
NLP FOR ESG INVESTING: EXTRACTING SUSTAINABILITY SIGNALS FROM FINANCIAL TEXTUAL DATA

Inaugural-Dissertation

zur

Erlangung des Doktorgrades der
Mathematisch-Naturwissenschaftlichen Fakultät
der Heinrich-Heine-Universität Düsseldorf

vorgelegt von

C. Fabian H. Billert

aus Aachen

Düsseldorf, Januar 2026

aus dem Institut für Informatik der
Heinrich-Heine-Universität Düsseldorf

Gedruckt mit der Genehmigung der
Mathematisch-Naturwissenschaftlichen Fakultät der
Heinrich-Heine-Universität Düsseldorf

Berichterstatter:

1. Prof. Dr. Stefan Conrad
2. Prof. Dr. Stefan Dietze

Tag der mündlichen Prüfung: 14.07.2026

ERKLÄRUNG

Ich versichere an Eides Statt, dass die vorliegende Dissertation von mir selbständig und ohne unzulässige fremde Hilfe unter Beachtung der „*Grundsätze zur Sicherung guter wissenschaftlicher Praxis an der Heinrich-Heine-Universität Düsseldorf*“ erstellt worden ist. Desweiteren erkläre ich, dass ich eine Dissertation in der vorliegenden oder in ähnlicher Form noch bei keiner anderen Institution eingereicht habe. Ich habe keinerlei andere Promotionsversuche unternommen.

Düsseldorf, Januar 2026

C. Fabian H. Billert

Dedicated to my family

ACKNOWLEDGEMENTS

I would like to start by thanking my supervisor, Prof. Dr. Stefan Conrad for his support during these last four years. Although I enjoyed a relatively high degree of autonomy in exploring diverse ideas and delving into captivating research areas, he consistently provided valuable guidance and critical insights on how to progress when necessary. In addition, I want to thank Prof. Dr. Stefan Dietze for acting as the secondary reviewer for my dissertation.

During the work on my thesis, I had the pleasure to collaborate with different teams and individuals. I want to thank my colleagues at GET Capital for their understanding and support during the more stressful periods in the last years, especially Pashutan, Saeed, Nima, Alex, Robert and Claudia. I also want to express my thanks to the pressrelations team for the cooperation on various projects. Without them this work would have been much more challenging. In addition, i want to thank Lefteris Loukas for the fruitful collaboration on edgar crawler, as well as countless discussions about our separate projects. Of course, I would also like to thank my university colleagues: Andreas Abels, Kirill Bogomasov, Manh Khoi Duong, Thomas Germer, Sergej Korlakov, Nina Liebrand, Mihir Mulye, Julia Romberg, Leonie Selbach and Boris Thome, and also the supporting staff, Guido Königstein and Lisa Lorenz.

It is no secret that the personal life can suffer during the stressful periods of a PhD ambition. For that reason, I also want to thank the people who have kept me sane in these difficult times: My parents, Ulrich and Valerie and my siblings Felix, Mathilde, Charlotte and Clara. Most of all, I want to thank my wife Yiqian for being so patient and supportive of me in the last years.

ABSTRACT

Together with the general evolution of the field of natural language processing (NLP), financial information extraction has undergone significant changes in recent years. With increasing access to data, the potential for informed decision making in financial investment processes has similarly risen. However, modern investors increasingly prioritize not only financial returns but also the ethical and societal impact of the companies. In particular, corporate sustainability has emerged as a key concern for both institutional and individual investors.

One significant challenge to sustainable investing is the lack of reliable, high-quality sustainability-related data. Traditional ESG (Environmental, Social, Governance) ratings have been widely used to quantify corporate sustainability practices. However, these ratings face growing criticism due to inconsistency across providers, lack of transparency, and concerns about greenwashing. As a result, there is an increasing interest in more transparent, flexible, and data-driven approaches to sustainability assessment.

This dissertation explores the use of NLP to extract sustainability-relevant information from alternative data sources such as financial news articles. Mirroring the recent development in NLP in general, we investigate different model architectures, including encoder-based representation models and decoder-based generative models, to extract relevant ESG-related content from unstructured text. Our goal is to support more nuanced, transparent, and adaptable sustainability assessments.

With the rising use of data-driven insights for financial decision-making, incorporating investor sentiment therein has also gained significant attention in academia and industry. Studies consistently demonstrate that sentiment can improve portfolio performance, and in response, major financial index providers have introduced sentiment-based indices. However, academic research on their construction remains limited. As a first step toward integrating ESG-related sentiment into investment processes, we start by creating a general sentiment-based version of the DAX-40 index, a stock market index that represents the 40 largest companies in Germany.

Building on this foundation, we then investigate how sentiment specifically related to sustainability topics can be integrated in portfolio optimization. Recognizing that the financial relevance of different ESG aspects may shift over time, we propose dynamic strategies that adjust to the evolving materiality of these factors. This line of research addresses a gap in the literature at the intersection of sustainable finance and sentiment analysis.

ZUSAMMENFASSUNG

Zusammen mit der allgemeinen Entwicklung im Bereich der Verarbeitung natürlicher Sprache (Natural Language Processing, NLP) hat sich auch die Extraktion finanzieller Informationen in den letzten Jahren stark verändert. Mit dem zunehmenden Zugang zu Daten ist auch das Potenzial für fundierte Entscheidungen in Investitionsprozessen gestiegen. Moderne Investoren legen jedoch zunehmend nicht nur Wert auf finanzielle Renditen, sondern auch auf die ethischen und gesellschaftlichen Auswirkungen von Unternehmen. Nachhaltigkeit generell ist zu einem zentralen Anliegen für institutionelle wie auch private Investoren geworden.

Ein wesentliches Hindernis für nachhaltiges Investieren liegt in der Verfügbarkeit und Qualität nachhaltigkeitsbezogener Daten. Traditionelle ESG-Ratings (Environmental, Social, Governance) werden häufig zur Quantifizierung von Nachhaltigkeitspraktiken in Unternehmen verwendet. Diese Ratings stehen jedoch zunehmend in der Kritik - insbesondere aufgrund von Inkonsistenzen zwischen verschiedenen Anbietern, mangelnder Transparenz und Bedenken hinsichtlich Greenwashing. Daher wächst das Interesse an transparenteren und datengetriebenen Ansätzen zur Bewertung von Nachhaltigkeit.

Diese Dissertation untersucht den Einsatz von NLP zur Extraktion nachhaltigkeitsrelevanter Informationen aus alternativen Datenquellen wie Finanznachrichten. Analog zur allgemeinen Entwicklung im Bereich NLP analysieren wir verschiedene Modellarchitekturen, darunter encoder-basierte Repräsentationsmodelle und decoder-basierte generative Modelle, um relevante ESG-bezogene Inhalte aus unstrukturierten Texten zu extrahieren. Unser Ziel ist es, differenzierte, besser nachvollziehbare und flexiblere Nachhaltigkeitsbewertungen zu ermöglichen.

Mit der zunehmenden Nutzung datengetriebener Erkenntnisse für finanzielle Entscheidungen hat auch die Einbeziehung von Investorensentiment in diesen Prozess erheblich an Bedeutung gewonnen – sowohl in der Wissenschaft als auch in der Praxis. Studien zeigen konsistent, dass Sentiment die Portfolio-Performance verbessern kann. Infolgedessen haben große Finanzindexanbieter stimmungsbasierte Indizes eingeführt. Die wissenschaftliche Forschung zu deren Konstruktion bislang jedoch begrenzt. Als ersten Schritt zur Integration ESG-bezogener Stimmungen in Investitionsprozesse entwickeln wir zunächst ein stimmungsbasiertes Pendant zum DAX-40-Index, einem Aktienindex welcher die 40 größten deutschen Firmen beinhaltet.

Aufbauend auf dieser Grundlage untersuchen wir anschließend, wie Stimmungen zu Nachhaltigkeitsthemen gezielt in die Portfolio-Optimierung eingebunden werden können. Da sich die finanzielle Relevanz einzelner ESG-Aspekte im Zeitverlauf verändern kann, schlagen wir einen dynamischen Ansatz vor, der sich an die sich wandelnde Mate-

rialität dieser Faktoren anpassen kann. Diese Arbeit schließt eine Lücke in der Literatur an der Schnittstelle zwischen nachhaltigen Finanzanlagen und Sentimentanalyse.

CONTENTS

1	Introduction	1
1.1	The Financial Domain	2
1.2	Contributions	3
1.3	Thesis Structure	3
2	Financial Information Extraction	5
2.1	Generative or Representational Language Models?	6
2.2	Parameter-Efficient Finetuning	8
2.3	Financial NLP	10
3	Financial Report Extraction	17
4	Modular Learning for Financial IR	25
4.1	News Genre Classification	26
4.2	ESG Key Issue Identification	34
4.3	ESG Impact Determination	41
5	LLMs for Financial IR	47
6	From Sentiments to Asset Representations	67
6.1	Building a BERT-based Sentiment Performance Index	69
6.2	ESG-Investing through Aspect Regimes with Dynamic Time Warping	74
7	Conclusion	83
7.1	Summary	83
7.2	Applicability Beyond the Financial Domain	84
7.3	Future Work	85
8	Publications	87
	Bibliography	89

1

INTRODUCTION

During the work on this thesis over the last few years, the field of natural language processing (NLP) has been fundamentally transformed by the rise of large, generative language models (GLMs). Since the release of GPT-3.5 in late 2022 [105], working with text using these models has become incredibly straightforward - even for non-experts. Tasks that once required carefully engineered pipelines and domain-specific resources can now often be solved by prompting a general-purpose language model [50, 79]. This recent shift in paradigm raises an important question: Is it still necessary to create specialized models or can single, large language models simply solve problems across different tasks and domains? What is the correct approach for extracting relevant information from textual data?

A glance at the Hugging Face model hub¹, a central platform for sharing machine learning models, reveals the scale of ongoing specialization. At the point of writing this thesis, more than 1.8 million models have been uploaded, many of which are tailored for specific tasks, domains, or datasets. This diversity illustrates a crucial point: while generalist models are powerful, they might not always be the best tool for every problem. Practitioners continue to train and fine-tune models that are purpose-built, whether to solve niche tasks or operate effectively within a specific domain.

The field of Information Extraction (IE), which focuses on identifying and extracting structured information from unstructured text, has been profoundly reshaped by the capabilities of GLMs. Traditionally, IE relied on rule-based systems or task-specific models trained to extract entities, relations, or events, often requiring extensive labeled datasets and domain expertise [128, 82]. GLMs, however, have introduced a paradigm shift by enabling flexible, prompt-based extraction that can adapt to diverse tasks without the need for task-specific training. Models such as GPT-4 [106] or LLaMA [150] have introduced a complementary approach, enabling flexible, prompt-based extraction that can adapt to diverse IE tasks without extensive retraining [167]. GLMs can perform diverse IE tasks through carefully crafted prompts, often achieving performance

¹<https://huggingface.co/models>

comparable to that of specialized systems. However, GLMs face challenges, including inconsistent structured outputs and potential hallucinations, particularly in niche domains where fine-tuned models benefit from tailored training [29, 67]. Conversely, fine-tuning requires high-quality labeled data and computational resources, which may limit accessibility. The case of IE thus highlights a critical trade-off: while LLMs offer versatility and ease of use, fine-tuned models provide precision and reliability in specific contexts.

In this thesis, we investigate the use of both types of language models on different Information Extraction tasks within the financial domain. The following chapters trace an arc in recent NLP developments: from training compact, supervised models to leveraging generative, larger language models with emerging techniques aimed at reducing costs - both computational and monetary - while retaining or improving performance. At the same time, parts of this thesis are application-oriented, aiming to develop tools or products that can be utilized in practical settings.

1.1 The Financial Domain

The stock market represents one of the largest and most dynamic social experiments ever created. Every day, millions of investors express their views, deliberately or unconsciously, on the value of assets by buying and selling them. These decisions reflect a diverse range of influences: from geopolitical shifts such as the war in Ukraine² to smaller developments like dying cacao trees that send commodity prices soaring³.

Financial decisions are deeply tied to opinion and interpretation, making finance a natural application area for NLP. Markets are driven by a multitude of signals - some structured, but many unstructured and deeply contextual. By analyzing the textual data generated daily across news outlets, financial filings, and social media, we can gain insights into investor sentiment, emerging risks, and market-moving events.

At the same time, finance, like NLP, has undergone a paradigm shift. The introduction of large, general-purpose language models has significantly simplified the processing of vast amounts of unstructured financial text. This advancement has sparked renewed interest in quantitative finance, where data-driven decision-making has become increasingly prevalent [80, 163, 171, 170].

Another important trend shaping modern finance is the growing emphasis on sustainability. Particularly in Europe, sustainable finance has gained substantial momentum, with more and more capital flowing into companies perceived as socially and environmentally responsible, although a few distractions have stalled this trend somewhat recently [97]. For many investors, financial performance is no longer the only concern - ethical considerations and personal values also guide their decisions. But this raises new challenges: how can we evaluate whether a company is genuinely sustainable? What data should we rely on? How can we measure alignment between a company's actions and its stated values?

In this dissertation, we address both of these threads. We explore how textual data, such as news articles, can inform general investment strategies, and we investigate how NLP techniques can help assess the sustainability practices of companies. Together,

²https://en.wikipedia.org/wiki/Russian_invasion_of_Ukraine

³<https://unctad.org/news/chocolate-price-hikes-bittersweet-reason-care-about-climate-change>

these studies showcase the recent evolution of the field of financial NLP and the resulting broader changes in how financial decisions are made in the age of language models.

1.2 Contributions

This thesis consists of several works which contribute to the field of financial NLP. Most of these works extract relevant information from news articles. To achieve this, we investigate both smaller, modular models using supervised learning, as well as larger, more general models without further training. At the same time, we also leverage the extracted information in order to investigate how to make more informed investment decisions.

In the following, we briefly list the most significant academic studies:

- First and foremost, in order to be able to work with textual data, it needs to be accessible. In [chapter 3](#), we create a tool which enables practitioners to work with company reports in the SEC’s EDGAR database. To make it easier to analyze the reports, we not only provide a crawler but also clean the data and export it into a better structured file.
- In [chapter 4](#), we investigate leveraging modular learning approaches to extract different properties of textual data via supervised learning. We experiment with various setups, using mainly bottleneck adapters in order to handle multilingual datasets and enrich representations through knowledge composition.
- To advance research on corporate sustainability event detection, we construct a dataset in [chapter 5](#) by analyzing news articles and identifying various properties pertinent to ESG decision-making, such as ESG-sentiment, ESG-aspects and ESG-based summaries of the articles. As opposed to the previous studies using supervised learning, we approach the problem using zero-shot extraction with an LLM in this work and evaluate a subset of the results with the help of human annotators. The constructed dataset presents one of the first open datasets with a time-series of ESG-sentiment as well as -aspect data.
- In [section 6.1](#), we investigate how sentiments mined from news articles can be used to construct a sentiment-driven financial index. A product like this adjusts its weight based on the investor opinion of different companies, as opposed to weighting them based on their size.
- Finally, we use the previously constructed ESG-dataset in [section 6.2](#) in order to detect points in time at which certain ESG-aspects are particularly likely to be relevant for different companies and explore how this information could be used for an ESG-based investment approach.

1.3 Thesis Structure

We start by briefly introducing several techniques from the field of Artificial Intelligence (AI) in [chapter 2](#). In [chapter 3](#), we present `edgar-crawler`, an open source

library which can download information from the United States Securities and Exchange Commission’s (SEC) EDGAR (Electronic Data Gathering, Analysis, and Retrieval) database. In addition, `edgar-crawler` also cleans and extracts the relevant information from the reports into a convenient `.json` format.

Next, we introduce several methods for identifying relevant information within news articles and other types of textual data.

In [chapter 4](#), we explore using modular learning approaches for extracting information from textual data. Here, we follow a supervised learning approach, where we experiment with various ways to use annotated datasets in order to train a model for three different use cases. Since all datasets are multilingual, a modular approach is beneficial for capturing knowledge in each language separately, combining it where it suits.

In contrast, in [chapter 5](#), we investigate a different approach using generative Large Language Models (LLMs) to extract information from news data without relying on annotated data. Here, we focus on the field of corporate sustainability, also called ESG (for Environmental, Social, Governance). As we validate our results by conducting an evaluation of the extracted data, this study also provides more clarity about the suitability of LLMs as annotators.

Following that, we apply the previous learnings to build two different approaches for portfolio optimization in [chapter 6](#). Our findings demonstrate that the extracted investor sentiments can improve financial performance, serving as reliable predictors of stock prices for the companies analyzed.

Finally, we conclude the thesis with [chapter 7](#), where we draw our conclusions and provide an outlook for future research directions.

2

FINANCIAL INFORMATION EXTRACTION

Information Extraction (IE) in Natural Language Processing is the automated process of identifying and extracting specific pieces of factual information from unstructured or semi-structured text [128]. This involves pinpointing and retrieving key elements such as named entities (e.g., people, organizations, locations), their attributes, the relationships between them, and the events in which they participate [55]. The overarching goal of IE is to transform the vast amounts of text available into a structured, organized, and machine-readable format [172].

The majority of the papers that make up this thesis apply IE in one way or another. Most of them aim to extract different opinions towards corporate entities from text. This sub-domain of IE is called Aspect-Based Sentiment Analysis (ABSA). However, we are not only interested in the sentiments expressed toward specific entities but also in understanding the sentiment related to a predefined topic.

One example of this is the sentiment toward corporate entities concerning sustainability. There are instances where actions may benefit a company's financial prospects but reflect poor sustainability practices, such as workforce layoffs¹.

In this chapter we will first separate the approaches for extracting information into two distinct fields: using encoder-like Language Models which create rich embeddings and using generative, decoder-like Language Models. We explore the advantages and disadvantages of both model types and summarize recent research developments.

Next, we introduce Parameter-Efficient Fine-Tuning (PEFT), a technique which we apply in several works in later chapters. PEFT allows for selective tuning of a model's weights in order to align it with a specific target task. We specifically focus on bottleneck adapters, a modular solution that allows the tuned parameters to be easily added to or removed from a base model as needed.

Finally, we summarize the general state of Financial NLP, first diving into Financial Sentiment Analysis (FSA) and breaking down recent research directions. We continue

¹Several cases of exactly this could be observed in the last few years: many big technology companies laid off significant parts of their workforce before their stock price rallied.

by exploring FSA applications, where sentiments are used as inputs for investment decisions. Finally, we give a general overview of the field of corporate sustainability.

2.1 Generative or Representational Language Models?

Since Vaswani et al. [154] introduced the transformer architecture in 2017, the majority of models building upon this technology have leveraged one of its two components in order to try to solve IE tasks². The encoder component of the transformer creates rich, contextual vector representations of an input, which can then be used to train further models depending on the desired output. In contrast, the decoder component takes an input and generates further text from it. In this section we provide a general overview of the advantages and disadvantages of the components.

2.1.1 Representational Language Models

Language Models which leverage the encoder part of the transformer are mostly based on the work of Devlin et al. [31] which introduced a model called BERT (Bidirectional Encoder Representations from Transformers). As indicated by the name, they use a bidirectional attention mechanism which looks at the whole input at once and thus can judge the context of each word while considering both preceding and succeeding words.

BERT-based models use Masked Language Modelling (MLM) as a pre-training method. Here, a certain percentage of tokens are “masked” (usually around 15%), meaning they are replaced by a special [MASK] token. Following that, the model is trained by attempting to predict the original token using a version of the cross-entropy loss-function [45, 92]. A secondary pre-training method is the Next-Sentence Prediction task, which aims to predict if two sentences were originally next to one another or not.

Following this first pre-training, the model can then be fine-tuned on the desired downstream task (e.g. Named Entity Recognition, Sentiment Analysis, ...) using supervised learning. This second step is significantly less training-intensive than the pre-training stage, so it is relatively simple to adapt the base model to a specific task.

Since BERT’s release in 2018, significant improvements have been made to the base method, leading to enhanced performance across various tasks. Models like RoBERTa [83] and DistilBERT [127] have made changes to the base model to enhance its capabilities and efficiency. Recently, ModernBERT [161] has taken a significant leap by utilizing rotary positional embeddings (RoPE) [142] to extend the context length from 512 tokens to 8192 tokens among other improvements. However, in general the paradigm in encoder-only models has not changed drastically from the first implementation [31].

2.1.2 Generative Language Models

Decoder-only Language Models can be traced back to the first GPT (Generative Pre-Trained Transformer) model, which was introduced by Radford et al. [119] in 2018. As

²For completion we would like to mention that models like T5 [121] leverage both components, but we will not explore them further in this thesis.

BERT, GPT was designed to be trained in two stages: an unsupervised pre-training stage followed by a supervised fine-tuning stage. However, unlike BERT, GPT uses a unidirectional attention mechanism as it aims to solve tasks by generating new tokens. For this, GPT was trained using Causal Language Modelling (CLM), which basically aims to predict the next token in a sequence by trying to minimize its Log-Likelihood in the loss function [45]. As with BERT, GPT can then be adapted to specific tasks by fine-tuning on a labeled dataset.

Similar to encoder-only models, the GPT series of models has been developed extensively over the recent years. But where BERT-like models primarily improved on the first iteration in smaller steps, occasionally making bigger jumps when a new technical improvement was found, GPT-like models have changed drastically since their initial inception.

During the development of GPT-2 [120] and GPT-3 [16], the authors gradually tested replacing the fine-tuning stage with a simple few-shot learning paradigm, in which the weights of the model are not updated anymore. By supplying a few samples to the model prior to asking for the answer of a sample, the model can adapt to different tasks. This shift to essentially an all purpose model that can be used for any task continued with InstructGPT [108], which leverages a technique called “Reinforcement Learning from Human Feedback” (RLHF). As the name suggests, RLHF further refines a model by applying Reinforcement Learning (RL) to teach the model to prefer generating answers which are more helpful, based on a human-annotated dataset. This trend continued with ChatGPT [105] and subsequent models like GPT-4 [106] and GPT-4.5 [107], which also added multimodality to their capabilities. At the same time, a large number of open source models were released as well, for example the Llama family of models [150, 151, 147] and the Qwen family [12, 169, 118, 148]. These models mostly scale by being larger and being trained on more high-quality data with more compute [68, 54].

2.1.3 Adapting to New Domains

Both of the approaches discussed above come with their own up- and downsides, especially when they are applied in particular domains (such as finance) or languages not included in their pretraining. First, we will look at the representational models discussed in [subsection 2.1.1](#).

Here, the fine-tuning part, which uses supervised learning, can present unique challenges [35, 22]. Annotated data in these areas is often sparse and difficult to obtain due to the need for domain expertise and the complexity of financial texts. To overcome those challenges, financial experts are needed to obtain high-quality annotations, which is expensive and time-intensive.

