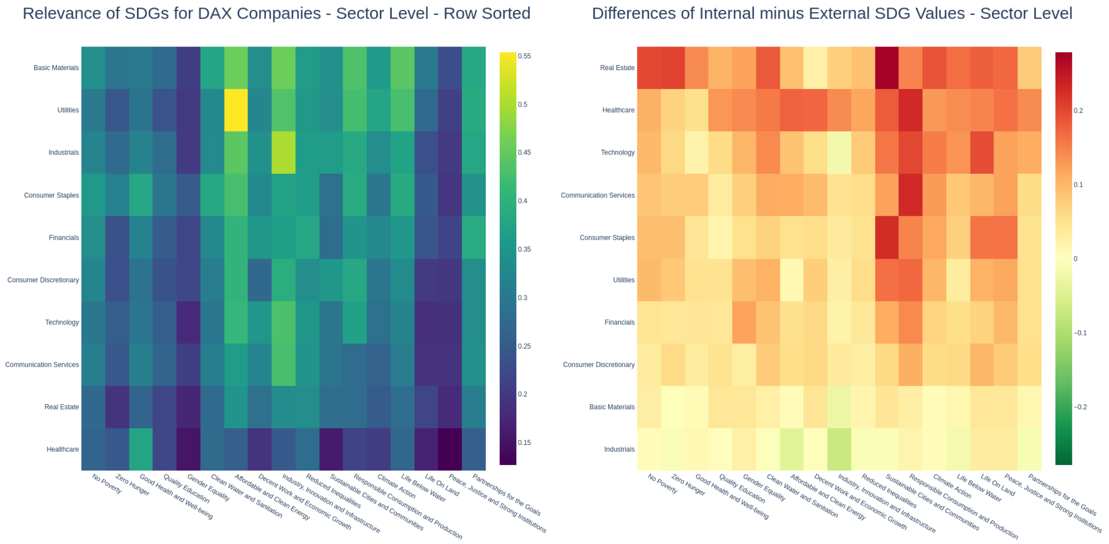

Figure 36: Relevance of SDGs for DAX companies at Sector Level (left-hand graph) and Differences of Internal minus External SDG Values at Sector Level (right-hand graph), Rows sorted by Sum of Row

similarity to each of the SDGs. Based on the mentioned results, a few key findings can be extracted. The most important insights are listed below:

- The SDGs that are most relevant internally, externally and thus also overall are 'Industry, Innovation and Infrastructure', 'Affordable and Clean Energy', and 'Responsible Consumption and Production'.
- It is considered as positive that the three internally most relevant SDGs correspond to the three externally most relevant SDGs.
- The three most relevant SDGs are consistent for an automotive group like BMW.
- The biggest discrepancies are in the SDGs 'Decent Work and Economic Growth', 'Clean Water and Sanitation', and 'Sustainable Cities and Communities'. In the case of the first two, the internal relevance is higher than the external relevance (effectively positive difference), and in the case of the last, the external relevance is higher than the internal relevance (effectively negative difference).

5.2.3. Most Relevant SDGs for Industries/Sectors

On sector and industry level, no detailed results for one sector as above for one company are displayed since the heatmaps contain all relevant information regarding SDG relevance and alignment. The left-hand graph of Figure 36 shows the importance of the various SDG targets per sector. The stronger the green or even yellow portion of a box's color, the more relevant an SDG is. For example, the goal previously identified as the most important SDG, 'Affordable and Clean Energy', appears to be particularly important for the utilities sector, which makes perfect sense. Overall, the SDG targets for the basic materials, utilities, and industrial sectors appear to be most prominent. On the other hand, SDG targets for the healthcare and real estate sectors seem to be of little relevance. Nevertheless, there are some SDGs that are also relevant for these sectors (e.g., 'Good Health and Well-being' for the healthcare sector or 'Affordable and Clean Energy' for the real estate sector).

In the right-hand graph of Figure 36 only a few green fields, but many red ones are visible in the graph. This means that at sector level, the SDG goals are in most cases more relevant internally than externally. Especially, the SDG 'Responsible Consumption and Production' shows a strong pattern of higher internal scores among almost all sectors.