If the domain-specific task is multilingual at the same time, the problem is complicated further. Since it can be challenging to obtain domain-specific annotated data in a single language, doing so for multiple different languages is even more difficult. However, several approaches have focused on facilitating the knowledge transfer of fine-tuned models between the different languages. We will expand on these approaches in [section 2.2](#).

There are several advantages of using representational models. Firstly, due to the bi-directional training, their contextual knowledge is richer which can result in higher

performances depending on the task [79, 44, 103]. Secondly, compared to modern generative models, they are more efficient and require less computational resources.

On the other side, generative models have the added issue that they can struggle with hallucinations [41, 64], especially in domains insufficiently represented in the pre-training dataset [78, 29]. Another problem of generative models is their dependency on the prompt used to perform a certain task. Although researchers have investigated (and made progress on) how to optimize prompts for performance, the best way to do this is still not clear [179, 156, 123, 81]. This means that it can be very costly (both in time and money) to optimize a prompt for a certain task.

However, modern generative models are very easy to use and to adapt to new tasks. Since they implicitly have some slight reasoning capabilities (which can be enhanced through clever prompting) [162, 177], it is often possible to achieve good performances even in a zero shot setting [79, 167].

At the same time, emerging research has found ways to mitigate hallucinations. If an underlying knowledge base is available, Retrieval Augmented Generation (RAG) [77] can provide documents there to the model so that it focuses on the information there instead of making up new information [136, 17]. This technique has also been found to reduce (but not completely solve) hallucination in specific domains, such as the medical [71], legal [89] and financial [129, 67, 78] domain.

Another approach for mitigating hallucinations is through the improvements of the prompt, such as using chain-of-thought prompting [162, 159], for the medical [71] and legal [74] domains and few-shot prompting for the financial domain [67, 78]. An approach which proved to be less performant for medical tasks [71], but effective for reducing hallucinations for financial tasks is fine-tuning [108] the base-model on domain-specific knowledge [125, 78].

2.2 Parameter-Efficient Finetuning

As mentioned in the previous section, it is possible to leverage pre-trained language models (LMs) which have learned general information during unsupervised training on large corpi by training them further on specific tasks. However, this comes with different challenges:

- As pre-trained LMs become larger, further finetuning becomes more computation heavy
- Continuous training of the same parameters can cause catastrophic forgetting, essentially leading the model to forget previously learned information when trained on new tasks [46]

To alleviate these points, parameter-efficient finetuning (PEFT) only adjusts a small set of parameters, either new or existing, while keeping the majority of them frozen [174]. Not only is this computationally more effective while achieving the same performances as full finetuning [168], but it also avoids catastrophic forgetting as the previous knowledge encoded in the LM is left mostly unchanged [57, 174].

A variety of different PEFT approaches have been studied [168, 174], however, in this thesis we will focus on bottleneck adapters [57], which are small neural network

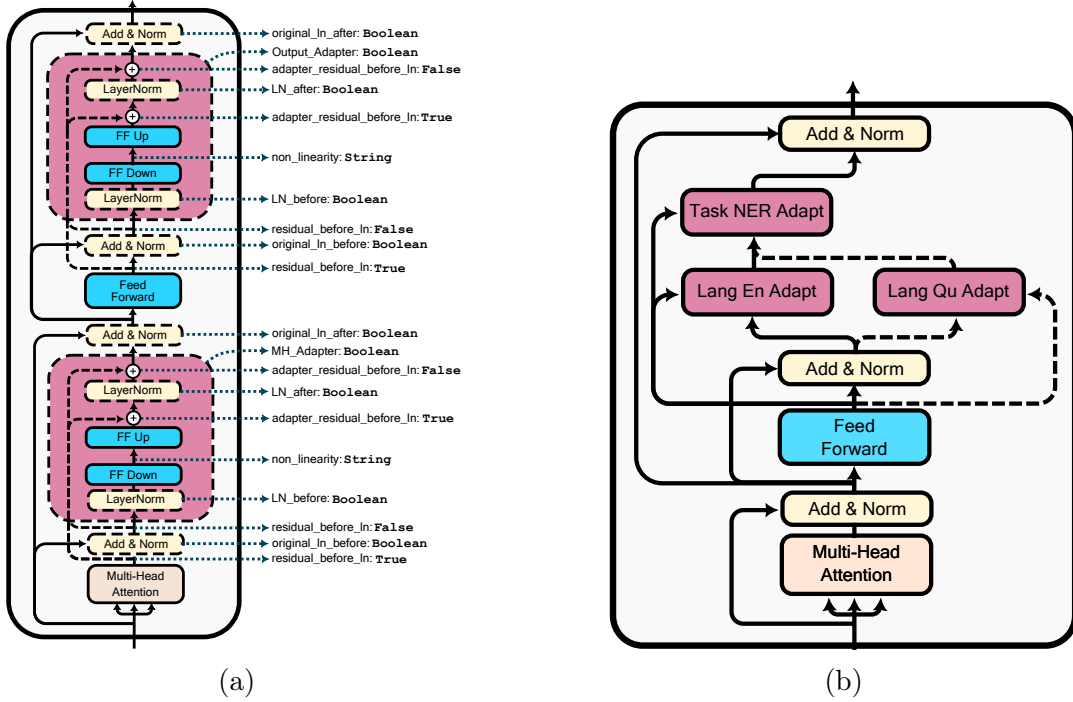


Figure 2.1: (a) Configuration options for bottleneck adapters. Dashed lines represent the different options for new weights (both up- & down-projections and LayerNorms) and residual connections. New weights are illustrated within pink boxes. Schema from [113]. (b) Transfer of knowledge learned in NER task in language EN to a different language (shown by the dashed lines). Schema from [114].

(NN) modules inserted within the layers of pre-trained transformer-based models. In order to improve the efficiency, the hidden states of the layer are first projected into a lower dimension using a matrix W_{down} , where a non-linearity f is applied. Finally, the states are projected back into the original dimension of the hidden layer with a second matrix W_{up} and combined with a residual connection r :

$$h \leftarrow W_{up} \cdot f(W_{down} \cdot h) + r$$

The adapters can be introduced at different locations within the transformer block, as shown in Figure 2.1a. Since the adapter-weights are fully separate from the original weights, they are modular and as such can be combined in interesting ways. An example for this is presented by Pfeiffer et al. [114], where the authors use both language-adapters created by training adapters in a specific language using the fill-mask task as well as other adapters. As illustrated by the schema in Figure 2.1b for the example of English and Quechua, combining a language adapter with a task adapter trained on data in language A (English in the example) allows for a transfer of the knowledge to a second language B (Quechua in this case), for which a training dataset might not exist.

Knowledge Composition with Adapters

The avoidance of catastrophic forgetting through adapters is a big advantage. However, combining the knowledge trained into the separate modules is not possible through the standard setup. In practice, the downstream task which uses the adapters is only able to employ one adapter at a time.

To enable the composition of modular knowledge present in the different adapters, Pfeiffer et al. [112] introduce AdapterFusion. Having trained the different adapters, AdapterFusion employs a secondary stage of training to learn how to apply combinations of the different adapters for a specific task. This is done by introducing a third set of parameters, Ψ , in addition to the parameters of the different adapters Φ as well as the base model parameters Θ . Then, the AdapterFusion parameters for a task m are determined by minimizing the loss function L_m when training on the task-data D_m :

$$\Psi_m \leftarrow \arg \min_{\Psi} L_m(D_m; \Theta, \Phi_1, \dots, \Phi_N, \Psi)$$

This means that the training dataset of m is used two times: one time for training the task adapters, and a second time for training the AdapterFusion.

In order to teach the new parameters how to compose the outputs from the different adapters, the original authors [112] define *Key* (K_l), *Value* (V_l) and *Query* (Q_l) matrices as parts of Ψ at each layer l . For the *key* and *value* transformations, the output $z_{t,l}$ of each adapter is used at each time-step t . For the *query*, the output of the feed-forward sub-layer is used. Finally, the activation of each adapter n based on the context is learned using

$$\begin{aligned} s_{l,t} &= \text{softmax}(h_{l,t}^\top Q_l \otimes z_{l,t,n}^\top K_l) \\ z'_{l,t,n} &= z_{l,t,n}^\top V_l \\ Z'_{l,t} &= [z_{l,t,0}, \dots, z_{l,t,N}] \\ o_{l,t} &= s_{l,t}^\top Z'_{l,t} \end{aligned}$$

The final output $o_{l,t}$ is then essentially the sum of the transformed adapter outputs weighted by the attention scores calculated with the softmax function.

Using AdapterFusion to apply learned knowledge from different tasks helps to improve results especially for tasks with little training data, as the model can leverage the knowledge gained from tasks with more training data [112]. However, AdapterFusion is a very versatile technique, and as we will see in [chapter 4](#), using it to compose knowledge learned from the same task in different languages can be quite beneficial as well.

2.3 Financial NLP

In this section we will outline the current state of financial NLP as well as key opportunities and challenges. We will then focus on financial sentiment analysis in particular, investigate how it is used for concrete investment decisions and finish with the topic of

corporate sustainability.

The role of NLP for the financial sector has rapidly increased over recent years. Chen et al. [22] note the rise in Google Trends for “FinTech” prior to 2020, along with the growing number of finance-related workshops collocated with NLP-related top-tier venues. This trend has only continued since then, as Du et al. [35] observe in the consistent growth of the number of NLP-related publications for different financial research topics. While topics such as “Risk management”, “Question Answering” and “Financial Sentiment Analysis” lead in the number of publications, others, such as “ESG and sustainable finance”, are comparatively under-researched but increasingly drawing attention.

Du et al. [35] separate Financial Textual data into three different fields: “Corporate Data” such as annual reports and other corporate disclosures, “Public News and Announcement”, which contains financial news and analyst reports, and finally “Social Media Data” from sources such as X (formerly Twitter), StockTwits and more.

While chapter 3 concerns itself mainly with the first type of data, the other chapters will overwhelmingly focus on the second source, public news.

2.3.1 Financial Sentiment Analysis

Financial Sentiment Analysis (FSA) is, as mentioned above, one of the topics in the Financial NLP domain which has garnered the most attention (and most publications) in recent years [35]. In general terms, FSA is concerned with determining investor sentiment (which contains beliefs about future investment risks) and textual sentiment (which considers how positive or negative a text is) [69]. Du et al. [34] divide FSA into two principal research tasks: Technique-driven FSA investigates determining the sentiment of financial texts, while application-driven FSA leverages the sentiments and studies how to use them in order to make informed investment decisions. We will start by expanding on the first research task and then follow up on the second one in the next subsection.

Predicting the sentiment of financial texts can be approached in different ways and has changed along with the entire field of NLP over the years. Earlier works attempted to use lexicon-based methods, which assign polarities to single words or phrases [49, 152, 144, 87, 104]. Those polarities can then be combined in different ways in a secondary step for interpretation [149], for example by counting positive or negative words in the text [60].

Following that, a more sophisticated way to determine the sentiment is to first extract numerical features from a text, and then train a machine learning (ML) model for sentiment prediction [122]. Feature extraction can happen through different methods which extract semantic information from the text, for example by using the above mentioned lexica. Other approaches extract n-grams [30], verbs or named entities [65]. Notably, a very popular technique for extracting features is Word2Vec, which is able to create dense and continuous vectors from text [95]. Word2Vec models can be trained in an unsupervised manner and capture the semantic information of words efficiently [96]. Once the Word2Vec model is available, the text can then be transformed to vectors, and a secondary model can be trained to predict the sentiment from these vector representations [135].

More modern approaches leverage the pre-trained language models we discussed in [section 2.1](#). Models like BERT [31] can be directly fine-tuned on a financial dataset to extract the sentiments [85]. Other approaches have continued pre-training BERT on financial data in order to increase performance on financial tasks. Araci et al. [10] for example have created a model they call FinBERT by pre-training further on the TRC-2 corpus by Thomson Reuters [25]. Seemingly simultaneously, Liu et al. [85] developed a different model using the same name, FinBERT. However, in this case the model is pre-trained from scratch on different datasets. Both models report an outperformance compared to simply fine-tuning a BERT-model on a financial task.

Finally, recent research [79] has conducted experiments on the performance of generative language models compared to the previously mentioned pre-trained, BERT-based language models. When using these generative models, the samples are simply fed to the model with an appropriate prompt. As we discussed above, prompting is not trivial and can make a huge difference in the final performance. By, for example, feeding examples to the model (this is called few-shot prompting [16]), the performance is generally improved [79].

It is noteworthy that although there is a lot of research on FSA - as discussed in this section - most of it builds upon three core datasets: The financial phrasebank dataset by Malo et al. [91] contains around 4840 sentences from financial news in the English language, which are categorized into negative, neutral or positive polarities. The SemEval-2017 Task 5 dataset [26] consists of several tracks which concern themselves with messages from microblogs such as StockTwits (1847 messages) and Twitter (1591 messages), as well as sentences from news headlines and texts (1780 sentences). Here, the samples are annotated with scores between -1 (very negative/bearish) to $+1$ (very positive/bullish). The Financial Opinion Mining and Question Answering (FiQA) dataset [90], and specifically the first task, takes a similar approach as the previous dataset but expands it to an Aspect-based Sentiment Analysis (ABSA) problem of 529 news headlines and 774 microblog messages.

A takeaway from this is that openly available, annotated data for FSA is still rather sparse. We will investigate solutions for improving the accessibility of financial data and even contribute with our own dataset over the course of this thesis.

2.3.2 FSA for Investment Decisions

Once the sentiments have been determined, using them in order to improve investment decisions is an obvious second step. The requirement for this to work is that there is a correlation between sentiments and future returns in the first place. We will first expand upon this in the first part of this subsection. If there is a correlation, the next logical step would be to try and predict the future prices. Finally, in reality, investors do not only buy into the single stock with the highest predicted performance, but they buy multiple stocks which are expected to perform well in the future - we will investigate using FSA for portfolio management in the last part of this subsection.

Causality

Several works have investigated the interdependency between news sentiments and stock prices. Wang et al. [158] identify relevant words that carry sentiment in financial

reports of US companies and finds a correlation between the occurrence of these words and future stock price volatility. Cosley et al. [27] instead focus on the platforms StockTwits (more of a social media platform) and Seeking Alpha (contains expert opinions) - they find that while content from Seeking Alpha correlates more with the stock performance than StockTwits, this correlation is small. However, by applying well defined filters to extract higher quality content, they could increase this correlation significantly. On the other hand, Hwang et al. [60] collect articles about North Korean stocks and Tesla and comes to the conclusion that stock prices influence the sentiment more than the inverse case. However, they only investigate data over a short time-frame - three months. Ranco et al. [122] investigate detecting sentiment “events” on Twitter and finds that the aggregate sentiment during events implies the direction of the future market evolution. They note that the data has to be carefully selected to observe this correlation. More recently, Wan et al. [157] similarly detect events with large changes in sentiment and finds that they not only induce abnormal return and volatility movements for the target company, but also for related companies from the same sector or group.

Direct Prediction

Stock price prediction has been studied extensively, mostly through using historical stock price data along with technical indicators and other economic variables [110]. However, with the increasing knowledge about the correlation between investor sentiment and the stock market, researchers started to include sentiment information into their features [34]. Ding et al. [33] train a Convolutional Neural Network (CNN) [131] in order to detect short-term and long-term influences of events detected in financial news titles. Sidogi et al. [137] train multiple LSTM models [53] in order to predict the stock price and find that using sentiments extracted with the FinBERT model [10] can best predict the future price of the US company Amazon. Ho et al. [52] train a CNN to predict the trend of the stock price development of five different US-based technology stocks over different time-frames and found that combining the sentiment with the price data performed best. Ruan et al. [126] incorporate sentiments mined from certain forums on Reddit³ into an LSTM architecture in order to predict the stock price in periods of anomalous returns. Kaeley et al. [66] develop a Transformer-based model and leverage both technical and sentiment data in order to predict the price development over different time periods. They find that their approach outperforms an LSTM baseline for long-term dependencies between sentiment and price.

Portfolio Management

Transforming stock price predictions to an optimal portfolio composition adds another layer of complexity to the previous step and has been addressed in different ways. Historically, modern portfolio theory (MPT) started with Harry Markowitz in 1952 [93], where he introduced the idea that a portfolio can be optimized based on two objectives: risk and expected return. Put briefly (more details can be found in Palomar et al. [109]), what Markowitz described is a bi-objective optimization problem where the expected return and the risk (via the volatility or variance) of the portfolio are the

³<https://www.reddit.com/>

two objectives. With the expected return being the weighted sum of expected returns of the singular portfolio constituents $w^T\mu$ and the risk being expressed through the variance $w^T\Phi w$, the optimization problem can be formulated as a weighted sum:

$$\max_w w^T\mu - \frac{\lambda}{2}w^T\Phi w, \quad (2.1)$$

where Φ is the covariance matrix of the singular returns and λ is a factor which expresses how risk-averse the portfolio aims to be. At the same time, a set of constraints can be introduced, as is often the case in reality. An example for this is the “5-10-40-rule” of managed funds in Germany, which says that stocks with weights above 5% must not surpass 40% in sum, all while a single stock should not have a weight above 10%⁴. Depending on the legal framework the investor has to work with, different kinds of constraints need to be adhered to while optimizing for returns.

Different studies have used the above mentioned mean-variance approach in order to build on top of their price-predictions. Muthivhi et al. [100] first predict the prices with an LSTM, before doing exactly this. They find that including sentiments in the LSTM prediction yields a superior performance compared to different sentiment-unaware strategies.

With the rise of bigger and more complicated models, the field of portfolio optimization has transformed as well. Modern approaches, for example by Koratamaddi et al. [72], increasingly leverage reinforcement learning (RL) [143] to train a model to choose optimal allocation based on different input features. Notably, recent studies have investigated using the financial sentiment as one of those features. The above mentioned work ([72]) find that the sentiment-aware RL-agent effectively maximizes return and minimizes risk and outperforms other sentiment-unaware baselines, including a Mean-Variance baseline. Other publications such as Liu et al. [84] include modern architectures such as the transformer into an RL-framework in order to combine market-, macro- and sentiment-features for portfolio optimization. Their ablation study shows that combining all features outperforms other variants with missing features. In particular, while the experiment excluding the sentiment achieve an only slightly lower performance, a sharp decline in sharpe ratio [134] shows that the investor sentiment plays a significant role in stabilizing the portfolio.

2.3.3 Corporate Sustainability

Corporate sustainability (CS), often framed through Environmental, Social, and Governance (ESG) criteria, is a framework that evaluates how businesses operate responsibly in relation to the environment, society, and their internal governance. ESG assesses a company’s impact on areas like climate change, employee welfare, and ethical leadership. The Environmental aspect focuses on reducing ecological harm, such as lowering carbon emissions or managing waste. The Social component examines how a company treats its employees, customers, and communities, emphasizing fairness, diversity, and human rights. Governance covers leadership practices, including transparency, anti-corruption measures, and ethical decision-making. Together, these elements guide companies toward sustainable practices that balance profit with positive societal impact.

⁴https://www.gesetze-im-internet.de/kagb/_206.html

ESG is important because it aligns business strategies with long-term societal and environmental goals, benefiting both companies and their stakeholders. For companies, adopting ESG practices can have monetary motivations. Research shows that strong ESG performance often correlates with better financial outcomes, such as higher profitability and lower risk [21]. For example, companies with robust ESG practices may attract more investors, reduce operational costs through energy efficiency, and build stronger customer loyalty by aligning with public values like sustainability. A 2015 meta-analysis of over 2,000 studies found that ESG-focused companies often achieve higher equity returns and reduced financial risks, proving that doing good can also be good for business [42]. Additionally, ESG helps companies maintain their “social license” to operate, earning trust from communities and regulators, which can lead to easier market access and growth opportunities [94].

In order to quantify the ESG practices of companies, stakeholders often rely on ESG-ratings provided by rating agencies. Those agencies analyze large amounts of both company-provided and public information in order to identify how sustainable a company is. One blog post [39] counts more than 600 different rating agencies near the end of 2023. However, most consumers rely on a small few agencies such as MSCI⁵, Sustainalytics⁶ or LSEG⁷ - all with slightly different approaches [37].

Despite its benefits, the current popular approach of ESG ratings faces several challenges. One major issue is the inconsistency of the ratings, as different rating agencies use varied methodologies, leading to discrepancies in how companies are evaluated [13]. For example, a firm might score highly on one agency’s environmental metrics but poorly on another’s due to differing criteria. This lack of standardization confuses investors and undermines trust. Additionally, skepticism about ESG persists, with critics arguing that some companies engage in green-washing, using ESG as a marketing tool without meaningful action [101]. Others question whether ESG truly delivers financial returns or if it distracts from core business goals [117], although novel research in this area suggests the opposite [21].

The advances in NLP described in the previous sections have also started to impact and accelerate research in the ESG domain. Using novel models and tools, it is possible to analyze vast amounts of textual data, such as corporate reports and news articles, to assess companies’ sustainability performance [102, 35]. They can help to identify patterns in how companies disclose ESG-related information, detect greenwashing (exaggerating environmental claims), and evaluate stakeholder sentiment independently from ESG-rating agencies. This would provide more transparency in the sector and thus shed light on much of the ESG-related skepticism described.

⁵<https://www.msci.com/data-and-analytics/sustainability-solutions/esg-ratings>

⁶<https://www.sustainalytics.com/>

⁷<https://www.lseg.com/en/data-analytics/sustainable-finance/esg-scores>

3

FINANCIAL REPORT EXTRACTION

As we mentioned in [subsection 2.3.1](#), data in financial NLP is sparse. For this reason, enabling access for more financial data plays an important role in advancing the field of financial NLP [34]. A popular data source which is publicly available but not used to its full potential due to several limiting factors is the EDGAR database of the United States Securities and Exchange Commission (SEC)¹. Here, American companies are required by law to upload different kinds of reports - the most important ones being yearly reports (10-Ks), quarterly reports (10-Qs) and current reports (8-Ks). These reports present a very valuable resource for financial NLP, but different issues make working with them in their raw state challenging:

1. In order to download reports, users have to visit the different company sites in EDGAR and find the reports they are interested in.
2. Reports are often very lengthy and contain numerous artifacts that make them difficult to process.
3. Over the years, the format of the reports has changed, making a unified tool for processing the reports more complex.

The company reports available on the EDGAR database are important for different reasons.

10-K (yearly) reports The 10-K is an annual report required by the SEC for publicly traded companies. It is filed within 60 to 90 days after the fiscal year-end and includes audited financial statements, management’s discussion and analysis (MD&A) as well as disclosures about business operations, legal proceedings and executive compensation.

10-Ks are an important resource both for investors and for researchers as they provide key information for evaluating a company’s long-term viability. For investors,

¹<https://www.sec.gov/edgar/search/>

the reports are a tool for informed decision-making while for researchers, they represent a rich dataset for empirical analysis, offering insights into corporate behavior, financial trends, and industry dynamics.

10-Q (quarterly) reports 10-Q reports are filed within 40 to 45 days after the end of each of the first three fiscal quarters (the 10-K being filed after the last one). They provide an update on financial performance and operations, including unaudited financial statements, MD&A, and disclosures about material changes since the last 10-K or 10-Q.

The 10-Q’s frequency and focus on recent performance make it essential for tracking ongoing developments. They fill the gaps between the more sparse yearly reports and can thus provide early warning signals, seasonal context and information on short-term trends.

8-K (current) reports 8-K reports are filed typically within four business days of a significant event that shareholders should know about. They can cover for example mergers, bankruptcies, lawsuits or executive changes. These reports are ideal for event studies as they provide precise timestamps for significant corporate actions.

Notably, the financial reports are also important regarding ESG practices, as they often contain information about the sustainability-related strategy of the companies. Ignatov et al. [61] for example finds that the ESG-related tone in 10-K reports correlates with the stock market returns of companies.

In order to provide the public with easier access to this free resource, we developed an open source tool which not only crawls the reports a user wants to download, but also extracts the different parts of the report in a readable `.json` format convenient to work with.

Paper: Lefteris Loukas, Fabian Billert, Manos Fergadiotis, Prodromos Malakasiotis and Ion Androutsopoulos: EDGAR-CRAWLER: From Raw Web Documents to Structured Financial NLP Datasets In *Companion Proceedings of the ACM Web Conference 2025*, pages 761–764, Sydney NSW, Australia. Association for Computing Machinery.

Personal Contribution: Lefteris Loukas developed the first version of the crawler with the help of Manos Fergadiotis and Prodromos Malakasiotis. Fabian Billert extended the project by adding support for more report types, refactored the existing code and made it more stable. Lefteris Loukas wrote the majority of the paper, while Fabian Billert wrote specific sections and advised for the rest. Prof. Ion Androutsopoulos supervised the work.

Status: Published

The reports are downloaded in three steps: First, the user defines which reports for what companies they want to download for which years. Afterwards, the corresponding reports are crawled and downloaded. The challenging task of cleaning and extracting the data is performed in the third step. Since the reports consist of different **Parts** and **Items** (as defined by EDGAR²), we can identify where the different sections start and end, and extract this information. This is valuable because both investors and researchers are often interested in specific sections as opposed to the whole report [18, 59, 9, 133]. Finally, this information is returned as a `.json` file - a convenient format to parse and work with.

`edgar-crawler` has already seen a wide adoption in the field of financial NLP, gathering over 450 stars on Github³, and being used to support a variety of different research projects, such as Crane et al. [28] and Xie et al. [166].

²<https://www.sec.gov/submit-filings/forms-index>

³<https://github.com/lefterisloukas/edgar-crawler>, as of 10.12.2025

EDGAR-CRAWLER: From Raw Web Documents to Structured Financial NLP Datasets

Lefteris Loukas
Department of Informatics, Athens
University of Economics and Business
Athens, Greece

Fabian Billert
Heinrich Heine University of
Düsseldorf
Düsseldorf, Germany

Manos Fergadiotis
Department of Informatics, Athens
University of Economics and Business
Athens, Greece

Prodromos Malakasiotis
Department of Informatics, Athens
University of Economics and Business
Athens, Greece

Ion Androutsopoulos
Department of Informatics, Athens
University of Economics and Business
Archimedes/Athena RC, Greece
Athens, Greece

Abstract

Acquiring data for resource-limited domains like financial NLP presents numerous challenges. One rich web resource of textual financial data is EDGAR, the Electronic Data Gathering, Analysis, and Retrieval system, where U.S. legislation requires publicly traded companies to upload economic reports for investor transparency. However, automating the process of downloading EDGAR filings is not enough. These long-form documents, often exceeding 100 pages each, include noisy text, images, and tables, requiring significant manual preprocessing. Moreover, researchers often require only specific sections of these lengthy documents to solve specific NLP tasks, further complicating the use of these filings. To address these challenges, we present EDGAR-CRAWLER, the first Open-Source Software (OSS) for automatically downloading and preprocessing EDGAR filings. Equipped with a variety of document mining options, EDGAR-CRAWLER can convert lengthy raw financial documents into clean and section-specific JSON files, which can be integrated seamlessly into NLP workflows. Supporting multiple filing types and offering extensive filtering options, EDGAR-CRAWLER facilitates the creation of structured large-scale datasets for financial research. Serving as a free alternative to premium web service data providers, EDGAR-CRAWLER has been widely adopted by both practitioners and academic researchers. Our software is available at <https://github.com/lefterisloukas/edgar-crawler>.

CCS Concepts

• **Information systems** → **World Wide Web**; *Digital libraries and archives*; • **Computing methodologies** → **Natural language processing**.

Keywords

Data Resource, Open-Source Software, EDGAR, Web Archive, NLP

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

WWW Companion '25, April 28-May 2, 2025, Sydney, NSW, Australia

© 2025 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 979-8-4007-1331-6/2025/04

<https://doi.org/10.1145/3701716.3715289>

ACM Reference Format:

Lefteris Loukas, Fabian Billert, Manos Fergadiotis, Prodromos Malakasiotis, and Ion Androutsopoulos. 2025. EDGAR-CRAWLER: From Raw Web Documents to Structured Financial NLP Datasets. In *Companion Proceedings of the ACM Web Conference 2025 (WWW Companion '25)*, April 28-May 2, 2025, Sydney, NSW, Australia. ACM, New York, NY, USA, 4 pages. <https://doi.org/10.1145/3701716.3715289>

1 Introduction

Natural Language Processing (NLP) for economics and finance in particular is a rapidly developing research area, as shown by the new numerous related conferences and workshops [1, 7], as well as the recent creation of a special interest group.¹ Extracting valuable information from financial data is crucial, and while numerical data is often reported in tables, textual data also holds significant insights. One free rich web resource of such textual data is the Electronic Data Gathering, Analysis, and Retrieval system (EDGAR) from the US Securities and Exchange (SEC) website, where publicly traded companies upload various reports describing their financial status and important events regularly, in order to maintain transparency in the open market [14, 18]. EDGAR's development began in 1993, and it is one of the most comprehensive and influential web archives of financial information globally for investors, analysts, and researchers, with 3,000 filings being uploaded there daily.²

Financial documents from EDGAR are useful in numerous machine learning (ML) and NLP applications. These include many tasks such as stock price prediction [15], risk analysis [14], and bankruptcy prediction [9]. Researchers using EDGAR documents often narrow their focus to specific years or company reports to tailor their experiments. Among the various types of EDGAR filings, the most widely used are:

- the **(10-K)** annual reports [10, 14, 17, 18, 20, 21], which provide a comprehensive overview of a company's financial performance, including audited statements, risk factors, management analysis, and operational details.
- the **(10-Q)** quarterly reports [3, 10, 16], which include financial data and updates on a company's financial condition and operations over the past three months.

¹<https://sigfintech.github.io/>

²<https://en.wikipedia.org/wiki/EDGAR>

- and the (8-K) important event reports [5, 15], which are event-driven reports disclosing significant corporate events, such as mergers, leadership changes, stock dilutions, or asset sales.

To use such financial data from EDGAR in their research, NLP practitioners often rely on (paid) curated-dataset providers. Examples include Wharton Research Data Services (WRDS)³ and NASDAQ Premium Data Publishers⁴, which require costly institutional licenses of thousands of dollars, or more affordable (but still paywall-blocked) options like AWS Marketplace⁵. These costs are often prohibitive for budget-limited organizations, especially small research groups. Free alternatives exist, but are limited to simply downloading the raw web documents from EDGAR without curating them (filtering or structuring them), which presents a significant challenge.

The raw financial documents on EDGAR can have different formats depending on the date of origin (.txt or .html), and are complex and lengthy, making them difficult to process. For instance, the 10-K reports average 55,000 words or 110 pages⁶, divided into 23 main item sections (Table 1) [6]. Quarterly reports (10-Q) and important event reports (8-K) contain 11 and 32 sections, respectively. Moreover, in practice, researchers typically focus on only a few specific sections relevant to their experiments (Table 1).⁷ For example, Katsafados et al. [13] combine just Item 7 and Item 1A to detect Initial Public Offering underpricing. Thus, the lack of free and curated financial data forces researchers to manually extract relevant item sections from raw documents found in the EDGAR web repository, a time-consuming and inefficient process that needs to be automated.

In this paper, we release EDGAR-CRAWLER, the first open-source toolkit that both downloads and cleans raw, unstructured documents from financial reports hosted on the EDGAR web repository from the US SEC. Our financial data mining toolkit converts each raw web document into a user-friendly JSON format, where all sections and subsections (called ‘items’) are segmented (Fig. 3). By automating the (until now) manual steps of data preprocessing and curation, EDGAR-CRAWLER not only accelerates the work of financial NLP practitioners, but also enables non-NLP-experts to start their research without relying on costly web data service providers. EDGAR-CRAWLER supports the quick creation of financial data resources and offers many configuration options, such as generating large text datasets for specific years or companies. Our software is free and open-sourced at GitHub: <https://github.com/lefterisloukas/edgar-crawler>

2 Related Libraries and Web Services

User-wise, there are three main requested functionalities when using data from EDGAR: (a) batch crawling documents (multiple years or companies), (b) crawling documents with specific filters such as stock tickers or CIKs (Central Index Key, the EDGAR company identifier), and most importantly (c) mining and extracting clean text from these lengthy and unstructured documents. The last

Table 1: The 23 different items of a financial annual report (10-K document), as uploaded by publicly traded companies in EDGAR’s web repository. These items often cover hundreds of pages and contain noisy text. Researchers and practitioners are mostly interested in few, very specific item sections depending on their task. EDGAR-CRAWLER currently supports mining (downloading and text cleaning) of annual, quarterly, and important event reports.

	Item	Section Name
Part I	Item 1	Business
	Item 1A	Risk Factors
	Item 1B	Unresolved Staff Comments
	Item 1C	Cybersecurity
	Item 2	Properties
	Item 3	Legal Proceedings
Part II	Item 4	Mine Safety Disclosures
	Item 5	Market
	Item 6	Consolidated Financial Data
	Item 7	Management’s Discussion and Analysis
	Item 7A	Quantitative and Qualitative Disclosures about Market Risks
	Item 8	Financial Statements
	Item 9	Changes in and Disagreements With Accountants
	Item 9A	Controls and Procedures
	Item 9B	Other Information
Part III	Item 9C	Disclosure Regarding Foreign Jurisdictions that Prevent Inspections
	Item 10	Directors, Executive Officers and Corporate Governance
	Item 11	Executive Compensation
	Item 12	Security Ownership of Certain Beneficial Owners
	Item 13	Certain Relationships and Related Transactions
	Item 14	Principal Accounting Fees and Services
Part IV	Item 15	Exhibits and Financial Statement Schedules Signatures
	Item 16	Form 10-K Summary

point is critical since, in practice, researchers are only interested in a few pages out of hundreds from the unstructured documents.

Several open-source toolkits (Table 2) have been developed to download financial filings from SEC’s EDGAR, with varying degrees of success. However, none of them supports the standardization and preprocessing of the downloaded web documents (splitting them into specific item sections), as EDGAR-CRAWLER does. We briefly discuss them below.

sec-edgar-downloader allows users to crawl data based on the company’s stock ticker and its CIK. However, it cannot clean or extract any data from the documents, making it less useful for NLP.

sec-edgar, as well as **edgartools**, allow users to crawl data from EDGAR in batches and support filtering. However, they do not support item section extraction, and users must write their own extraction code to integrate the data into their NLP pipelines.

py-edgar allows users to crawl data based on CIKs and does not support the search of companies via stock tickers. It also lacks the ability to parse and extract any data from the documents.

sec-api is a **closed-source** paid service offering batch and filtered crawling. It also recently introduced the ability to parse the documents and produce clean data, like EDGAR-CRAWLER (ours). However, as a paid service, it limits its usability by small business entities or research groups with limited budgets.

All the above libraries and services have limitations (functionality or cost-wise) when it comes to crawling and parsing data from EDGAR. Our library, EDGAR-CRAWLER, stands apart with its ability to (a) crawl in batch, (b) crawl with filters (stock tickers and CIKs), and (c) extract clean text data in a standardized and easy-to-use JSON format, without any cost.

³<https://wrds-www.wharton.upenn.edu/pages/about/data-vendors/>

⁴<https://data.nasdaq.com/publishers>

⁵<https://aws.amazon.com/marketplace/pp/prodview-k65p3l3pdyqfw#offers>

⁶A document of that size cannot fit in the context of most large language models.

⁷www.investopedia.com/articles/basics/10/efficiently-read-annual-report.asp

Table 2: EDGAR-related mining OSS libraries and (closed-source) paid web services, along with their functionalities and costs.

Name	Batch Crawling	Filtered Crawling	Extract Sections	Pricing
sec-edgar-downloader	✗	✓✓	✗	Free
sec-edgar	✓	✓✓	✗	Free
edgartools	✓	✓✓	✗	Free
py-edgar	✗	✓	✗	Free
sec-api.io (<i>Paid web service</i>)	✓	✓✓	✓	Subscription Paywall
EDGAR-CRAWLER (ours)	✓	✓✓	✓	Free

3 EDGAR-CRAWLER Architecture

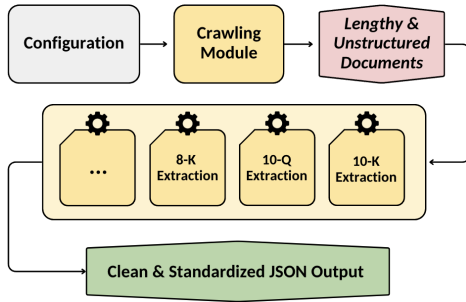


Figure 1: EDGAR-CRAWLER’s architecture. First, the user specifies (Configuration) what data they want (companies, years, filing types) and the software downloads them. Then, the item extraction pipeline extracts the item-specific sections and converts them to a standardized JSON format.

EDGAR-CRAWLER consists of two Python modules:

download_filings.py is responsible for crawling and downloading financial reports for publicly-traded companies. It supports multiple input arguments provided by the user, such as the start and end year, quarters, filing types, and a list of company identifiers to download reports for. We utilize BeautifulSoup4 for the crawling and we set a maximum rate of 10 HTTP GET requests per second, as per EDGAR’s website policy.

extract_items.py cleans and extracts the text of all or particular items from downloaded documents, regardless of file format (e.g., older .txt or modern .html filings). Users can select which sections to extract and optionally remove numerical tables. Special characters, extra spaces/newlines, HTML tags etc. are automatically removed to obtain the clean text data. The extracted data is saved as JSON files, with each item’s text stored as a separate key-value pair (Fig. 3), for easy integration in NLP pipelines.⁸

The extraction mechanism uses a comprehensive collection of regular expressions to identify unique and subsequent item headers (e.g., Item 9A – Controls and Procedures) in the text of the target reports.⁹ We then assign the matched content located between the specified item headers as the item sections. Since a specific item can also be mentioned inside paragraphs of other items (not as a

header), we take several precautions to ensure that we do not cut an item section short or detect an item header in the wrong place. For example:

- We require each item header to be preceded explicitly by a newline character (\n).
- We first conduct a case-sensitive search and employ a lower-case search only as a fallback scenario.
- In the rare case where we identify multiple possible matches, we keep the longest one.

Also, since 10-Q filings (quarterly reports) are separated into two parts whose items can have the same names, we follow the above approach, but hierarchically: we first identify the two parts, and then search each part separately for its item sections.¹⁰

4 Example usage of EDGAR-CRAWLER

4.1 Configuration

Before running any module, one needs to configure the toolkit via the config.json file. Followingly (Fig. 2), we provide one example scenario of how to use EDGAR-CRAWLER. More examples are available on GitHub.

```

1 {"download_filings": {
2   "start_year": 2014,
3   "end_year": 2020,
4   "quarters": [1, 2, 3, 4],
5   "filing_types": ["10-K"],
6   "cik_tickers": ["MSFT"],
7   "user_agent": "your_name@mail.com",
8   "raw_filings_folder": "RAW_FILINGS",
9   "indices_folder": "INDICES",
10  "filings_metadata_file": "FILINGS_METADATA.csv",
11  "skip_present_indices": true},
12 "extract_items": {
13   "raw_filings_folder": "RAW_FILINGS",
14   "extracted_filings_folder": "EXTRACTED_FILINGS",
15   "filings_metadata_file": "FILINGS_METADATA.csv",
16   "items_to_extract": ["1", "1A", "1B", "7", "7A", "8"],
17   "remove_tables": true,
18   "skip_extracted_filings": true}}
  
```

Figure 2: Example Configuration showing how to download 10-K filings for MSFT (Microsoft ticker) from 2014 to 2020 and how to extract items 1, 1A, 1B, 7, 7A, 8.

4.2 JSON Output

After running the toolkit, there will be multiple item-segmented JSON files, one for each filing, containing the clean text data. An example of the JSON output is shown in Figure 3.

5 Contribution to other NLP and AI work

EDGAR-CRAWLER, currently has 300+ stars on Github. Apart from its adoption by practitioners, it is also worth noting that it

⁸Multi-threaded extraction is powered by the pathos library for efficiency.

⁹These regular expressions were meticulously created over many iterations of manually analyzing company filings over various years, starting from 1993.

¹⁰More details of our data extraction pipeline are available on our GitHub.

```

1 {
2   "filename": "92116_10K_2004.html",
3   "cik": "92116",
4   "year": "2004",
5   "...": "...",
6   "item_1": "Item 1. Business ...",
7   "item_1A": "Item 1A. Risk Factors: ...",
8   "item_1B": "Item 1B. Unresolved Staff Comments: None.",
9   "...": "...",
10  "item_7": "Item 7. Management's Discussion and Analysis: ...",
11  "item_7A": "Item 7A. Quantitative and Qualitative Disclosures: ...",
12  "...": "...",
13  "item_15": "Item 15. Exhibits and Financial Statement Schedules: ...",
14  "item_16": "Item 16. Form 10-K Summary: ..."
15 }

```

Figure 3: Example of a financial filing downloaded and extracted to a JSON format by EDGAR-CRAWLER. Multiple filing types are supported.

has been used in multiple NLP and, more generally, AI studies. For example, the U.S. National Bureau of Economic Research used EDGAR-CRAWLER to research customer capital [11], while the U.S. Federal Reserve Board tracked specific 8-K item sections to predict layoffs [8]. Researchers from Goldman Sachs performed business XML tagging by leveraging custom BERT models [18] pre-trained on data generated by our software [20]. Moreover, EDGAR-CRAWLER has enabled the creation of several financial datasets, including the largest financial NLP corpus [17]¹¹ and datasets for specific tasks like fraud detection [21].

In addition, EDGAR-CRAWLER has contributed in advancing NLP research benchmarks. For instance, Cao et al. [4] benchmarked LLMs from OpenAI and Anthropic on long-form document summarization using data from EDGAR-CRAWLER, while Hillebrand et al. [12] utilized Word2vec embeddings [19] trained on data generated by our toolkit to perform relation extraction. Finally, Ahn et al. [2] developed a network graph-based investing method using EDGAR-CRAWLER data, outperforming the S&P 500 Index¹².

6 Conclusions

We introduced EDGAR-CRAWLER, the first open-source software (OSS) library for extracting item-segmented raw documents from EDGAR, the official repository of publicly traded companies in the US. EDGAR-CRAWLER efficiently downloads financial filings, segments them by item, and exports structured text sections in JSON format, enabling seamless integration into data processing pipelines. EDGAR-CRAWLER serves as a free and open-source alternative to paying proprietary web services for financial data. As a result, EDGAR-CRAWLER has gained mass adoption both from practitioners in Github, as well as the academic community. The software is available at <https://github.com/lefterisloukas/edgar-crawler>.¹³

References

- [1] 2024. ICAIF '24: *Proceedings of the 5th ACM International Conference on AI in Finance* (Brooklyn, NY, USA). Association for Computing Machinery.
- [2] J. C. J. Ahn, D. Gorduz, and S. Park. 2024. Hidden neighbours: extracting industry momentum from stock networks. *Financial Markets and Portfolio Management* 38 (2024), 415–441. <https://doi.org/10.1007/s11408-024-00455-4>
- [3] Steven Balsam, Eli Bartov, and Carol Marquardt. 2002. Accruals management, investor sophistication, and equity valuation: Evidence from 10-Q filings. *Journal of Accounting Research* 40, 4 (2002), 987–1012.
- [4] Tianyu Cao, Natraj Raman, Danial Dervovic, and Chenhao Tan. 2024. Characterizing Multimodal Long-form Summarization: A Case Study on Financial Reports. arXiv:2404.06162 [cs.CL]. <https://arxiv.org/abs/2404.06162>
- [5] Mary Ellen Carter and Billy S Soo. 1999. The relevance of Form 8-K reports. *Journal of Accounting Research* 37, 1 (1999), 119–132.
- [6] Richard A. Cazier and Ray J. Pfeiffer. 2016. Why are 10-K Filings So Long? *Accounting Horizons* 30, 1 (Mar 2016), 1–21. <https://doi.org/10.2308/acch-51240>
- [7] Chung-Chi Chen, Hen-Hsen Huang, Hiroya Takamura, and Hsin-Hsi Chen (Eds.). 2022. *Proceedings of the Fourth Workshop on Financial Technology and Natural Language Processing (FinNLP)*. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (Hybrid). <https://aclanthology.org/2022.finnlp-1.0>
- [8] Leland Crane, Emily Green, Molly Harnish, Will McClennan, Paul Soto, Betsy Vrankovich, and Jacob Williams. 2024. Tracking Real Time Layoffs with SEC Filings: A Preliminary Investigation. *Finance and Economics Discussion Series* (04 2024), 1–30. <https://doi.org/10.17016/FEDS.2024.020>
- [9] Priyank Gandhi, Tim Loughran, and Bill McDonald. 2019. Using Annual Report Sentiment as a Proxy for Financial Distress in U.S. Banks. *Journal of Behavioral Finance* 20, 4 (October 2019), 424–436. <https://doi.org/10.1080/15427560.2019.15533171>
- [10] Paul A Griffin. 2003. Got information? Investor response to Form 10-K and Form 10-Q EDGAR filings. *Review of Accounting Studies* 8 (2003), 433–460.
- [11] Bianca He, Lauren I Mostrom, and Amir Sufi. 2024. Investing in Customer Capital. 33171 (November 2024). <https://doi.org/10.3386/w33171>
- [12] Lars Patrick Hillebrand, Tobias Deußer, Tim Dilmaghani Khameneh, Bernd Kliem, Rüdiger Loitz, Christian Bauckhage, and Rafet Sifa. 2022. KPI-BERT: A Joint Named Entity Recognition and Relation Extraction Model for Financial Reports. *2022 26th International Conference on Pattern Recognition (ICPR)* (2022), 606–612. <https://api.semanticscholar.org/CorpusID:251280287>
- [13] Apostolos G. Katsafados, Ion Androutsopoulos, Ilias Chalkidis, Manos Fergadiotis, George N. Leledakis, and Emmanouil G. Pyrgiotakis. 2020. *Textual Information and IPO Underpricing: A Machine Learning Approach*. MPRA Paper 103813. University Library of Munich, Germany.
- [14] Shimon Kogan, Dmitry Levin, Bryan R. Routledge, Jacob S. Sagi, and Noah A. Smith. 2009. Predicting Risk from Financial Reports with Regression. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Boulder, Colorado, 272–280. <https://aclanthology.org/N09-1031>
- [15] Heeyoung Lee, Mihai Surdeanu, Bill MacCartney, and Dan Jurafsky. 2014. On the Importance of Text Analysis for Stock Price Prediction. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. Reykjavik, Iceland, 1170–1175.
- [16] Yen-Jung Lee. 2012. The effect of quarterly report readability on information efficiency of stock prices. *Contemporary Accounting Research* 29, 4 (2012).
- [17] Lefteris Loukas, Manos Fergadiotis, Ion Androutsopoulos, and Prodromos Malakasiotis. 2021. EDGAR-CORPUS: Billions of Tokens Make The World Go Round. In *Proceedings of the Third Workshop on Economics and Natural Language Processing*. Association for Computational Linguistics, Punta Cana, Dominican Republic, 13–18. <https://doi.org/10.18653/v1/2021.econlp-1.2>
- [18] Lefteris Loukas, Manos Fergadiotis, Ilias Chalkidis, Eirini Spyropoulou, Prodromos Malakasiotis, Ion Androutsopoulos, and Georgios Paliouras. 2022. FINER: Financial Numeric Entity Recognition for XBRL Tagging. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Dublin, Ireland, 4419–4431. <https://doi.org/10.18653/v1/2022.acl-long.303>
- [19] Tomás Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. In *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*, Yoshua Bengio and Yann LeCun (Eds.).
- [20] Soumya Sharma, Subhendu Khatuya, Manjunath Hegde, Afreen Shaikh, Koustuv Dasgupta, Pawan Goyal, and Niloy Ganguly. 2023. Financial Numeric Extreme Labelling: A dataset and benchmarking. In *Findings of the Association for Computational Linguistics: ACL 2023*. Association for Computational Linguistics, Toronto, Canada, 3550–3561. <https://aclanthology.org/2023.findings-acl.219>
- [21] Elias Zavitsanos, Dimitris Mavroeidis, Konstantinos Bougiatiotis, Eirini Spyropoulou, Lefteris Loukas, and Georgios Paliouras. 2022. Financial Misstatement Detection: A Realistic Evaluation. In *Proceedings of the Second ACM International Conference on AI in Finance (Virtual Event) (ICAIF '21)*. Association for Computing Machinery, New York, NY, USA, Article 34, 9 pages. <https://doi.org/10.1145/3490354.3494453>

7 Acknowledgments

We sincerely thank the Hellenic Artificial Intelligence Society (EETN) (<https://www.eetn.gr/en/>) for their sponsorship.