The left-hand graph of Figure 37 shows the importance of the various SDG targets per industry. The stronger the green or even yellow portion of a box's color, the more relevant an SDG is. For example, the goal previously identified as the most important SDG, 'Affordable and Clean Energy', appears to be particularly important for the utilities - diversified industry, which seems plausible. Overall,

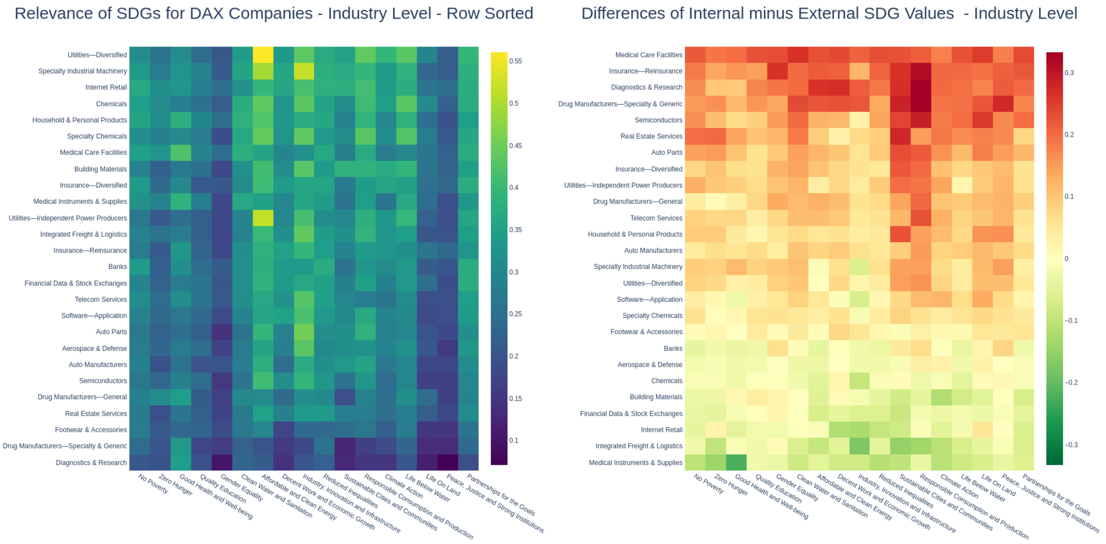

Figure 37: Relevance of SDGs for DAX Companies at Industry Level (left-hand graph), and Differences of Internal minus External SDG Values at Industry Level (right-hand graph), Rows sorted by Sum of Row

the SDG targets for the utilities - diversified, specialty industrial machinery, and internet retail industries appear to be most relevant. On the other hand, SDG targets for the diagnostics & research and drug manufacturers industries seem to be of little relevance.

When analyzing the differences in the internal and external SDG scores per industry, additional patterns can be observed. As the right-hand graph of Figure 37 above is sorted by the sum of each row, the industries at the bottom of the plot show the largest negative discrepancy, colored in intensifying red. A negative discrepancy is assumed if the internal values are higher than the external values. Especially the medical instrument & supplies and integrated freight & logistics industries show the largest negative inconsistency between internal and external company-related texts in regard to the SDG descriptions.

5.3. Summary Stage IV

Using sentence embeddings, SDG alignment was modeled between company-related texts and SDG descriptions using standard cosine-similarity. A variety of patterns was observed not only on company but also on sector and industry level. Overall, the SDGs 'Affordable and Clean Energy', 'Industry, Innovation and Infrastructure' and 'Responsible Consumption and Production' seem to be most important. However, especially for 'Responsible Consumption and Production', differences of the internal and external SDG alignment scores were identified. In particular, the external scores were larger than the internal SDG embeddings. As the earth overshoot day for Germany (May 4, 2023) lies already in the past, greenwashing is possibly present [8]. Nevertheless, as overconsumption and -production are structural problems, binding global frameworks would be needed. As another example at company level, a SDG analysis for BMW was performed, and it was considered as positive that the three internally and externally most relevant SDGs correspond. On sector level, SDG goals for sectors basic materials, utilities, and industrial sector seem to be the most prominent, which seems plausible given their importance for the German economy. A similar pattern is apparent when analyzing the SDG at industry level, however, when considering differences between internal and external scores, a possible starting point for future research is detected. Few industries, led by medical instrument & supplies, show higher internal scores for all SDGs. To investigate this circumstance in more detail might be both interesting and challenging.