¹¹An early version of EDGAR-CRAWLER was developed specifically for that old work of ours. However, the current paper is the first publication that discusses the toolkit in depth. If you use EDGAR-CRAWLER, please cite this current paper.

¹²https://en.wikipedia.org/wiki/S%26P_500

¹³EDGAR-CRAWLER is continuously developed. Please check our Github for any updates.

4

MODULAR LEARNING FOR FINANCIAL IR

In this chapter, we explore using a modular learning approach in order to train bottleneck adapters on top of representational encoder-only LMs and extract key information from textual data this way. As we described in [section 2.2](#), adapters present a cost-effective solution to adjusting a pre-trained model for a specific target task. But the modular approach holds significant other advantages over simple fine-tuning. In the following few sections we present our case for using multiple different bottleneck adapters in combination for knowledge composition when various models need to be trained. We show that by using Adapter Fusion (introduced in the second half of [section 2.2](#)) in order to be able to combine the knowledge of the different models at runtime, we can nearly always achieve a better performance compared to simply using a singular model.

After having introduced a tool for working with financial reports in [chapter 3](#), we move on to the main type of textual data we will work with for the remainder of the thesis: news articles. We first look at determining the type of news article - can we detect if an article is satire, or if it represents the opinion of a singular reporter? Next, we will delve further into the field of corporate sustainability, as we attempt to extract different ESG-aspects from text. Does an article concern environmental, social or governance related ESG-issues? Finally, we will build on this and further extract the type of ESG-impact an article represents. In all three works, we present different architectures of adapters for extracting the desired information. We find that by being creative with the presented challenge and combining the adapters in different ways, we can generally outperform an uninspired baseline which directly trains on the different classes.

Whereas all three works concern simple classification tasks, every one of them is multilingual, which both complicates the tasks, but also offers an opportunity, as generally, more data is available. We find that applying Adapter Fusion across the different languages can enhance the results compared to using just a monolingual adapter.

4.1 News Genre Classification

Paper: Fabian Billert and Stefan Conrad. 2023. HHU at SemEval-2023 Task 3: An Adapter-based Approach for News Genre Classification. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 1166–1171, Toronto, Canada. Association for Computational Linguistics.

Personal Contribution: The presented method was developed and implemented by Fabian Billert. All parts of the paper were written by Fabian Billert under the supervision of Stefan Conrad

Status: Published

Fake news and propaganda in online media have been a much discussed topic in recent years. From 2009 to 2023, the number of active fact-checking initiatives rose from 17 to 457 [141], but has started to decline slightly in 2024, down to 439. Together with the news that big companies such as Meta have stopped employing fact-checkers on their website¹, the nature of validating online content seems to be changing. Another approach in handling fake news is to enable users to report posts which are not accurate, a feature called “Community Notes” by the popular social media platform X (formerly Twitter). While research suggests that such crowdsourced initiatives are effective [115], they have also been criticized by others as detecting fake news at an early stage (before mass distribution) is not possible and they rely on the users to possess the knowledge necessary to refute the false post in the first place [58]. In addition, as Slaughter et al. [138] find, although the notes do seem to reduce the amount of engagement with the marked posts as well as the number of likes they get, the number of views they get is not affected as much. A majority of research on fake news focuses on social media platforms, but even in more conventional media sources, misleading content can be found. Interestingly, Horne et al. [56] find that fake news articles found there are more similar to satire than to real news containing true information.

To tackle the above mentioned problems among others, employing modern NLP approaches to analyze news articles has garnered increasing attention in recent years [58, 175, 180, 178]. In the following paper, we expand on previous research by training a system to identify if a news article represents objective reporting, an opinion piece, or is satire. Although this problem is a fairly straightforward classification task, the main challenge in this case is caused by the multilingual nature of the data: we work with six different languages in the training dataset, as well as two unseen “surprise” languages designated to test zero-shot performance of the developed system.

We show that for tasks like this, where the data is fragmented but regards the same task, using knowledge composition through adapter fusion [111], which we introduced in section 2.2, is very suitable. In addition, as shown by Pfeiffer et al. [114], by using a combination of language- and task-adapters, it is possible to transfer knowledge for similar tasks between different languages, which is obviously helpful for zero-shot prediction in new languages.

By developing models which can separate objective reporting from opinion pieces

¹<https://www.nytimes.com/live/2025/01/07/business/meta-fact-checking>

and satire, we hope to advance research on combating issues such as fake news. In addition, this could also be helpful in a financial context. Araci et al. [11], for example find that while negative fake news does drive stock prices down (meaning, investors believe the fake news and act accordingly), positive fake news does not contribute positively to the stock price of companies. As a result, identifying false financial news could be beneficial for a sentiment-based portfolio approach.

HHU at SemEval-2023 Task 3: An Adapter-based Approach for News Genre Classification

Fabian Billert and Stefan Conrad

Heinrich-Heine University of Düsseldorf
{fabian.billert, stefan.conrad}@hhu.de

Abstract

This paper describes our approach for Subtask 1 of Task 3 at SemEval-2023. In this subtask, task participants were asked to classify multilingual news articles for one of three classes: Reporting, Opinion Piece or Satire. By training an AdapterFusion layer composing the task-adapters from different languages, we successfully combine the language-exclusive knowledge and show that this improves the results in nearly all cases, including in zero-shot scenarios.

1 Introduction

In SemEval-2023 Task-3 (Piskorski et al., 2023), the authors intend to find out what makes a news article persuasive. To this end, they designed three subtasks: News Genre Categorisation, Framing Detection and Persuasion Techniques Detection. In this paper, we present our results for the first subtask, which aims to classify the genre of a news article into one of three classes; reporting, opinion piece or satire. The news articles are available in different languages, meaning this is a multilingual problem. Furthermore there are a few languages without training data in order to test the performance of the models in a zero-shot scenario. The results for the different languages were evaluated separately while taking the F1-macro score as the evaluation metric.

In the following, we present our approach for English, Italian, German and Russian, as well as the zero-shot languages Spanish and Greek. We use an adapter-based setup (Pfeiffer et al., 2020b), which enhances BERT-type architectures (Devlin et al., 2018) with modular weights. We train a task-adapter for each of the languages before using AdapterFusion (Pfeiffer et al., 2020a) to combine what the language-specific task adapters learned. We show that using AdapterFusion largely improves the performances of the single-language task-adapters, which is a behavior that has already

been observed for low-resource tasks. With an F1-macro score of 0.5942, we achieved 4th place out of 27 teams for the English dataset. In addition, for the zero-shot language Greek, we came in third place out of 15 teams with an F1-macro score of 0.7504. For the other languages, we ranked near 10th place, meaning that our approach still shows some instabilities.

2 Background

The training data for subtask-1 consists of whole texts in six different languages: English, Italian, German, French, Polish and Russian. As we noted above, we only submitted results for English, Italian, German and Russian. We will elaborate on the reason for this in subsection 3.4. However, we still used the French and Polish datasets to augment our data, as we explain in the next section.

The main challenge of this subtask is presented by the limited availability, as well as the imbalance of the training data. Here, the opinion label is over-represented (76% of the labels) while the satire label is the clear minority class (5.8% of the labels). On top of that, the texts are often longer than the 512 token-limit BERT-type systems can handle. Consequently, utilizing all of the available data requires preprocessing the training data to address these issues.

Since the dataset for each language is fairly small but we have data for multiple languages, it makes sense to try and combine the distributed knowledge that can be extracted from the training data. To fully exploit the information contained in all languages, we use AdapterFusion (Pfeiffer et al., 2020a), which combines task-specific representations using a task-composition layer. We will elaborate more on adapters and AdapterFusion in the next section.

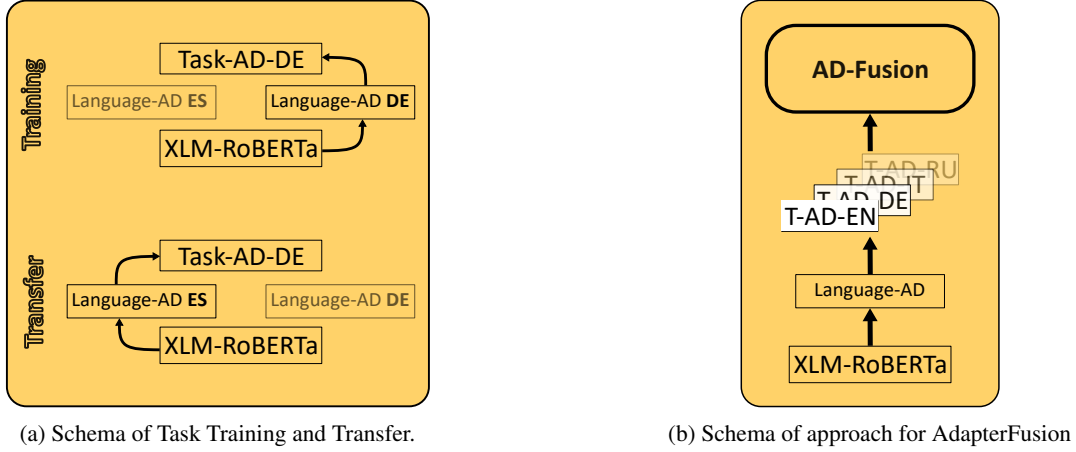


Figure 1: Approaches for zero-shot predictions with a pre-trained task-adapter and for combining task adapters for AdapterFusion. Note that these are schemata, the adapters are actually located within the XLM-R model, behind the feed-forward network (see (Pfeiffer et al., 2020a)) (a) Top: The task-adapter is trained using German data. Bottom: To do a zero-shot prediction for a different language (in this case, Spanish), the language-adapter is exchanged with the one from the zero-shot-language. (b) Several task adapters can be combined in parallel to train an AdapterFusion layer which learns to assign different weights to the task adapters depending on the input it is trying to predict.

3 System Overview

As mentioned, we use an adapter-based model to classify the news data. Adapter-based frameworks have shown to be very useful in dealing with multilingual problems, as their modular structure makes it possible to adapt task-specific representations to different languages (Pfeiffer et al., 2020c). Before we get into the details of our experimental setup, we want to elaborate on different measures we took in order to augment our data, as well as introduce the concepts of adapters and adapter fusion.

3.1 Data Augmentation

As a first step to augment our data, we decided to translate the training data from each language to all the languages we aim to build a model for. For this, we used the OPUS-MT models published in (Tiedemann and Thottingal, 2020) from the huggingface-hub¹. We translated each sentence separately, using nltk’s sentence-tokenizer to split the texts into sentences (Bird et al.). For some language pairs, no translation models were available. In these cases, we did not translate the data. For English, we translated the data from all other languages. For German, we were missing the Russian translation model, for Italian in addition to the Russian, we were missing the model for translation to Polish. Finally, for Russian, the only translation models available on the huggingface-hub are the ones to

English and French.

Since the average text of the training data is longer than the 512 tokens BERT-type models can work with, we further augmented our data by dividing each text into multiple subtexts, depending on the length of the text. Each of these subtexts was then used as a new row with the same label as the original text in our training data.

3.2 Adapters

Adapters offer a modular and parameter-efficient solution for fine-tuning a base-model for specific tasks (Houlsby et al., 2019) or transfer task-specific knowledge to different languages (Pfeiffer et al., 2020c). Since subtask-1 consists of training data in multiple different languages, but the amount of data for a single language is rather small, we approached this problem with the plan to use adapters as a vehicle to combine the knowledge available in each language.

Adapter modules are added into the layers of pre-trained models and aim to learn one specific task without changing the weights of the pre-trained model (Pfeiffer et al., 2020b). Similarly to training adapters for tasks, it is possible to train language-specific adapters by adding an adapter to a multilingual base model and subsequently training the adapter using MLM (Pfeiffer et al., 2020c). When training a task-adapter with a multilingual base model, it is useful to also use a fixed language-adapter as it captures language-specific knowl-

¹<https://huggingface.co/Helsinki-NLP>

edge, thus improving the final performance (Pfeiffer et al., 2020c). Transferring the knowledge from the trained task adapter to a different language is quite straight-forward: One simply needs to exchange the language-adapter with the language-adapter of the target language. We will use this setup for the zero-shot languages. A schema for this approach is displayed in Figure 1a.

3.3 Adapter Fusion

As explained in the previous subsection, adapters are modular and can be used in tandem with one another. But stacking them, as we are doing with language- and task-adapters, is only one display of this. A different way of combining adapters is called AdapterFusion. In AdapterFusion, different adapters are used in parallel in order to combine their task-specific knowledge (Pfeiffer et al., 2020a). AdapterFusion has been shown to be especially beneficial for tasks with small datasets, as it can rely more on the tasks that were trained on the larger datasets. However, our augmented data-sizes are fairly similar for most of the languages. Consequently, it is not clear if AdapterFusion will lead to performance gains or not.

Our approach for training AdapterFusion is to first train a task-adapter for the data in each language as explained in the previous subsection. Afterwards, we train AdapterFusion by stacking all the task-adapters in parallel on top of a language-adapter, as shown in Figure 1b.

3.4 Experimental Setup

To build the structure of our model, we used the AdapterHub introduced in (Pfeiffer et al., 2020b). This hub offers a wide range of pre-trained adapters, as well as a library to create new adapters and train them. Importantly, one can find many language-adapters in the AdapterHub. As our base-model, we use XLM-RoBERTa, which has shown good performances among other multilingual models (Conneau et al., 2019).

For each language, we then train a task-adapter by first stacking a language-adapter on top of the XLM-R model, before stacking a task-adapter on top of the language-adapter (this setup is explained with more details in (Pfeiffer et al., 2020c)). Because the AdapterHub does not have language-adapters with the XLM-R architecture for French or Polish, we only trained adapters for the other four languages.

As a baseline to the multilingual approach, we also

trained several task-adapters which used monolingual models as their base. This baseline should yield similar results to simply fine-tuning the base-model itself, although it doesn't permit us to use a language-adapter for predictions on the zero-shot languages. We used the following base models: For English we used RoBERTa-base (Liu et al., 2019), for Russian we used ruBERT (Kuratov and Arkhipov, 2019) and for Italian² and German³ we used models from the huggingface-hub.

We trained each task adapter using a learning rate between $5 \cdot 10^{-5}$ and 10^{-4} , with a batch size of 8. We used a weighted cross-entropy loss function to account for the imbalanced data.

After training a task-adapter for each language, we trained AdapterFusion by stacking all the task-adapters in parallel on top of a language adapter (see Figure 1b). We trained an AdapterFusion layer for each language, because to train a single AdapterFusion for all languages, one would have to change the model configuration by switching out the language-adapters during training, which is not straightforward. For the training of AdapterFusion, we applied a learning rate of $5 \cdot 10^{-5}$, a batch size of 8 and again used a weighted loss function.

4 Results

In Table 1 we show the results of the various model-configurations on the evaluation- and test-set and highlight the results we submitted to the task organizers with a dagger-symbol. In addition to the approaches with a multilingual base-model we show the results of the baseline which uses a monolingual base-model and fine-tunes a task-adapter stacked on top of it.

In order to be as clear as possible, we divide this section into two parts - the first describes results submitted before the task-deadline while the second describes results achieved after the deadline.

4.1 Pre-Deadline Results

The results of the models for the evaluation-set are mixed: For English and Russian, the configuration using AdapterFusion achieves the best performance by far. However, for Italian and German, the AdapterFusion approach is worse than the approach which uses simply the task-adapter. The

²<https://huggingface.co/dbmdz/bert-base-italian-cased>

³<https://huggingface.co/bert-base-german-cased>

Configuration	EN	IT	DE	RU	ES	GR	Average
Evaluation-Set							
XLM-ADTask	0.3615	0.6101	0.7333	0.4895			0.5486
XLM-ADFusion	0.4769	0.5595	0.6530	0.5551			0.5611
Ro/BERT/a-ADTask	0.4189	0.4221	0.6889	0.4762			0.5015
Test-Set							
XLM-ADTask	0.4436	0.4552[†]	0.6109 [†]	0.3561	0.3267 [†]	0.7504[†]	0.4905
XLM-ADFusion	0.5942[†]	0.4375	0.6996	0.4257[†]	0.5795	0.7502	0.5811
Ro/BERT/a-ADTask	0.5767	0.4112	0.6374	0.3738			0.4998
Final Rank	4	10	11	11	10	3	8

Table 1: F_1 -macro scores of our different model configurations on the evaluation (top) and test-set (bottom). Scores marked by a dagger-symbol were submitted before the task-deadline and represent our scores on the leaderboard.

baseline approach generally achieved similar performances to the task-adapter, although it is much worse for the Italian language.

We always used the model that achieved the best performance on the evaluation-set, marked in bold in Table 1, to predict on the test-set. Comparing the results between evaluation- and test-set, the absolute value of the F_1 -macro score changes a lot for all languages. For English, it is more than 10 percentage-points higher while it is significantly lower for all other languages, with up to 16 points for Italian. This suggests that the evaluation and test data are composed differently, however, since we do not have access to the test labels, we cannot investigate this further.

Overall, the German task-adapter achieved the highest F_1 -macro score on the evaluation set, so we also used this adapter for the zero-shot languages. We applied the standard approach from (Pfeiffer et al., 2020c), replacing the German language-adapter with the zero-shot language-adapter and then predicting. The results of the zero-shot languages for the test-set were varied, with a high F_1 -macro score for Greek, but a low value for Spanish. We didn’t predict for Georgian, as, again, there is no language-adapter with the correct configuration (XLM-R) available for this language on the adapter-hub.

4.2 Post-Deadline Results

Since the F_1 -macro-scores for the test-dataset were only released after the deadline, we did some more tests to see how different configurations performed. The AdapterFusion setup of the zero-shot languages displayed in Table 1 uses the German AdapterFusion as its base while only replacing the language-adapter. We can see that in the case of

German, the configuration that performed the best on the evaluation set actually performs worse on the test set. The same can be observed when using the German task-adapter to predict for the zero-shot language Spanish, as the AdapterFusion approach shows a better result. Overall, the setup using AdapterFusion performs best on the test set in four out of six cases, although the performance in the Greek case is fairly similar for both configurations. In the last column of the table, we calculated the average score for all languages with values. We can see that for both the evaluation, as well as the test-set, AdapterFusion has the highest average score.

5 Performance Analysis

5.1 Performance in Zero-Shot scenario

We did a series of experiments to test the zero-shot predictions when transferring knowledge from different languages and show the results in Table 2. For each language, we transferred once using simply the task-adapter, and once using the AdapterFusion setup. In addition, we calculated the average values of all results from the task-adapters and the AdapterFusions and show them in the last column. For almost all results, using AdapterFusion improves the results in the zero-shot language. There are only two cases where this is not the case: For Greek, when transferring from German and when transferring from Italian, although it is very close in the first case.

In general, one would expect the performances of the different AdapterFusions to be similar, since they all consist of the same task-adapters. The only difference between them should be the language of the training data with which they were trained. In Table 2, we can see that the AdapterFusion re-

Zero-Shot Language	EN		DE		IT		RU		Average	
	Task	Fusion	Task	Fusion	Task	Fusion	Task	Fusion	Task	Fusion
ES	0.4066	0.4325	0.3267	0.5795	0.3353	0.5045	0.4318	0.4832	0.3751	0.4999
GR	0.3573	0.6079	0.7504	0.7502	0.6807	0.6521	0.3258	0.6110	0.5286	0.6553

Table 2: Comparison of the F_1 -macro scores when transferring knowledge to the zero-shot languages (rows) from different origin-languages (columns).

sults for a zero-shot language are in a similar range: For Spanish, they vary between 0.4325 and 0.5795 and for Greek, the values lie between 0.6079 and 0.7502. However, considering they use the same task-adapters in their configuration, this is still a rather large difference. Because of this, we decided to take a closer look at the AdapterFusion attentions with regards to the different task adapters.

5.2 AdapterFusion Attentions

The AdapterFusion layer attends over the different task-adapters when an input is passed through it and activates them accordingly. By investigating these attentions, we can find out how exactly the Fusion layer creates its predictions depending on the input data.

First, we look at the AdapterFusion activations when predicting for the test-data. In Figure 2, we have depicted the mean attentions over the first eleven XLM-R layers for the test-data. We excluded the last layer, since (Pfeiffer et al., 2020a) found that its attentions show some unpredictable behavior because they are located right before the prediction head and not, like the other layers, between the frozen pretrained layers of XLM-R.

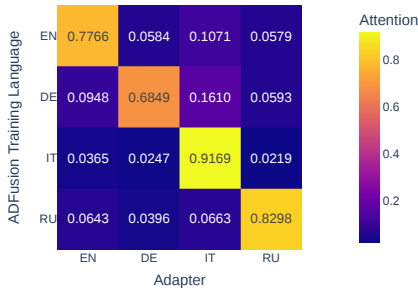


Figure 2: AdapterFusion attentions for the test-data.

We can see that generally, the AdapterFusion tends to activate the adapter of the language it is currently training with. However, especially for Italian, the fusion layer nearly doesn't activate any other task-adapters at all. It appears that the adapters of the other tasks are detrimental in this case - this

is supported by the results in Table 1, where the task-adapter for Italian achieves better results than its AdapterFusion.

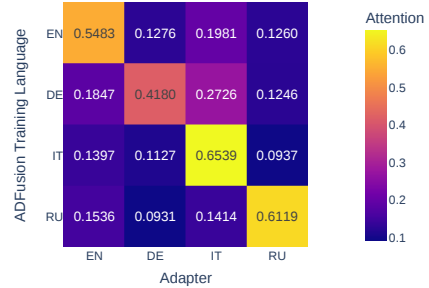


Figure 3: AdapterFusion attentions when transferring to greek.

In addition to the activations for the test-set, we take a closer look at the AdapterFusion attentions while predicting for one of the zero-shot languages, Greek, in Figure 3. We see a similar effect as in Figure 2, meaning the AdapterFusion attends most to the task-adapters of the language it was trained with, which explains the discrepancy between the AdapterFusion-results in Table 2. However, those attentions are generally lower than when predicting for its own language on the test-set. In contrast, the activations of the non-diagonal task-adapters are now much higher, indicating that the fusion now combines the different task-adapters more effectively.

6 Conclusion

We successfully investigate multiple approaches to determine the genre of news articles. We find that using a configuration which utilizes AdapterFusion to combine the knowledge learned from datasets in various languages delivers the best results for most languages. By comparing the attentions of the AdapterFusion between a zero-shot language and a trained language, we see that it effectively composes the learned knowledge when used to predict for an unseen language.

Further work could try to improve the setup of the

AdapterFusion, in order to make it possible to train it with multilingual data at once. In addition, including more languages (for example French and Polish, after training language-adapters for them), might improve the results, as more knowledge would be available.

Jörg Tiedemann and Santhosh Thottingal. 2020. [OPUS-MT – Building open translation services for the World](#). *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*.

References

Steven Bird, Ewan Klein, and Edward Loper. *Natural Language Processing with Python*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. [Unsupervised Cross-lingual Representation Learning at Scale](#). *arXiv*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. [BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding](#). *arXiv*.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. [Parameter-Efficient Transfer Learning for NLP](#). *arXiv*.

Yuri Kuratov and Mikhail Arkhipov. 2019. [Adaptation of Deep Bidirectional Multilingual Transformers for Russian Language](#). *arXiv*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. [RoBERTa: A Robustly Optimized BERT Pretraining Approach](#). *arXiv*.

Jonas Pfeiffer, Aishwarya Kamath, Andreas Rücklé, Kyunghyun Cho, and Iryna Gurevych. 2020a. [AdapterFusion: Non-Destructive Task Composition for Transfer Learning](#). *arXiv*.

Jonas Pfeiffer, Andreas Rücklé, Clifton Poth, Aishwarya Kamath, Ivan Vulić, Sebastian Ruder, Kyunghyun Cho, and Iryna Gurevych. 2020b. [AdapterHub: A Framework for Adapting Transformers](#). *arXiv*.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2020c. [MAD-X: An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer](#). *arXiv*.

Jakub Piskorski, Nicolas Stefanovitch, Giovanni Da San Martino, and Preslav Nakov. 2023. SemEval-2023 Task 3: Detecting the Category, the Framing, and the Persuasion Techniques in Online News in a Multilingual Setup. In *Proceedings of the 17th International Workshop on Semantic Evaluation, SemEval 2023*, Toronto, Canada.

4.2 ESG Key Issue Identification

Paper: Fabian Billert and Stefan Conrad. 2023. Team HHU at the FinNLP-2023 ML-ESG Task: A Multi-Model Approach to ESG-Key-Issue Classification. In *Proceedings of the Fifth Workshop on Financial Technology and Natural Language Processing and the Second Multimodal AI For Financial Forecasting*, pages 146–150, Macao.

Personal Contribution: Fabian Billert developed and implemented the presented approach. Fabian Billert wrote the paper under the supervision of Stefan Conrad.

Status: Published

We introduced the term ESG, short for Environmental, Social and Governance in [subsection 2.3.3](#). Those three aspects are the key pillars of most ESG-frameworks used in the industry in order to classify pieces of information. The reason for looking at different aspects rather than corporate sustainability as a whole is rooted in the concept of materiality: the significance of specific ESG issues varies considerably across industries due to their unique operational characteristics, risk exposures, and stakeholder landscapes [70, 88]. This means that for companies belonging to certain industries, specific ESG-events or factors are far more impactful than others.

For example, a bank will probably not pollute the environment with harmful substances directly from its operations. Its environmental impact is often indirect, primarily through its financing activities, and can exhibit non-linear relationships with market value. At the same time, a bank might make sustainable investments which, as Ersoy et al. [36] show, could increase the value of the company. Another example is shown by Wang et al. [160], who find that for heavily polluting industries in China, there is a significant correlation between ESG scores and the amount and quality of published patents. This implies that these companies are driven by their ESG commitments to foster an environment that encourages innovation, leading to the development of new technologies and processes. This can be attributed to factors such as reduced agency problems, enhanced information disclosure, and improved internal corporate governance, which collectively spur the creation of patents related to cleaner operations, sustainable materials, or improved efficiencies.

To summarize, not every aspect of sustainability has the same importance for a company. For this reason, different frameworks attempt to further categorize ESG in various sub-aspects. One popular framework is the MSCI ESG-rating methodology, which judges companies based on more than thirty key issues². In order to determine an ESG rating for a company, different weights are assigned to these key issues depending on the company’s industry (accounting for the fact that some ESG-related news are more impactful for a company than others).

In this paper, we examine ESG-events and attempt to determine which of the 35 key-issues they belong to. As in the previous section, we use an adapter-based setup and explore different approaches. Again, the problem consists of a multilingual dataset and in addition, the ESG-issues build an interesting tree-type structure with multiple

²At the time of the writing of this paper, MSCI defined 35 key issues, however, issues might be added or removed according to changes in MSCI’s methodology.

levels (first, the coarse E/S or G aspects and then the more fine-grained sub-issues). This allows for the development of creative prediction systems.

Team HHU at the FinNLP-2023 ML-ESG Task: A Multi-Model Approach to ESG-Key-Issue Classification

Fabian Billert and Stefan Conrad

Heinrich-Heine University of Düsseldorf

{fabian.billert, stefan.conrad}@hhu.de

Abstract

In this paper, we discuss our submission to the Multi-Lingual ESG Issue Identification (ML-ESG) 2023, where we classify news articles into different ESG key-issues defined by MSCI. We use an adapter-based approach and evaluate different approaches and configurations, finally showing that it is advantageous to use multiple models in order to first classify articles into E/S/G classes before determining the final sub-issues.

1 Introduction

The surge in Environmental, Social, and Governance (ESG) research over the past few years is a testament to the growing importance of these issues in the corporate world (Zumente and Bistрова, 2021). Companies are increasingly recognizing that ESG-related matters can pose significant risks if not addressed properly (Aue et al., 2022). Beyond risk management, ESG topics are also crucial for a company’s reputation, as they often reflect the company’s values and commitment to sustainable practices (Schramm-Klein et al., 2016), (Islam et al., 2021).

Investors, too, are becoming more attentive to the ESG behaviors of companies. One common method of evaluating a company’s ESG practices is through the human-curated scores provided by major rating agencies like MSCI¹ or Sustainalytics². These agencies assess whether a company adheres to good ESG practices and assign a numerical value to represent the company’s ESG performance. MSCI does this by considering 35 key-issues which they combine in different ways depending on the specific industry a company operates in (Nagy et al.). The final weight of a

key-issues in the ESG score calculation is determined by quantitatively assessing each industry and consulting with investment practitioners.

Upon closer scrutiny of the evaluation metrics employed by the different ESG rating agencies, it becomes evident that these metrics do not completely incorporate sustainability principles into their process of assessing corporate sustainability (Escrig-Olmedo et al., 2019). In addition, (Crona, 2021) raises several concerns with the traditional rating agencies. One point of critique mentioned by them is that companies might self report data on positive environmental initiatives that are not connected to their negative environmental impact, but are similarly considered by the rating agencies. On the other hand, scoring mechanisms like the one used by MSCI are problematic in the sense that the weighting mechanism might not consider key ESG issues, depending on how the weights were created. These uncertainties in the evaluation process underscore the need for more comprehensive and nuanced methods of assessing ESG practices.

In order to create independent analyses, machine learning techniques, particularly those in Natural Language Processing (NLP), can be used. Over the past few years, NLP research has seen a significant uptick, with advancements in this field offering promising solutions for more in-depth ESG analysis (Min et al., 2021), (Chen et al., 2022), (Fischbach et al., 2022). By leveraging NLP, investors can conduct their own research to determine the sustainability of potential investment companies.

The Machine Learning for ESG (ML-ESG) task (Chen et al.) aims to motivate research in this direction and has annotated news articles in English, French, and Chinese based on the 35 key-issues used by MSCI. The task challenges participants to develop a system capable of classifying

¹<https://www.sustainalytics.com/esg-data>

²<https://www.msci.com/our-solutions/esg-investing/esg-ratings>

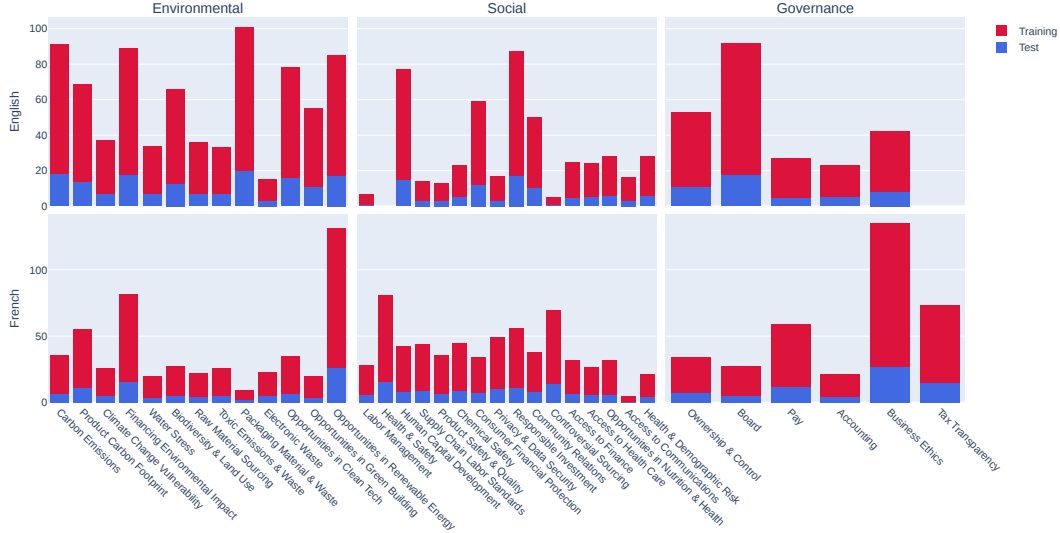


Figure 1: Occurrences of the different labels in the training- and test-data of the task. The top row represents the English data, the bottom row the French data. Each column represents one out of E/S/G as written above the figure. Training data is shown in red, test data in blue. **Note:** The test data was released after the task deadline and was not used during training unless mentioned.

articles in these languages into the appropriate ESG issues. This represents a significant stride towards more nuanced and comprehensive ESG analysis, ultimately enabling more informed and sustainable investment decisions.

In this paper, we present our solution for the ML-ESG task for the English and French datasets. We train a set of adapters for each language and try out different approaches to classify the news articles, ultimately showing that it is preferable to first classify a news article into one of the three main classes (Environmental, Social or Governance), before further classifying into the key-issues belonging to each category. Our approach achieves third place for the French language and nineteenth for the English language.

2 Task Description and Dataset

The dataset contains 1200 French and 1199 English news articles. Each article has the following properties: "URL", "news_title", "news_content", "ESG_label" (Chen et al.). The "ESG_label" is one of the 35 key-issues described in the ESG Industry Materiality Map of MSCI³. Each key-issue is attributed to one of the three top ESG components, "environmental", "social" and "governance". In

³<https://www.msci.com/our-solutions/esg-investing/esg-industry-materiality-map>

Figure 1, different histograms for each ESG component show the occurrence of all the key-issues per language. Most news articles are classified as one of the environmental key-issues while the least articles belong to governance key-issues.

3 Experimental Approach

3.1 Adapters

Adapters are an efficient and flexible method for fine-tuning a foundational model for unique tasks (Houlsby et al., 2019) or transferring task-specific knowledge across different languages (Pfeiffer et al., 2020b). These tools are particularly useful when dealing with a dataset composed of multiple languages.

Adapter modules, which are incorporated into the layers of pre-existing models, are designed to master a particular task without altering the weights of the original model (Pfeiffer et al., 2020a). They are more parameter efficient than fine-tuning the full model while achieving nearly the same performance (Houlsby et al., 2019). Much like adapters that are trained for specific tasks, we can also train language-specific adapters. This is achieved by adding an adapter to a multilingual base model and then training it using Masked Language Modeling (MLM) (Pfeiffer et al., 2020b). If a task adapter is being trained with a multilingual base model, it is beneficial to also utilize a fixed language adapter because it captures and applies language-specific

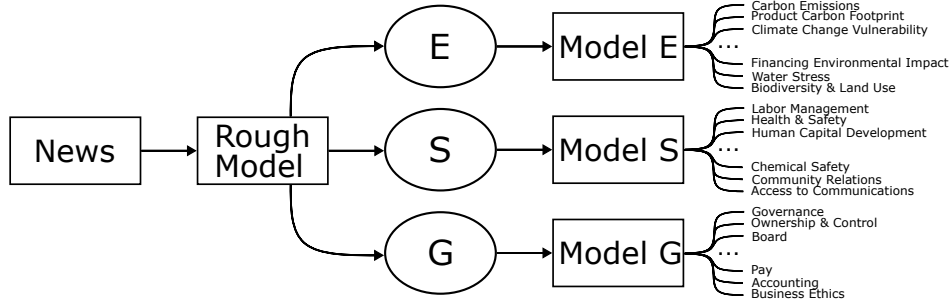


Figure 2: Schematic description of method 3. We first train a model to determine if a news article describes environmental, social or governance thematics. A second model then classifies the article with regards to the different key-issues belonging to that component. Each model in this case is constituted by a base-model, a language-adapter and a task-adapter as explained in subsection 3.1.

knowledge, which can enhance the final performance (Pfeiffer et al., 2020b).

3.2 Title or Content?

Classific. Target	Title	Content
Key-issues	0.63	0.41
Components	0.88	0.80

Table 1: Comparison of the F_1 -macro scores when using the title vs using the content of the news articles for the French dataset. The first row shows results when directly classifying for the 35 key-issues while the second row classifies only for the rough E/S/G components.

We performed several experiments in order to determine if it is better to use the title or the content in order to classify the news articles. In Table 1, we show the F_1 -macro scores when classifying the 35 key-issues directly in the first row, and the results for classifying the rough components (E/S/G) in the second row for the French dataset. In both cases, we achieve the best result when simply using the title to train the adapter. Since we observed a similar result for the English dataset, we decided to continue working without the "news_content" element.

3.3 Data Augmentation

Classific. Target	EN	FR
No Augmentation	0.58	0.69
Augmentation	0.68	0.67

Table 2: Comparison of the F_1 -macro scores when augmenting the data by translating from the dataset in the other language and training on classifying the 35 key-issues directly. The top row shows the results without augmentation, the bottom row with augmentation.

Since we have data in two languages, we tried augmenting the data of each language by translating the data of the other language. For this, we use the OPUS-MT models published by (Tiedemann and Thottingal, 2020) from the huggingface-hub⁴. In Table 2, we show results on the key-issue classification for French and English. The results are conflicting, as we can see an improvement for the English dataset, while the French dataset performs slightly worse with the augmentation.

3.4 Configurations

We designed three different configurations:

- Method 1: Train an adapter on the 35 key-issues directly.
- Method 2: Augment the data by translating from the other language, then train on the 35 key-issues.
- Method 3: First train an adapter to classify an article into the ESG component, then train a set of three adapters, one for each ESG component, in order to classify the key-issues.

A rough schema for the third approach is shown in Figure 2. Since we achieved a better performance when translating for the English dataset (see Table 2), we decided to augment the data for the third approach for this language.

For all configurations, we evaluate the approach on 10% of the original training dataset. This is without augmentation, meaning the size of the eval dataset is only 5% the size of the total dataset for method 2.

As a base model, we use mBERT (Devlin et al.,

⁴<https://huggingface.co/Helsinki-NLP>

Language	Pre-Deadline			Post-Deadline	
	Method 1	Method 2	Method 3	Method 1	Method 3
EN	0.56	0.35	0.57	0.61	0.61
FR	0.75	0.73	0.77	0.78	0.80

Table 3: Weighted F_1 -scores of the trained adapters for the test-set. On the left side, the official results. On the right side, the post-deadline results in which we evaluate on the test set during training.

2018). We then stack a pre-trained language adapter with fixed weights (from the AdapterHub⁵, (Pfeiffer et al., 2020a)) on top of it followed by a task adapter (this setup is explained with more detail in (Pfeiffer et al., 2020b)). For training, we used a learning rate of $5 \cdot 10^{-5}$ and a simple cross-entropy loss-function.

4 Results

The submitted results of the three approaches are displayed on the left side of Table 3. Note that we display the weighted F_1 -score here as opposed to F_1 -macro we used in the previous section. For both languages method 3 shows the best performance, followed by method 1. However, while we placed third for French, the final placement for English was much worse. Our first guess at a reason for this was the imbalance of the training dataset, which we did not consider during the training. However, as we can see in Figure 1 in blue, the test data (which was released after the task-deadline) is similarly distributed as the training data.

4.1 Augmenting Key-Issues

Classific. Target	EN	FR
No Augmentation	0.60	0.77
Aug. Key-issue Model	0.58	0.78
Aug. Rough Model	0.59	0.77
Augmentation Both	0.57	0.77

Table 4: Comparison of weighted F_1 -scores of the test-set when augmenting different parts of the data for method 3 (Figure 2). The first row shows the results without augmentation, the second row when augmenting only during training of the models classifying into the key-issues, the third row when augmenting only the first (rough) model and the fourth row when augmenting for all models.

In Table 3 we can see that the approach augmenting the data with the dataset of the other language (method 2) performs the worst for both languages.

But since we augmented the English data for the method 3 and still measure good results (compared to method 1), we are unsure of the impact of the augmentation here. For that reason, we performed several tests where we train a model using augmented data at different stages. The results are shown in Table 4. We can see that the results are very similar among each language, especially for the French dataset where the configuration which augments the second models (classifying the key-issues) performs slightly better than the rest of the configurations. For the English dataset, the configuration without augmentation shows the best performance, while augmenting both models performs worst.

4.2 Evaluation on Test-Set

In order to determine the best performance possible with our setup, we train adapters on the whole training set, using the labelled test set to evaluate. The results are displayed in the right part of Table 3. Because method 2 performed the worst before, we do not include it here anymore. In addition, we do not augment the English approach for method 3 since we saw a better performance not augmenting in Table 4. We observe that the results improve slightly, but don't account for the difference in F_1 -scores between the two languages.

5 Conclusion

We successfully trained several configurations capable of classifying news articles into the 35 key-issues defined by MSCI, showing that using the title instead of the content of the news article is more performant. We also tried to augment the datasets by translating from the other language but saw that this has little impact, even decreasing the performance in some cases. Out of the three different approaches, we observe that it is generally best to first classify the news articles into their rough ESG components (environmental, social & governance) before using a second model in order to determine the final key-issue.