6. Conclusion

Stage I: The DAX ESG media data set requires sophisticated preprocessing. The content of the documents is raw, and the data cleaning has to be implemented with care. Extensive data exploration can be used to find possible issues. Furthermore, polarity comparisons showed sentiment medians and averages of around 0.2 to 0.3, indicating at stage one that the majority of the documents have a positive sentiment.

Stage II: A feasible manual annotation strategy is crucial. Even comparable annotation strategies can yield different manual sentiments, possibly depending on the amount of manually annotated sentences, but also due to individual judgement. Furthermore, fine-tuned LLM, such as GPT-3.5-turbo, seemed ideal to implement a few-shot strategy. However, due to reasons of cost and computational inefficiency of the OpenAI API, the next best performing Hugging Face model was used for the annotation of the entire data set.

Stage III: In stage three of the study, a sentiment classifier was trained using data annotation from stage two. The model was trained to produce continuous predictions between 0 and 1, following task instructions. An approach utilizing a pre-trained LLM was selected, specifically the RoBERTa model, which yielded the most promising results among other LLMs considered. However, evaluation of the metrics revealed unsatisfactory results throughout the training and evaluation phases. The training was conducted on two different data sets to compare the impact of class balance on the model's performance. Ultimately, the sentiment analysis did not detect any significant differences between internal and external documents, leading to the conclusion that no greenwashing was detected in the analyzed data.

Stage IV: The SDGs 'Affordable and Clean Energy', 'Industry, Innovation and Infrastructure' and 'Responsible Consumption and Production' seem to be most important for DAX companies. However, especially for the latter, differences of the internal and external SDG alignment scores were identified. In particular, the external scores were larger than the internal SDG embeddings. As the earth overshoot day for Germany (May 4, 2023) lies already in the past, greenwashing is possible with respect to this SDG. Nevertheless, as overconsumption and -production are structural problems, ideally binding global framework conditions should be implemented. In addition, few industries show a large discrepancy between internal and external scores. These industries, led by medical instrument & supplies, show higher internal scores for all SDGs. To investigate this circumstance in more detail might be both, interesting and challenging.

References

[1] SwissText, Detecting greenwashing signals through a comparison of esg reports and public media, 2023. URL: https://drive.google.com/file/d/1FSQ-3EBtCtIxR1hYY32kT9JrUjhdwo4f/view.

[2] Kaggle, Dax esg media dataset, 2023. URL: https://www.kaggle.com/datasets/equintel/dax-esg-media-dataset.

[3] Zhiyuan Liu, Yankai Lin, Maosong Sun, Sentence representation, 2022. URL: https://www.researchgate.net/figure/The-Pre-trained-language-model-family_fig4_342684048.

[4] Yinhan Liu, Myle Ott, Naman Goyal, Jingrey Du, Mandar Joshi, Danqi Chen, Omer Levy, Mikel Lewis, Luke Zettlemoyer, Veselin Stoyanov, Roberta: A robustly optimized bert pretraining approach, 2019. URL: https://doi.org/10.48550/arXiv.1907.11692.

[5] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, Quoc V. Le, Xlnet: Generalized autoregressive pretraining for language understanding, 2020. URL: https://doi.org/10.48550/arXiv.1906.08237.

[6] Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, Sayak Paul, Peft: State-of-the-art parameter-efficient fine-tuning methods, 2022. URL: https://github.com/huggingface/peft.

[7] Hugging Face, Sentence-transformer: all-mpnet-base-v1, 2023. URL: https://huggingface.co/flax-sentence-embeddings/all_datasets_v3_mpnet-base.

[8] Global Footprint Network, Country overshoot days, 2023. URL: https://www.overshootday.org/newsroom/country-overshoot-days/.

Appendix: Individual Contribution

Table 5
Individual Contribution

Task	Arian	Tim	Levin
`Code Stage I`	ind.	ind.	ind.
`Code Stage II`	ind.	ind.	ind.
`Code Stage III`	5%	95%	
`Code Stage IV`	40%		60%
`Report Intro`	95%		5%
`Report Stage I`	95%		5%
`Report Stage II`	95%		5%
`Report Stage III`	80%	20%	
`Report Stage IV`	5%		95%
`Report Conclusion`	25%		75%
`Report Formatting`			100%
`Report & Code Review`	33%	33%	33%