⁵<https://adapterhub.ml/>

References

- Tanja Aue, Adam Jatowt, and Michael Färber. 2022. [Predicting Companies' ESG Ratings from News Articles Using Multivariate Timeseries Analysis](#). *arXiv*.
- Chung-Chi Chen, Hen-Hsen Huang, Hiroya Takamura, and Hsin-Hsi Chen. 2022. [An Overview of Financial Technology Innovation](#). *Companion Proceedings of the Web Conference 2022*, pages 572–575.
- Chung-Chi Chen, Yu-Min Tseng, Juyeon Kang, Anaïs Lhuissier, Min-Yuh Day, Teng-Tsai Tu, and Hsin-Hsi Chen. Multi-Lingual ESG Issue Identification. In *Proceedings of the Fifth Workshop on Financial Technology and Natural Language Processing (FinNLP) and the Second Multimodal AI For Financial Forecasting (Muffin)*.
- Beatrice Crona. 2021. [Sweet Spots or Dark Corners? An environmental sustainability examination of Big Data and AI in ESG](#). *SSRN Electronic Journal*.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. [BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding](#). *arXiv*.
- Elena Escrig-Olmedo, María Ángeles Fernández-Izquierdo, Idoya Ferrero-Ferrero, Juana María Rivera-Lirio, and María Jesús Muñoz-Torres. 2019. [Rating the Raters: Evaluating how ESG Rating Agencies Integrate Sustainability Principles](#). *Sustainability*, 11(3):915.
- Jannik Fischbach, Max Adam, Victor Dzhagatspanyan, Daniel Mendez, Julian Frattini, Oleksandr Kosenkov, and Parisa Elahidoost. 2022. [Automatic ESG Assessment of Companies by Mining and Evaluating Media Coverage Data: NLP Approach and Tool](#). *arXiv*.
- Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. [Parameter-Efficient Transfer Learning for NLP](#). *arXiv*.
- Tahir Islam, Rauf Islam, Abdul Hameed Pitafi, Liang Xiaobei, Mahmood Rehmani, Muhammad Irfan, and Muhammad Shujaat Mubarak. 2021. [The impact of corporate social responsibility on customer loyalty: The mediating role of corporate reputation, customer satisfaction, and trust](#). *Sustainable Production and Consumption*, 25:123–135.
- Bonan Min, Hayley Ross, Elinor Sulem, Amir Pouran Ben Veyseh, Thien Huu Nguyen, Oscar Sainz, Eneko Agirre, Ilana Heinz, and Dan Roth. 2021. [Recent Advances in Natural Language Processing via Large Pre-Trained Language Models: A Survey](#). *arXiv*.
- Zoltan Nagy, Linda-Eling Lee, and Guido Giese. [ESG Ratings: How the Weighting Scheme Affected Performance](#).
- Jonas Pfeiffer, Andreas Rücklé, Clifton Poth, Aishwarya Kamath, Ivan Vulić, Sebastian Ruder, Kyunghyun Cho, and Iryna Gurevych. 2020a. [AdapterHub: A Framework for Adapting Transformers](#). *arXiv*.
- Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2020b. [MAD-X: An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer](#). *arXiv*.
- Hanna Schramm-Klein, Joachim Zentes, Sascha Steinmann, Bernhard Swoboda, and Dirk Morschett. 2016. [Retailer Corporate Social Responsibility Is Relevant to Consumer Behavior](#). *Business & Society*, 55(4):550–575.
- Jörg Tiedemann and Santhosh Thottingal. 2020. [OPUS-MT – Building open translation services for the World](#). *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*.
- Ilze Zumente and Jūlija Bistropa. 2021. [ESG Importance for Long-Term Shareholder Value Creation: Literature vs. Practice](#). *Journal of Open Innovation: Technology, Market, and Complexity*, 7(2):127.

4.3 ESG Impact Determination

Paper: Fabian Billert and Stefan Conrad. 2023. Exploring Knowledge Composition for ESG Impact Type Determination. In *Proceedings of the Sixth Workshop on Financial Technology and Natural Language Processing*, pages 79–83, Bali, Indonesia. Association for Computational Linguistics.

Personal Contribution: Fabian Billert implemented the presented approach and gathered the results. Fabian Billert wrote the paper while being supervised by Stefan Conrad.

Status: Published

Having determined which ESG-aspect an event is relevant for, an obvious follow-up question is whether the event represents a risk or an opportunity for companies with regards to sustainability. Established ESG-rating agencies have their own methods to do this. However, a common part of their approach consists of human sustainability experts and their analysis [86]. As we discussed in [subsection 2.3.3](#) citing previous research such as Berg et al. [13], ESG-ratings of established agencies have some deficiencies, so one of the main goals of this work is to advance independent systems which allow the public to create their own impressions of a company’s sustainability practices without relying on human analysts.

Our approach for this problem follows the previous two sections and uses bottleneck adapters. However, we take advantage of one of the benefits of building systems with modular models and leverage the adapters trained in [section 4.2](#) in order to include aspect-specific knowledge in the prediction of the impact-type.

Exploring Knowledge Composition for ESG Impact Type Determination

Fabian Billert^{1,2} and Stefan Conrad¹

¹Heinrich-Heine University of Düsseldorf

²GET Capital AG

{fabian.billert, stefan.conrad}@hhu.de

Abstract

In this paper, we discuss our (Team HHU's) submission to the Multi-Lingual ESG Impact Type Identification task (ML-ESG-2). The goal of this task is to determine if an ESG-related news article represents an opportunity or a risk. We use an adapter-based framework in order to train multiple adapter modules which capture different parts of the knowledge present in the training data. Experimenting with various Adapter Fusion setups, we focus both on combining the ESG-aspect-specific knowledge, and on combining the language-specific-knowledge. Our results show that in both cases, it is possible to effectively compose the knowledge in order to improve the impact type determination.

1 Introduction

The substantial rise in Environmental, Social, and Governance (ESG) research over the past few years underscores the growing significance of sustainability in the corporate world (Zumente and Bistрова, 2021). Both investors and companies have a growing interest in ESG issues, as it becomes clearer that they are vital for a company's brand and, consequently, also value (Schramm-Klein et al., 2016), (Islam et al., 2021). Studies show that investor interest in a company's stock depends significantly on whether an estimation of their ESG-practices is available or not (Zumente and Lāce, 2021).

Different agencies have developed a variety of ESG-scoring mechanisms in order to quantify the sustainability practices of companies which offers investors an easier way to determine the ESG-related risk a company represents in their portfolio. Popular examples of this are MSCI¹ and Sustainalytics². However, inconsistencies in

the scoring practices and discrepancies between different score-providers cast doubt on the available ESG-scores (Berg et al., 2019), (Clément et al., 2023). This doubt could be alleviated by providing a more detailed overview of the various aspects of sustainability a company is involved in. Researchers in the field of natural language processing (NLP) have recently made efforts to enhance conventional ESG scoring methodologies by incorporating alternative data sources, aimed at offering a clearer understanding of a company's sustainability practices. (Nicolas et al., 2023) underscores the impact of ESG-risk events found in social media data, illustrating how they can result in adverse financial returns. Meanwhile, (Aue et al., 2022) focuses on analyzing news articles to construct a predictive model for determining ESG scores.

The ML-ESG series of shared tasks aims to advance research in this area of using natural language processing to provide more transparency in ESG topics. ML-ESG-1 focused on determining the ESG-aspects of news articles (Chen et al., 2023b). In the next step of the ML-ESG series, ML-ESG-2 aims to find the impact type of the information in news articles, meaning detecting if they represent opportunities or risks (Chen et al., 2023a).

In this paper, we present our submission for the French and English part of the ML-ESG-2 shared task. We train adapters on different parts of the dataset and experiment with several approaches in order to combine their knowledge using Adapter Fusion. We show that this knowledge composition yields a significant improvement of the performance compared to simply finetuning an adapter for the whole dataset. Our submitted approaches achieve fourth place for French and fifteenth place for English, however we investigated additional approaches after the task deadline which improve on

¹<https://www.msci.com/our-solutions/esg-investing/esg-ratings>

²<https://www.sustainalytics.com/esg-data>

those results.

2 Task Description and Dataset

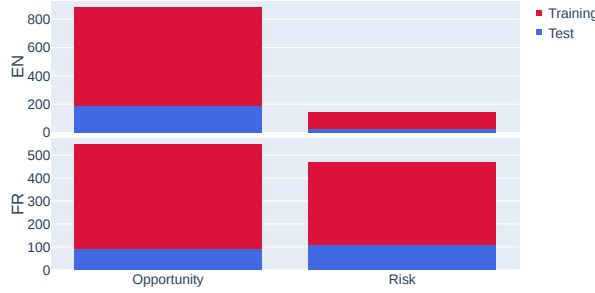


Figure 1: Occurrences of the different labels in the training- and test-data of the task. The top row represents the English data, the bottom row the French data. Training data is shown in red, test data in blue. **Note:** The test data was released after the task deadline and was not used during training.

ML-ESG-2 presents data in four different languages: English, French, Japanese and Chinese. Because the Japanese and Chinese labels are slightly different than the English and French ones, we chose to only participate in the latter languages. We display information about the training dataset in Figure 1 in red. For English, the training dataset contains 808 news articles while there are 818 news articles for French. Each news article has a news title, a news content and an associated impact type label. Different samples can have the same news title but a different news content and a different impact type, which happens a total of 100 times in the two datasets. For this reason, we didn't use the news title in our approaches.

The impact type is labelled as either "Opportunity" or "Risk". While the data is quite imbalanced for English, with the "Opportunity" label being the majority class, it is approximately even for French. The average length of the news content is 412 characters for English and 565 characters for French, which should fit well within the maximum length of 512 tokens of BERT-based models. Otherwise, the samples are truncated.

2.1 ESG-Aspect Dataset

The previous ML-ESG shared task focused on determining one out of 35 key issues defined by MSCI³, where each key issue belongs to one ESG-aspect (E/S or G) (Chen et al., 2023b). The dataset

³<https://www.msci.com/our-solutions/esg-investing/esg-industry-materiality-map>

shared there consists of 1200 news articles in English and French each, some of which overlap with the news articles used in the dataset we work on here. We present some approaches of using the knowledge available in the ML-ESG-1 dataset to improve the determination of "Opportunity" or "Risk" of the current task in subsection 3.3.

3 Experimental Approach

The basis of our approach consists of mBERT (Devlin et al., 2018), a multilingual BERT-based model. For fine-tuning, we use bottleneck-adapter-modules (Houlsby et al., 2019), which are a small set of weights inserted into the layers of the base-model. In order to account for the specific language we are training in, we implement an approach described in (Pfeiffer et al., 2020b), where a pre-trained language adapter is inserted before the adapter that is being fine-tuned (known as the task-adapter). We use the language adapters from the adapterhub (Pfeiffer et al., 2020a), which are available for all task-languages for the mBERT architecture. Adapter-modules allow us to be more flexible in training our model, as we will describe in subsection 3.2.

To train the adapters, we split the dataset and use 10% as the evaluation set. We use a learning rate of $5e-5$. Finally, to account for the label imbalance in the training data we can observe in Figure 1, we use a weighted loss function.

3.1 Data Augmentation

To augment the data available in the two different languages, we translate both of them to the other language respectively. Since translating the whole text at once sometimes produced artifacts, we split the news content into sentences using nltk's sentence tokenizer (Bird et al., 2009) and translate a single sentence at a time. In order to translate, we use the OPUS-MT models (Tiedemann and Thottingal, 2020) from the huggingface-hub⁴.

3.2 Adapter Fusion

In section 3, we explain our adapter-based approach, where we use a base model, a pre-trained language adapter and finally a task-adapter for fine-tuning. In this configuration, the different components are aligned in series. Adapters can also be used in parallel, which allows for the combination of the task-specific knowledge of each adapter.

⁴<https://huggingface.co/Helsinki-NLP>

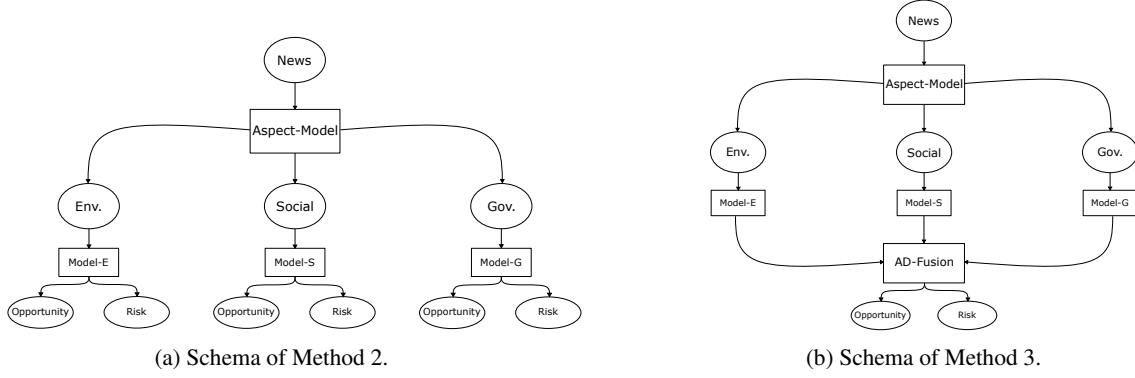


Figure 2: Schemata for Methods 2 and 3. The Aspect Model is obtained using data from ML-ESG-1 (Chen et al., 2023b). The E/S/G-Models in Method 3 are the same ones we train in Method 2.

This is known as Adapter-Fusion. Adapter-Fusion demonstrates particular advantages in scenarios involving limited datasets, as it can place greater emphasis on the knowledge obtained from the task-adapters trained on larger datasets (Pfeiffer et al., 2021). In addition, Adapter-Fusion has shown to lead to performance-improvements in multilingual-scenarios as it can use knowledge from training in other languages (Billert and Conrad, 2023). Since the concept of "Opportunity" and "Risk" can mean different things depending on the context, we aim to use the learnings from ML-ESG-1 (Chen et al., 2023b) in order to train aspect-specific adapters. These adapters will learn more specifically to determine "Opportunity" and "Risk" in the environmental, social and governance context. In practice this means that we will first train a model to determine the rough ESG-aspect of a news-article (environmental, social, governance), before training three different aspect-specific adapters to determine the ESG impact type for articles of each aspect. We achieve an F_1 -Macro score of around 0.86 for English and 0.79 for French for the adapters classifying the ESG-aspect, meaning that there is still the possibility of misclassifying the aspect in the first place, which might be detrimental for the downstream adapters predicting the impact type.

After training the three aspect-specific adapters, it is then possible to train an Adapter Fusion layer to make a final prediction for the ESG impact type.

3.3 Configurations

In order to test the premise that the ESG impact type depends on the aspect (as we describe in the previous subsection), we designed three different configurations.

- Method 1: Train an adapter on the two impact type labels directly.
- Method 2: Determine the ESG-aspect first, then train an adapter for the news articles belonging to each aspect separately in order to determine the impact type.
- Method 3: Use the same approach as in Method 2, but use Adapter-Fusion in order to combine the knowledge of the three separate adapters.

Method 2 and method 3 are depicted in Figure 2. Note that for the governance aspect-adapters, we augment the evaluation data with the governance-evaluation data from the other language respectively, because the evaluation sets have a very small amount of samples.

4 Results

The test data is shown in Figure 1 in blue. For English, the label imbalance in the training set is also present in the test set. For French, while there is still no clear imbalance, the "Risk" label now occurs more often as opposed to the "Opportunity" label in the training set.

In Table 1, we show the F_1 -Macro scores of the various approaches. The three methods we submitted can be seen on the left side of the table. For English, method 1 and method 3 have the same score, with method 2 being much worse than both of them. For French, method 1 shows the best performance, with method 3 being slightly worse. Again, method 2 trails far behind.

Interestingly, method 2 shows a big discrepancy to method 3, even though the same adapters are used, the only difference being the Adapter-Fusion layer.

	Pre-Deadline Results			AD-Fusion Experiments	
Language	Method 1	Method 2	Method 3	w/o gov	Lang-Fusion
EN	<u>0.8098</u>	0.4225	<u>0.8098</u>	0.8771	0.8557
FR	<u>0.7548</u>	0.6169	0.7457	0.7726	0.8084

Table 1: F_1 -Macro scores of the trained adapters for the test-set. On the left side, the submitted results. On the right side, there are additional experiments that were done after the deadline. Bold values represent the best achieved score, while underlined values represent the best submitted score for each language.

4.1 Aspect-Fusion

Since there was not a lot of training data for the governance adapter in methods 2 and 3, we suspect that it is detrimental to the adapter fusion performance. We do several experiments without the governance adapter, using only the environmental and social adapters for the Adapter Fusion. The results are depicted on the right side of Table 1. We can see that the performance surpasses both the previous score of method 3 as well as the score of method 1, which confirms our previous suspicions about the governance adapter.

For English, we note an improvement of almost 7 points while for French, we see an increase slightly lower than 2 points. In both cases, the performance is significantly higher than the previous highest result, demonstrating that the concept of ESG impact type depends on the ESG aspect and that the trained Adapter Fusion layer successfully composes the knowledge learned by our aspect-specific adapters.

4.2 Language-Fusion

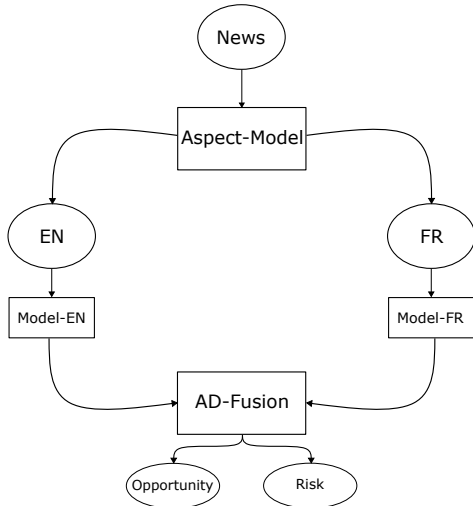


Figure 3: Adapter Fusion on the different language adapters from Method 1.

As mentioned in subsection 3.2, it is also possible to use Adapter Fusion to combine informa-

tion learned from tasks in different languages. To test if this also holds true in our case, we train an Adapter Fusion layer consisting of both language-specific adapters of method 1. The setup is displayed schematically in Figure 3. This is a similar approach as was used in (Billert and Conrad, 2023), although we only work with two languages in this case.

In the rightmost column of Table 1, we can see that this configuration achieves an outperformance of about 5 points over just using the language-specific adapters of method 1. Compared to using the aspect-fusion approach described in subsection 4.1, the results are conflicting. For English, the aspect-fusion shows better results, but for French, using the language-fusion seems superior.

5 Conclusion

In this work, we present multiple approaches using adapters to determine the ESG impact type of news articles. We show that it is possible to gain significant performance increases by using Adapter Fusion, which allows us to combine the knowledge present in the separate adapter modules, by experimenting with implementing Adapter Fusion in two different ways: Firstly, using the prior knowledge about ESG aspects, we train an adapter for each aspect and fuse them to determine the ESG impact type. Secondly, we fuse the adapters trained on the different languages in order to combine the knowledge learned from the different language datasets. Further studies could focus on how to combine the two Adapter Fusions, which we did not attempt in this work due to time constraints. In addition, it might be interesting to use the Japanese and Chinese datasets to try to improve the language-fusion approach even further. Finally, putting more focus on augmenting the training data, e.g. by using paraphrases, could bridge the gap between this study and the top-placed systems.

References

- Tanja Aue, Adam Jatowt, and Michael Färber. 2022. [Predicting Companies' ESG Ratings from News Articles Using Multivariate Timeseries Analysis](#). *arXiv*.
- Florian Berg, Julian F Kölbl, and Roberto Rigobon. 2019. [Aggregate Confusion: The Divergence of ESG Ratings](#). *SSRN Electronic Journal*.
- Fabian Billert and Stefan Conrad. 2023. [HHU at SemEval-2023 Task 3: An Adapter-based Approach for News Genre Classification](#). *Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 1166–1171.
- Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python*. O'Reilly Media Inc.
- Chung-Chi Chen, Yu-Min Tseng, Juyeon Kang, Anaïs Lhuissier, Min-Yuh Day, Teng-Tsai Tu, and Hsin-Hsi Chen. 2023a. [Multi-Lingual ESG Impact Type Identification](#). In *Proceedings of the Sixth Workshop on Financial Technology and Natural Language Processing (FinNLP)*.
- Chung-Chi Chen, Yu-Min Tseng, Juyeon Kang, Anaïs Lhuissier, Min-Yuh Day, Teng-Tsai Tu, and Hsin-Hsi Chen. 2023b. [Multi-Lingual ESG Issue Identification](#). In *Proceedings of the Fifth Workshop on Financial Technology and Natural Language Processing (FinNLP) and the Second Multimodal AI For Financial Forecasting (Muffin)*, pages 111–115, Macao.
- Alexandre Clément, Élisabeth Robinot, and Léo Trepspeuch. 2023. [The use of ESG scores in academic literature: a systematic literature review](#). *Journal of Enterprising Communities: People and Places in the Global Economy*, ahead-of-print(ahead-of-print).
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. [BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding](#). *arXiv*.
- Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. [Parameter-Efficient Transfer Learning for NLP](#). *arXiv*.
- Tahir Islam, Rauf Islam, Abdul Hameed Pitafi, Liang Xiaobei, Mahmood Rehmani, Muhammad Irfan, and Muhammad Shujaat Mubarak. 2021. [The impact of corporate social responsibility on customer loyalty: The mediating role of corporate reputation, customer satisfaction, and trust](#). *Sustainable Production and Consumption*, 25:123–135.
- Maxime L D Nicolas, Adrien Desroziers, Fabio Caccioli, and Tomaso Aste. 2023. [ESG Reputation Risk Matters: An Event Study Based on Social Media Data](#). *arXiv*.
- Jonas Pfeiffer, Aishwarya Kamath, Andreas Rücklé, Kyunghyun Cho, and Iryna Gurevych. 2021. [AdapterFusion: Non-Destructive Task Composition for Transfer Learning](#). *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 487–503.
- Jonas Pfeiffer, Andreas Rücklé, Clifton Poth, Aishwarya Kamath, Ivan Vulić, Sebastian Ruder, Kyunghyun Cho, and Iryna Gurevych. 2020a. [AdapterHub: A Framework for Adapting Transformers](#). *arXiv*.
- Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2020b. [MAD-X: An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer](#). *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7654–7673.
- Hanna Schramm-Klein, Joachim Zentes, Sascha Steinmann, Bernhard Swoboda, and Dirk Morschett. 2016. [Retailer Corporate Social Responsibility Is Relevant to Consumer Behavior](#). *Business & Society*, 55(4):550–575.
- Jörg Tiedemann and Santhosh Thottingal. 2020. [OPUS-MT – Building open translation services for the World](#). *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*.
- Ilze Zumente and Jūlija Bistrova. 2021. [ESG Importance for Long-Term Shareholder Value Creation: Literature vs. Practice](#). *Journal of Open Innovation: Technology, Market, and Complexity*, 7(2):127.
- Ilze Zumente and Natalja Lāce. 2021. [ESG Rating—Necessity for the Investor or the Company?](#) *Sustainability*, 13(16):8940.

5

LLMs FOR FINANCIAL IR

In [chapter 4](#), we presented several methods of training representational language models in order to extract different types of information from textual data. We now want to investigate a different approach: Instead of annotating a dataset in order to train a specialized LM, can we simply use a generative LM and ask it to extract the information of a specific text for us? This would hold several advantages. First of all, annotating samples is very costly. The price of either deploying and using LLMs, or using a closed source LLM has decreased continuously over the past few years, so if the resulting quality of annotations is comparable, it might be reasonable to start with an LLM, and train a smaller, more specialized LM once enough data has been gathered.

Secondly, the context length of LLMs has continued to increase. The most recent models of google’s Gemini family [\[146\]](#) have a context length of 1 million tokens. Compared to that, representational LMs still lag far behind - the original models based on BERT [\[31\]](#) had a token limit of 512. This has increased over the year of 2024, as architectures such as mGTE [\[176\]](#), BGE [\[23\]](#) and ModernBERT [\[161\]](#) can handle up to 8192 tokens. While being much larger than the original amount, this is still too small for many news articles, which means it would be necessary to truncate the article and lose potentially valuable information or working with chunks of the article which might also lose context.

Of course, this approach hinges on the question if the quality of the extracted information is comparable of using more specialized LMs in the first place. This question has been researched extensively ever since ChatGPT was first released more than two years ago [\[105\]](#). Whereas preliminary works found that GPT 3.5 created sub-par annotations [\[32\]](#), more recent research is suggesting that modern generative language models have similar, or even better annotation qualities compared to an average human annotator [\[50\]](#), although they still perform worse than human experts. Even in the financial domain specifically, GPT-4 outperforms non-expert crowd workers [\[8\]](#).

Analyzing News Articles to identify Corporate Sustainability Practices

Paper: Fabian Billert and Stefan Conrad. Nano-ESG: Extracting Corporate Sustainability Information from News Articles. In *Advances in Information Retrieval, 47th European Conference on Information Retrieval, ECIR 2025*, pages 324–338 Lucca, Italy, April 6–10, 2025, Proceedings, Part IV

Personal Contribution: The idea, methodology, dataset and implementation of this paper were developed solely by Fabian Billert. The evaluation guidelines were defined by Fabian Billert, who also prepared the datasets for annotation and set up the annotation pipeline. The paper was written by Fabian Billert under the supervision of Stefan Conrad.

Status: Published

In [section 4.2](#) and [section 4.3](#), we created two separate systems to extract different sustainability-related properties from textual data, more specifically, snippets from news articles. In the following, we explore a unified approach for sustainability-related information extraction using generative LMs. As explained above, a major advantage is the context length - in reality, we are not investigating snippets of text, but complete news articles which are much longer. On the other hand, a downside of these modern systems is that the most performant models (at the time of conducting this research) are often closed-source, meaning each token within a sample costs money¹. Therefore, a major focus of the paper is to reduce the number of articles by applying various filtering methods. For now, we concentrate on German companies listed in the DAX-40 financial index² as of April 2024, as well as a few companies which have recently dropped out. In addition, we only work with English and German web media sources. Lowering processing costs would enable us to include more companies and media in more languages in future analyses.

Although generative LMs have higher initial costs, a multi-task approach significantly reduces expenses, as extracting additional tasks beyond the first incurs only marginal costs, with the primary expense stemming from the news article provided to the LM. For this reason, it makes sense to extract as much information as possible in one pass - if the performance does not degrade so much as to outweigh this benefit. Recent research has investigated if this “single-prompt, multi-task” (SPMT) information extraction leads to decreased performance compared to single-task IE. Gozzi et al. [48] evaluates this with multiple open source LMs and finds mixed results, but no clear winner. Tao et al. [145] on the other hand finds that the SPMT approach achieves a higher accuracy on different questions concerning scientific papers in the medical domain, compared to providing questions to the model separately. They hypothesize that this might be because having access to all the questions at the same time might resemble chain-of-thought prompting [162]. In conclusion, although this might vary on

¹The pricing of OpenAI’s models is available here: <https://platform.openai.com/docs/pricing>

²The components can be seen here: <https://de.wikipedia.org/wiki/DAX#Zusammensetzung>

a case-by-case basis, there is no clear argument against using an SPMT approach.

We want to note that the inherent cost-per-token of closed-source models decreased significantly during the work on this project - at the start, the most powerful available model (GPT-4) had a cost-per-token of 30\$ (60\$) per 1 million input (output) tokens. In the end, by using a newer model (`gpt-4o-2024-05-13`) along with OpenAI's batch-API, this cost went down to 2.5\$ (7.5\$) per 1 million input (output) tokens. So while the focus on cost reduction was certainly necessary at the start of this project, cheaper pricing for new models might allow us to scale up the following research in the future.

Nano-ESG: Extracting Corporate Sustainability Information from News Articles

Fabian Billert^[0009–0005–0311–2176] and Stefan Conrad^[0000–0003–2788–3854]

Heinrich-Heine University of Düsseldorf, Düsseldorf, Germany
 {fabian.billert, stefan.conrad}@hhu.de

Abstract. Determining the sustainability impact of companies is a highly complex subject which has garnered more and more attention over the past few years. Today, investors largely rely on sustainability-ratings from established rating-providers in order to analyze how responsibly a company acts. However, those ratings have recently been criticized for being hard to understand and nearly impossible to reproduce.

An independent way to find out about the sustainability practices of companies lies in the rich landscape of news article data. In this paper, we explore a different approach to identify key opportunities and challenges of companies in the sustainability domain. We present a novel dataset of German and English news articles which were gathered for major German companies between January 2023 and September 2024. By applying a mixture of Natural Language Processing techniques, we first identify relevant articles, before summarizing them and extracting their sustainability-related sentiment and aspect using Large Language Models (LLMs). Furthermore, we conduct an evaluation of the obtained data and determine that the LLM-produced answers are accurate. We release both datasets at <https://github.com/Bailefan/Nano-ESG>.

Keywords: Sustainability · Dataset · News Articles · ESG Extraction · Financial NLP.

1 Introduction

With the growing fear of climate change related issues in the recent past, sustainability has become an important topic not only for the general public, but also in the corporate world [33], where it is coined as “ESG”, standing for Environmental, Social and Governance topics. Those involved in the corporate world, both on the investor and on the company side, desire clarity about how sustainable a company is [14,34].

Third party rating agencies offer products which aim to represent the sustainability of a company in a single value, an “ESG-score”. While convenient for investors, recent criticism has pointed out several problems with this approach. For one, different rating agencies can arrive at a variety of scores for the same company, which confuses users and raises concerns about the derived scores [1]. In addition, the various complex steps involved in determining a single value for

a company make it hard for the end user to truly understand why a company has a certain ESG-rating [6].

With the recent advancements in the field of Natural Language Processing (NLP) and the rise of Large Language Models (LLMs), analyzing vast amounts of data has become more feasible [20,19,27,28]. One textual data source which offers important insights for corporate sustainability practices are news articles. They provide a real-time, unfiltered source of information that reflects public perception and current events related to corporate sustainability practices. Unlike formal reports or company disclosures, which may be delayed or subject to selective reporting, news coverage can capture immediate reactions to a company’s actions, controversies, or initiatives [23,8]. Leveraging NLP techniques in order to extract the ESG-information available in this public data source can enable different stakeholders to track corporate sustainability more transparently, without relying solely on ratings that may obscure the complexity of ESG performance [18]. Furthermore, it is possible to trace a sentiment back to a concrete source and thus evaluate each ESG event on its own if desired. Finally, this approach also enables continuous monitoring, capturing shifts in perception that would otherwise go unnoticed between formal ESG score updates.

To further research in this direction, we analyze a large number of online news articles in order to extract information relevant to a company’s ESG practices in this work. By leveraging multiple modern Machine Learning (ML) models to filter out as much noise as possible, we can process a small amount of news articles with a powerful, but expensive LLM to create a rich dataset we call **Nano-ESG** - named for the level of detail each company can be investigated with. This dataset currently contains companies in the German DAX 40 financial index, as well as a few companies that recently dropped out of it. After determining if a news article is relevant for a given company in the context of sustainability, we summarize it and determine the sentiment as well as which ESG-aspect (meaning Environmental, Social or Governance) it belongs to. To maximize the quality of the final responses, we use GPT-4o¹ to extract them.

Nano-ESG contains news articles from both German and English media sources from January 2023 until September 2024. By further leveraging LLMs, the dataset contains summaries in both German and English. While we cannot release the original news articles themselves for copyright reasons, we release the URLs of each news article along with the concerned company, the discussed summaries, the determined ESG-sentiment and the ESG-aspect. As a result this dataset can be used to train smaller, ESG-specific models in the future. Since the publication date of each article is also included, this is, as far as we know, the first open source dataset containing a time-series of ESG-sentiments and ESG-aspects of companies, which can help practitioners to analyze the correlation of the here detected ESG-events with other data sources and thus enable work on a variety of downstream applications.

For example, multiple works such as [22] and [17] have previously focused on analyzing sustainability reports from companies. These reports are often self-

¹ <https://openai.com/index/hello-gpt-4o/>

published by the companies themselves, which raises the question of whether the content is accurately reported. By using a secondary source of information to verify the reported findings, it might be possible to detect companies which do not report their sustainability practices to satisfaction.

Another indicator for the state of a company is its stock price. Some previous research has analyzed the interdependency of ESG events and price movement (e.g. [8] and [24]). However, the underlying data is typically not publicly available. In addition, the correlation between specific ESG-aspects and the stock price has been mostly neglected until now. **Nano-ESG** offers the possibility for more clarity in this research domain.

While we do not investigate the above mentioned use cases further, we do present a first approach for a deeper analysis through topic detection in section 6. Moreover, in order to be able to judge the reliability of our dataset and the quality of the final responses, we conduct a rigorous evaluation of a subset of the generated data in section 5. Finally, we analyze unsatisfactory responses, examine why they were generated and discuss our findings in section 7.

2 Related Work

Large Language Models for Data Annotation In the last few years the capabilities of both closed and open source LLMs have skyrocketed, with novel LLMs being capable of complex reasoning tasks [27,28,19,26]. As a result, LLMs have become popular as tools to process textual data and extract desired properties in an automatic way [25]. [7] study different approaches in which a GPT-3-based model, `text-davinci-003`, is used to annotate various datasets. The authors come to the conclusion that GPT-3, although still lacking behind human annotators, shows potential in performance while being significantly cheaper. [16] investigates the performance of the improved GPT-4 in annotating low-resource languages and finds impressive results of the LLM, breaking even or outperforming human annotators in many cases. [12] compares GPT-4 with an elaborate Amazon Mechanical Turk (MTurk) setup to annotate a subset of the CODA-19 dataset [13] and finds that GPT-4 achieves a higher performance. [29] comes to a similar conclusion when analyzing annotations of political twitter messages from [30] between GPT-4, MTurk workers and Experts.

Datasets for ESG-Research Datasets in the ESG domain are scarce, but have been gaining interest recently. The FinNLP series of workshops² has published multiple ESG-centric datasets and organized shared tasks for researchers. They have focused on identifying ESG-aspects [2], ESG-sentiments [3] and ESG impact duration [4] and training specialized models for these tasks. However, these datasets are usually fairly small and do not focus on specific companies, making them unsuitable for further analysis of company sustainability practices.

A more extensive dataset was created by [9]. Here, the authors gathered more than 450000 headlines from the twitter account of *The Guardian*³, checked the

² <https://aclanthology.org/venues/finnlp/>

³ <https://www.theguardian.com/>

tags for ESG-relation and mapped relevant headlines to ESG-aspects manually. They then trained models to classify the Relevancy, ESG-aspect and ESG-sentiment of the headlines. While being a fairly big dataset, the downsides here are that it is constrained to one source and only analyzes the headlines.

[21] improves upon this by creating a corpus of nearly 4000 news articles about 10 companies in the FTSE financial index with ESG-relevance as well as ESG-aspects labels. This work uses the whole article content for their classification approaches, but is not as extensive as the previous approach.

Notably, both of the two former datasets do not include timestamps, making it impossible to analyze the therein detected ESG-events in context with other external factors. For this reason, we believe that **Nano-ESG** can help to further sustainability research in a variety of directions.

3 Data Processing Pipeline

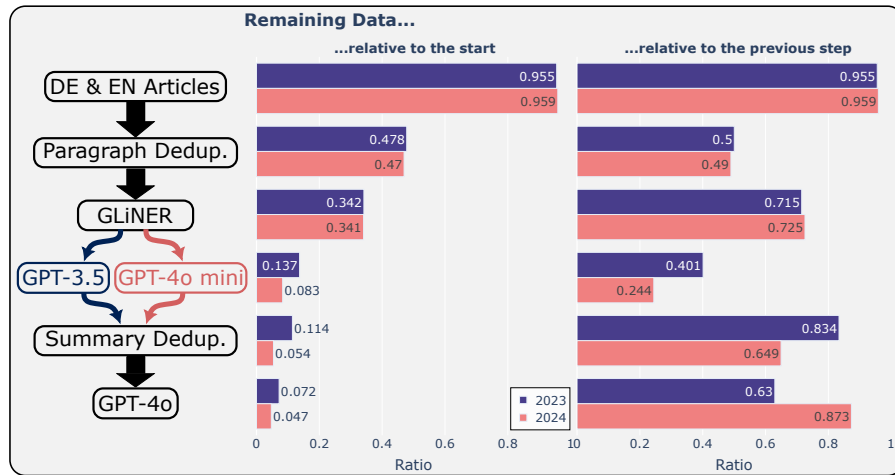


Fig. 1. Processing-pipeline (left) used for the creation of the dataset and how much each step reduces the amount of data relative to the starting point (middle) and relative to the previous step (right).

In this section, we describe the pipeline used to obtain the final information contained in the **Nano-ESG** dataset. To reduce the cost of processing the articles with the more expensive LLM in the last step as much as possible, we apply different techniques to filter out articles unlikely to be relevant before then. A schema of our workflow along with how much of the initial data is filtered out (compared to the total at the start as well as each previous step) is displayed in Figure 1. As the data processing pipeline for 2024 differs slightly from the one for 2023, we display them separated in two colors in the graph.

3.1 Raw Data

The original data is obtained by crawling the top 100 English and German online media websites (meaning 200 in total). Social media sources such as Twitter and Facebook are not considered. As this dataset is currently constrained to German companies, we defined a list of keywords⁴ for each company in the German DAX 40 financial index and search for articles which contain any of these keywords. After this, we are left with around 850,000 articles for the considered timeframe from January 2023 until the 16th of September 2024. Some of the crawled articles were in other languages - we decided to proceed with only English and German articles for simplicity and lower costs.

3.2 Paragraph Deduplication

During the initial data exploration we noticed that online sources often publish similar articles for a specific event. In order to condense the displayed information, we employ an embedding model to find similar articles in a set timeframe. We use a bilingual embedding model which was trained on English and German data [15], `jina-embeddings-v2-base-de`⁵ to embed the paragraphs of the articles which contain any one of the company-specific keywords. We then group together articles published within a week of each other for which we calculate a cosine-similarity of 0.8 or more. We decided on this value after experimenting with different cutoffs for various groups of articles. Once we find articles similar to one another, we keep the first article in the dataset and discard the following articles. This step allows us to reduce the amount of articles by around 50%.

3.3 Entity-based Filtering

Some companies have associated keywords which can also refer to unrelated information. To deal with these cases, we use a multilingual zero-shot named entity recognition model called GLiNER, more specifically `gliner_multi-v2.1` [32]. To filter out articles which do not refer to any of the companies we are interested in (for example where “Allianz” refers to a pact between entities rather than the German company), we apply this model to the sentences in which we detected a keyword. We discard the article if none of the keywords refer to an organization or company.

3.4 Filter Language Model

As depicted in Figure 1 in the middle plot, the previously described techniques can reduce the amount of data by around two thirds. However, in order to reduce costs further, we decided to first use a relatively cheap language model to filter out irrelevant articles with regards to either the target company or ESG thematics. We chose to use OpenAI’s `gpt-3.5-turbo-0125` for its ease of use and competitive pricing. While we were working on processing the data, a

⁴ See `company_info.py` in our Github repository for the full list

⁵ <https://huggingface.co/jinaai/jina-embeddings-v2-base-de>

more performant and cheaper model, `gpt-4o-mini-2024-07-18` was released, so we switched to this model instead. The result is that for this step, data from 2023 was processed with GPT-3.5-turbo, and data from 2024 was processed with GPT-4o-mini.

The prompt we use is included in the supplementary material. In an attempt to force the model to reflect about its answer similar to a chain-of-thought approach [31], we ask it to first generate a first answer for the relevancy of the article. We then prompt it to explain its answer as well as generate a summary for the article, which we will use in a later step. We noticed that the model often inferred a relevance for ESG issues from loosely related topics (e.g. reasoning that a change of the stock price can have effects on the governance of the company), so we also ask the model to determine if ESG issues are being directly addressed in the article. After this, we let the model determine the relevancy again and refer it to its own answers for the explanation, summary and the direct ESG connection. Finally, we continue with all articles which directly address ESG issues and were judged to be **relevant**. This step manages to reduce the amount of data the most - by 60% when using GPT-3.5-turbo in 2023 and by more than 75% with GPT-4o-mini in 2024. Although the underlying data is different for the two models, it seems that the latter LLM is more adept at determining relevant articles.

3.5 Summary Deduplication

At this point, we still have some articles in our dataset which are quite similar to one another. To further reduce the cost of the final determination of each article's properties, we use the summary generated in the previous step in order to remove similar articles with the same approach as in subsection 3.2. Again, we can observe a significant difference as to how much data is filtered out by this step - for 2024 nearly 20% more articles are removed here compared to 2023. A reason for this could be the worse relevancy-detection in the previous step by the older model, allowing more irrelevant articles of varying topics to still be in our dataset at this point.

3.6 Final Determination Model

In the final step we decide to use a more powerful language model to create high quality data which can potentially be used in order to train small, but specialized models in the future. We chose the model based on the chatbot arena [5], which evaluates Large Language Models based on human preference. When we were working on this step, the highest performing model there was the GPT-4o model from OpenAI, more specifically `gpt-4o-2024-05-13` - all final responses are generated with this model. At the time of writing this, a few other models perform better, but the version of GPT-4o we used still places in the top ten.

In this step we again ask the model to summarize the article regarding ESG-topics about the target company as well as determine if the article is relevant

or not. If it is, we also ask for the ESG-sentiment, meaning the sentiment of the ESG-information for the target company, and the ESG-aspect. Again, the full prompt used in this step is included in the supplementary material. Although we instruct the model to reply in German, we decided to formulate the prompt in English since this results in less tokens and OpenAI uses a pay-per-token pricing model. To the best of our knowledge, this does not degrade the performance. Finally, in order to make sure the generated summaries do not violate German copyright law, we shorten and rephrase them.

This step is the most expensive step of our pipeline, so our goal was to have as few irrelevant articles remaining as possible. As shown on the right in Figure 1, the newer GPT-4o-mini left significantly less irrelevant articles in the dataset compared to the older model - only around 13% of the articles processed by GPT-4o were deemed to be irrelevant.

4 Nano-ESG

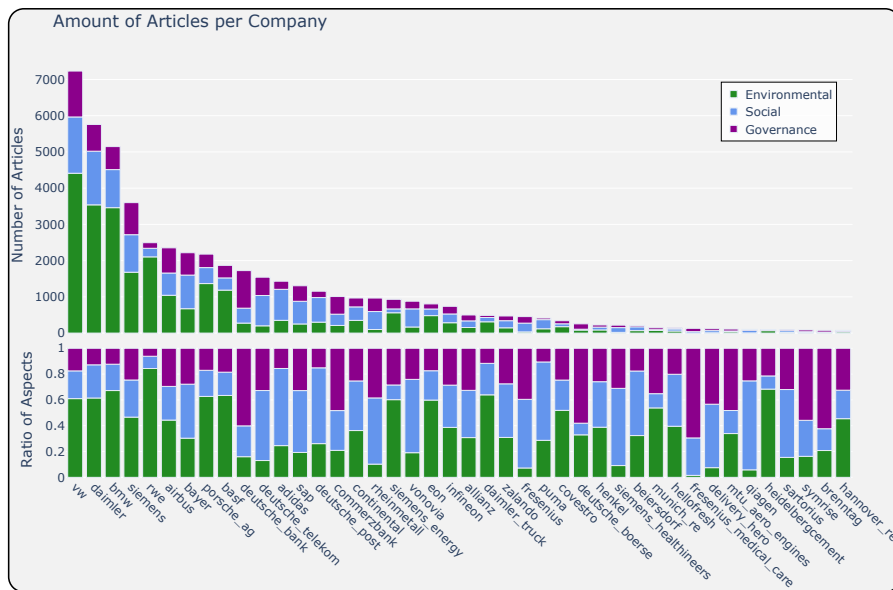


Fig. 2. Number of articles with different ESG-aspects per company. Top: total number. Bottom: Ratio of each aspect per company.

After analyzing roughly 850,000 articles, the remaining dataset consists of 51,087 relevant articles with ESG-information. Each article has a timestamp of publication, a URL as well as the fields mentioned in subsection 3.6. In addition, we also asked the model to score the relevancy of the article and determine a list

Table 1. Top row: The total amount of articles for each sentiment and ESG-aspect. Bottom row: The ratio of articles per sentiment and ESG-aspect averaged for all companies.

	Negative	Neutral	Positive	E	S	G
Total	17668	6115	27304	24546	15086	11455
Avg. Perc. per Company	0.3274	0.1290	0.5436	0.3574	0.3430	0.2995

of keywords it deems important for each article. We will investigate these fields in section 6. Finally, we also provide an English version of the summary which was created by translating with `gpt-4o-mini-2024-07-18`.

Each article is associated with one company. An article can appear multiple times in the final dataset, but the summary differs as it is always composed with regards to a single target company. Table 1 shows the total number of articles categorized by sentiment and ESG-aspect, along with the distribution of these categories averaged over all companies. While the distribution of ESG-aspects per company is approximately similar, the amount of neutral articles is significantly lower than negative and positive articles. At the same time, as is shown in the top part of Figure 2, it is worth noting that the total amount of articles between companies varies significantly. The fewest ESG-relevant articles we find for a company is 77 articles for Hannover Re, while for the company with the most articles (Volkswagen) we find 7233 relevant articles. In the bottom part of the figure we display the ratio of articles for each company which belong to each aspect. As can be suspected from the bottom row of Table 1, the companies with the most articles in general all have significantly more environmental articles than articles for other ESG-aspects.

5 Data Evaluation

In order to assess the quality of the generated results, we evaluated a subset of the data manually. We asked five sustainability experts to annotate the data and used Argilla as our annotation-interface⁶. The results are evaluated in two steps: First, we check if the summaries were correctly created by GPT 4-o. In a second, separate task, we evaluate the sentiment and aspect determined by the LLM. The instructions for the the two annotation tasks are included in the supplementary material.

5.1 Evaluation Dataset

To create a dataset to evaluate the summarization task, we randomly picked one article for each (company, sentiment) group, ending up with 123 summaries. 38 of these samples, roughly one per company, were assigned to all annotators.

For the more complex second task we decided to evaluate more samples. In order to pick diverse summaries, we applied k-means clustering (with $n = 3$)

⁶ <https://docs.argilla.io/latest/>

to each (company, sentiment) subgroup using their embeddings and chose the nearest article from each cluster center. In total, 367 samples were evaluated. Of these, 82 articles (two per company, with approximately the same number of articles for all sentiments) were shown to all annotators.

5.2 Summary Evaluation

We first analyze the samples all annotators evaluated. For 35 of the 38 samples, all annotators agreed on the correctness of the summary. In the other three cases, only one annotator disagreed. After consulting with the annotators, it became clear that the problem for the three cases lies in the article not containing any relevant ESG-information for the target company in the first place. Looking at all samples, in 95.9% of the cases all annotators determined the summary to be **correct**. We conclude that the LLM generally is able to correctly summarize the article according to the instructions, but can struggle to assess its relevancy. We investigate this further in the second evaluation task.

5.3 Classification Evaluation

We discuss the different fields that were labeled separately:

Relevancy: Of the articles evaluated by all five sustainability experts, 68.7% are deemed as relevant by all annotators. For 83.1% of the samples, a majority of 3/5 annotators decided the sample was relevant. Finally, 98.8% of the samples were determined as relevant by at least one annotator.

Sentiment: Evaluating the sentiment raised some problems. Some annotators marked all sentiments appearing in parts of the summary while others simply marked the most relevant sentiment. We decided to simplify the sentiments according to the following rules: If an annotator selected **neutral** along with either **positive** or **negative**, we assign the non-**neutral** value. If an annotator selected both **positive** and **negative**, we set the sentiment to **neutral**, which we also do if they were **not sure**. We then measure Fleiss' kappa [10] for all samples all annotators deemed relevant to be $\kappa = 0.818$. From these findings it seems that the task of determining the ESG-sentiment for a summary is not straight-forward, but the annotators had a generally reasonable agreement as to the sentiment of a sample. The final sentiment was determined by taking the value picked by most annotators or following a similar approach as above if the values are tied. We then measure an accuracy of 79.9% of the LLM annotations compared to the the user annotations.

Aspects: We further investigate the aspect evaluations. We again calculate Fleiss' kappa by taking all samples all annotators found relevant and arrive at a value of $\kappa = 0.427$. Annotators evidently found this task more difficult to judge than the sentiment. To determine the aspect the annotators most agree about for a sample, we first look at the aspect most selected among all annotators. If there is a draw, we look at the aspect the annotators deemed the most relevant individually and select the most selected aspect from there. From this, we calculate an accuracy of 78.5% for the aspect determined by the LLM compared to

the one determined by the annotators. Notably, the most disagreement appeared between samples the LLM determined as **social** while the annotators selected **governance**.

6 Results

6.1 Development of the ESG-Aspects

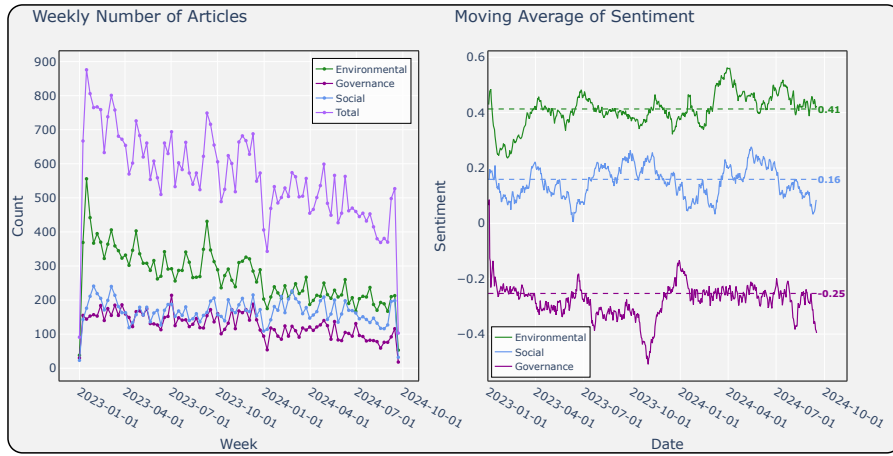


Fig. 3. Left: Weekly number of articles in total and for each ESG-aspect. Right: 30-day moving average of the sentiment per ESG-aspect.

Over the 21-month period investigated in this work, the total number of ESG-relevant news articles we detect for the observed companies generally decreases. As we can see in Figure 3 on the left, which shows the weekly number of news articles for each ESG-aspect, this decrease can mostly be attributed to the environmental and governance aspects while the number of social articles remains approximately stable.

In the right figure of Figure 3, we first calculate the daily mean sentiment and then apply a 30-day moving-average for each ESG-aspect. We find that the environmental aspect is generally the most positive with a mean sentiment of 0.41. Articles with a social ESG-aspect are only slightly positive at a mean sentiment of 0.16. Overall, governance is the only consistently negative aspect during our observed timeframe, with a mean sentiment of -0.25 .

6.2 Can LLMs Judge Degrees of Relevance?

Since we also prompted the LLM to provide a relevance-score between 1 and 10 for each article, we investigate here whether the returned score is a meaningful

Table 2. Top: Amount of articles for each relevance score in the full dataset. Bottom (two rows): Articles deemed irrelevant by at least one annotator per Relevance Score as judged by the LLM (irrelevant/total articles and associated ratio).

Relevance Score	2	3	4	5	6	7	8	9	10
All Data	2	122	983	1380	10465	18219	17113	2780	23
Eval - Total	-	-	3/6	3/8	33/63	19/101	7/141	1/50	-
Eval - Ratio	-	-	0.50	0.38	0.52	0.19	0.05	0.02	-

measure of relevancy. The top row in Table 2 displays the total amount of articles for each relevance score as assigned by GPT-4o. Most of its responses are centered around the score of 7 - the overall mean relevance score is 7.12. The model only rarely responds with low scores - values of 1 are never assigned, and there are just two cases of 2. Similarly, the highest score of 10 is only given to 23 articles.

The bottom two rows of Table 2 show the annotation results in the evaluation set. In the middle row we show how often at least one annotator deems an article irrelevant out of the total number of articles present in the evaluation set. The corresponding ratio in the bottom row indicates that articles with a lower relevance scores are more likely to be determined as irrelevant. At the same time, higher relevance scores are rarely determined as irrelevant - for a score of 9, there is only one out of 50 annotated articles with this label. We conclude from these findings that the LLM is generally able to determine to which degree an article is relevant in the here investigated context of ESG topics.

6.3 Detecting Relevant Topics



Fig. 4. Amount of positive and negative articles regarding the “Forced Labor” topic of Volkswagen over time for the different detected aspects.

We use **Bertopic**, a popular library for topic modeling, to identify relevant topics for each company [11]. **Bertopic** identifies topics through the following steps: First, the dimensionality of the embeddings of the underlying documents

is reduced before their clusters are detected. To determine good representations for the identified clusters, the default algorithm applies a bag-of-words approach for each cluster and then uses a modified TF-IDF technique to create a set of representing words per cluster.

In order to encourage the detection of sustainability-specific topics, we decided to augment the article summaries with the keywords the LLM determined in subsection 3.6 while removing mentions of the companies which are sometimes included in them. We simply append the keywords to the end of the summaries and use the same embedding model as in subsection 3.2 to embed the resulting text.

In the previous subsection, we identified the relevance score as a meaningful measure for the degree of relevancy of the articles. By calculating the mean relevance score of each topic cluster, we can thus determine relevant topics for the companies, as well as check if a topic is more likely to represent an opportunity or a risk through the mean sentiment. In addition, by looking at the amount of detected news articles for topics over time, it is possible to identify when critical events take place and how companies attempt to address them. Taking Volkswagen as an example, one of its most relevant topics we detected concerns reports of forced labor in China’s Xinjiang province taking place in plants of the company. As shown in Figure 4, this is a generally negative topic for the company in 2023. In the end of 2023, an audit is conducted in its plants and finds no problems. However, shortly afterwards, in February 2024, new reports emerge, showing that the problem persists with Chinese partners of Volkswagen.

In this way, it is possible to explore different topics of the covered companies and even contextualize them with other real world events⁷.

7 Limitations & Error Analysis

During our evaluation, we identified several limitations in our current approach. One significant issue is the inability to differentiate conflicting sentiments within a single news article. Currently, we only capture an overarching sentiment, meaning that if multiple perspectives exist, we may only capture one, thus omitting valuable nuance.

Additionally, the LLM sometimes struggles to isolate ESG sentiment pertaining specifically to the target company, confusing it with the overall tone of the article. This is particularly problematic for reinsurance companies impacted by increased natural disasters. Such articles might be classified as (negative, environmental), which does not necessarily reflect the sentiment relevant to the reinsurance firm itself.

Another challenge is handling articles about multiple related companies. For instance, news about both **Siemens** and its subsidiary **Siemens Energy** may be marked as relevant to both companies, even when the article predominantly

⁷ A notebook in the supplementary material allows users to analyze the different topics of our covered companies further on their own.

concerns one. Despite attempts to address this by including related companies in the prompt, this approach has had limited success.

8 Conclusion & Future Work

We present **Nano-ESG**, a novel dataset containing corporate sustainability information extracted from news articles. By determining an ESG-sentiment and ESG-aspect for each relevant article, we are able to provide an alternative way to investigate corporate sustainability events - we present one approach for this through the detection of relevant topics in subsection 6.3. Our pipeline for information extraction is described in detail in this work and can be applied to other domains.

We will continue to extend the work on this project by regularly adding more data to monitor the state of German corporate sustainability practices. In the future, we will aim to build an interactive application to allow users to conveniently access up-to-date information in this domain. In addition, we would like to extend the covered companies to more countries beyond Germany. To make this second goal feasible, it might be necessary to train more specialized models for the different data extraction steps, which should be possible due to the high-quality data gathered in this work. This would allow us to reduce costs further and thus increase the amount of data we can continuously monitor.

Acknowledgments. We want to thank GET Capital AG for the support in infrastructure they provided.

References

1. Berg, F., Kölbel, J.F., Rigobon, R.: Aggregate Confusion: The Divergence of ESG Ratings. SSRN Electronic Journal (2019). <https://doi.org/10.2139/ssrn.3438533>
2. Chen, C.C., Tseng, Y.M., Kang, J., Lhuissier, A., Day, M.Y., Tu, T.T., Chen, H.H.: Multi-Lingual ESG Issue Identification. In: Proceedings of the Fifth Workshop on Financial Technology and Natural Language Processing (FinNLP) and the Second Multimodal AI For Financial Forecasting (Muffin). pp. 111–115. Macao (8 2023), <https://aclanthology.org/2023.finnlp-1.11>
3. Chen, C.C., Tseng, Y.M., Kang, J., Lhuissier, A., Seki, Y., Day, M.Y., Tu, T.T., Chen, H.H.: Multi-Lingual ESG Impact Type Identification. Proceedings of the Sixth Workshop on Financial Technology and Natural Language Processing pp. 46–50 (2023). <https://doi.org/10.18653/v1/2023.finnlp-2.6>
4. Chen, C.C., Tseng, Y.M., Kang, J., Lhuissier, A., Seki, Y., Lee, H., Day, M.Y., Tu, T.T., Chen, H.H.: Multi-Lingual ESG Impact Duration Inference. Proceedings of the Joint Workshop of the 7th Financial Technology and Natural Language Processing, the 5th Knowledge Discovery from Unstructured Data in Financial Services, and the 4th Workshop on Economics and Natural Language Processing @ LREC-COLING 2024 pp. 219–227 (5 2024), <https://aclanthology.org/2024.finnlp-1.22>

5. Chiang, W.L., Zheng, L., Sheng, Y., Angelopoulos, A.N., Li, T., Li, D., Zhang, H., Zhu, B., Jordan, M., Gonzalez, J.E., Stoica, I.: Chatbot Arena: An Open Platform for Evaluating LLMs by Human Preference. *arXiv* (2024). <https://doi.org/10.48550/arxiv.2403.04132>
6. Clément, A., Robinot, E., Trespeuch, L.: The use of ESG scores in academic literature: a systematic literature review. *Journal of Enterprising Communities: People and Places in the Global Economy* **ahead-of-print**(ahead-of-print) (2023). <https://doi.org/10.1108/jec-10-2022-0147>
7. Ding, B., Qin, C., Liu, L., Chia, Y.K., Li, B., Joty, S., Bing, L.: Is GPT-3 a Good Data Annotator? Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) pp. 11173–11195 (2023). <https://doi.org/10.18653/v1/2023.acl-long.626>
8. Dorfleitner, G., Zhang, R.: ESG News Sentiment and Stock Price Reactions: A Comprehensive Investigation via BERT. *Schmalenbach Journal of Business Research* **76**(2), 197–244 (2024). <https://doi.org/10.1007/s41471-024-00185-3>
9. Fischbach, J., Adam, M., Dzhagatspanyan, V., Mendez, D., Frattini, J., Kosenkov, O., Elahidoost, P.: Automatic ESG Assessment of Companies by Mining and Evaluating Media Coverage Data: NLP Approach and Tool. 2023 IEEE International Conference on Big Data (BigData) **00**, 2823–2830 (2023). <https://doi.org/10.1109/bigdata59044.2023.10386488>
10. Fleiss, J.L.: Measuring nominal scale agreement among many raters. *Psychological Bulletin* **76**(5), 378–382 (1971). <https://doi.org/10.1037/h0031619>
11. Grootendorst, M.: BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv* (2022). <https://doi.org/10.48550/arxiv.2203.05794>
12. He, Z., Huang, C.Y., Ding, C.K.C., Rohatgi, S., Huang, T.H.K.: If in a Crowdsourced Data Annotation Pipeline, a GPT-4. *arXiv* (2024). <https://doi.org/10.1145/3613904.3642834>
13. Huang, T.H.K., Huang, C.Y., Ding, C.K.C., Hsu, Y.C., Giles, C.L.: CODA-19: Using a Non-Expert Crowd to Annotate Research Aspects on 10,000+ Abstracts in the COVID-19 Open Research Dataset. *arXiv* (2020). <https://doi.org/10.48550/arxiv.2005.02367>
14. Islam, T., Islam, R., Pitafi, A.H., Xiaobei, L., Rehmani, M., Irfan, M., Mubarak, M.S.: The impact of corporate social responsibility on customer loyalty: The mediating role of corporate reputation, customer satisfaction, and trust. *Sustainable Production and Consumption* **25**, 123–135 (8 2020). <https://doi.org/10.1016/j.spc.2020.07.019>
15. Mohr, I., Krimmel, M., Sturua, S., Akram, M.K., Koukounas, A., Günther, M., Mastrapas, G., Ravishankar, V., Martínez, J.F., Wang, F., Liu, Q., Yu, Z., Fu, J., Ognawala, S., Guzman, S., Wang, B., Werk, M., Wang, N., Xiao, H.: Multi-Task Contrastive Learning for 8192-Token Bilingual Text Embeddings. *arXiv* (2024). <https://doi.org/10.48550/arxiv.2402.17016>
16. Nasution, A.H., Onan, A.: ChatGPT Label: Comparing the Quality of Human-Generated and LLM-Generated Annotations in Low-Resource Language NLP Tasks. *IEEE Access* **12**, 71876–71900 (2024). <https://doi.org/10.1109/access.2024.3402809>
17. Ni, J., Bingler, J., Colesanti-Senni, C., Kraus, M., Gostlow, G., Schimanski, T., Stambach, D., Vaghefi, S.A., Wang, Q., Webersinke, N., Wekhof, T., Yu, T., Leippold, M.: CHATREPORT: Democratizing Sustainability Disclosure Analysis through LLM-based Tools. Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations pp. 21–51 (2023). <https://doi.org/10.18653/v1/2023.emnlp-demo.3>

18. Oliveira, U.R.d., Menezes, R.P., Fernandes, V.A.: A systematic literature review on corporate sustainability: contributions, barriers, innovations and future possibilities. *Environment, Development and Sustainability* **26**(2), 3045–3079 (2024). <https://doi.org/10.1007/s10668-023-02933-7>
19. OpenAI: GPT-4 Technical Report. arXiv (2023)
20. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C.L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., Lowe, R.: Training language models to follow instructions with human feedback. arXiv (2022). <https://doi.org/10.48550/arxiv.2203.02155>
21. Pavlova, M., Casey, B., Wang, M.: ESG-FTSE: A corpus of news articles with ESG relevance labels and use cases. arXiv (2024). <https://doi.org/10.48550/arxiv.2405.20218>
22. Rouen, E., Sachdeva, K., Yoon, A.: Sustainability Meets Substance: Evaluating ESG Reports in the Context of 10-Ks and Firm Performance. SSRN Electronic Journal (2024). <https://doi.org/10.2139/ssrn.4227934>
23. Salmones, M.d.M.G.d.l., Herrero, A., Martínez, P.: Determinants of Electronic Word-of-Mouth on Social Networking Sites About Negative News on CSR. *Journal of Business Ethics* **171**(3), 583–597 (2021). <https://doi.org/10.1007/s10551-020-04466-9>
24. Serafeim, G., Yoon, A.: Stock price reactions to ESG news: the role of ESG ratings and disagreement. *Review of Accounting Studies* **28**(3), 1500–1530 (2022). <https://doi.org/10.1007/s11142-022-09675-3>
25. Tan, Z., Beigi, A., Wang, S., Guo, R., Bhattacharjee, A., Jiang, B., Karami, M., Li, J., Cheng, L., Liu, H.: Large Language Models for Data Annotation: A Survey. arXiv (2024). <https://doi.org/10.48550/arxiv.2402.13446>
26. Team, L.: The Llama 3 Herd of Models (7 2024), <https://ai.meta.com/research/publications/the-llama-3-herd-of-models/>
27. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., Lample, G.: LLaMA: Open and Efficient Foundation Language Models. arXiv (2023). <https://doi.org/10.48550/arxiv.2302.13971>
28. Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Ferrer, C.C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., Fuller, B., Gao, C., Goswami, V., Goyal, N., Hartshorn, A., Hosseini, S., Hou, R., Inan, H., Kardas, M., Kerkez, V., Khabsa, M., Kloumann, I., Korenev, A., Koura, P.S., Lachaux, M.A., Lavril, T., Lee, J., Liskovich, D., Lu, Y., Mao, Y., Martinet, X., Mihaylov, T., Mishra, P., Molybog, I., Nie, Y., Poulton, A., Reizenstein, J., Rungta, R., Saladi, K., Schelten, A., Silva, R., Smith, E.M., Subramanian, R., Tan, X.E., Tang, B., Taylor, R., Williams, A., Kuan, J.X., Xu, P., Yan, Z., Zarov, I., Zhang, Y., Fan, A., Kambadur, M., Narang, S., Rodriguez, A., Stojnic, R., Edunov, S., Scialom, T.: Llama 2: Open Foundation and Fine-Tuned Chat Models. arXiv (2023). <https://doi.org/10.48550/arxiv.2307.09288>
29. Törnberg, P.: ChatGPT-4 Outperforms Experts and Crowd Workers in Annotating Political Twitter Messages with Zero-Shot Learning. arXiv (2023). <https://doi.org/10.48550/arxiv.2304.06588>
30. Vliet, L.v., Törnberg, P., Uitermark, J.: The Twitter parliamentary database: Analyzing Twitter politics across 26 countries. *PLoS ONE* **15**(9), e0237073 (2020). <https://doi.org/10.1371/journal.pone.0237073>

-
31. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., Zhou, D.: Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. arXiv (2022). <https://doi.org/10.48550/arxiv.2201.11903>
 32. Zaratiana, U., Tomeh, N., Holat, P., Charnois, T.: GLiNER: Generalist Model for Named Entity Recognition using Bidirectional Transformer. arXiv (2023). <https://doi.org/10.48550/arxiv.2311.08526>
 33. Zumente, I., Bistrova, J.: ESG Importance for Long-Term Shareholder Value Creation: Literature vs. Practice. *Journal of Open Innovation: Technology, Market, and Complexity* **7**(2), 127 (2021). <https://doi.org/10.3390/joitmc7020127>
 34. Zumente, I., Lăce, N.: ESG Rating—Necessity for the Investor or the Company? *Sustainability* **13**(16), 8940 (2021). <https://doi.org/10.3390/su13168940>

6

FROM SENTIMENTS TO ASSET REPRESENTATIONS

In the previous two chapters, we showed different approaches for extracting financial information from news articles. In [subsection 2.3.1](#) we investigated if the consumer sentiment found in news articles can be used as an indicator for future financial returns. Various studies [\[122, 157\]](#) find that there is a significant correlation in many cases so that including the sentiment for investment predictions can yield better returns in portfolios compared to relying on other indicators alone [\[72, 84\]](#). In the following works, we build on this research and rely on the sentiment alone to steer a universe of single assets without taking other variables (such as previous price development or macro-economic factors) into account.

We briefly want to introduce some of the different types of “asset representations” that we will talk about in the following sections.

Index-Based Representations Indices generally are non-investible constructs that track the performance of a specific group of assets. A group of assets could be the biggest german companies (DAX-40), the biggest german technological companies (TecDAX) or the biggest companies in the world (MSCI-World). The assets within the index are weighted based on the methodology of the index. Examples for weighting schemata are market capitalization-based (e.g. DAX-40, S&P500), price-based (Nikkei 225, Dow Jones Industrial Average) or simply equal-weight versions where every stock has the same weight [\[14\]](#).

Passive Funds (ETFs) Passive funds such as ETFs generally build on top of Indices and try to replicate their performance in an investible vehicle [\[43\]](#). The weights of the single stocks in the universe are changed according to the changes within the tracked index, without any active decisions to optimize performance.

Actively Managed Funds (Portfolios) As opposed to ETFs, actively managed portfolios try to outperform a specific benchmark, which could simply be an index, by optimizing the weights of the constituents based on the investment decisions of a fund manager. Contrary to expectations, research suggests that actively managed funds often do not outperform passive ETFs [38]. According to S&P's last SPIVA report at the end of 2024, 65% of active large-cap U.S. equity funds underperformed the S&P500 in that year [62]. Over the 15 prior years, 90% of active managers performed worse than the S&P500.

In the following sections we will investigate various approaches for asset allocations based on different signals.

6.1 Building a BERT-based Sentiment Performance Index

Paper: Fabian Billert and Stefan Conrad. A Framework for the Construction of a Sentiment-Driven Performance Index: The Case of DAX40. In *Proceedings of the AAAI-2023 Workshop on Multimodal AI for Financial Forecasting*, Washington DC, USA.

Remark: The paper was presented at the Workshop, but the Workshop did not publish the proceedings until now. For this reason, the camera-ready version of the paper was uploaded to ArXiv instead.

Personal Contribution: The research idea was developed solely by Fabian Billert. All written parts of the paper were authored by Fabian Billert with the supervision of Stefan Conrad.

Status: Published

We first elaborate on index-based representations and attempt to construct a novel type of index based on the DAX-40 universe of stocks in this section. The general idea here is that instead of weighting the constituents by their market capitalization (which is the normal method the DAX-40 follows), we weigh them by their consumer sentiment.

Sentiment-based indices have gained popularity in recent years, as many index providers have created suites of sentiment-related products. MSCI, for example, launched a sentiment index in April 2025 using an “analyst sentiment score” which is based on different analyst forecasts for stocks, such as how much of them predict an upward trend vs a downward trend, or a revision of the targeted price of the stocks [98, 99]. S&P instead relies on Ravenpack¹ as their data provider and leverages news articles in order to determine the index composition [140, 139] - S&P manages this index since March 2024.

In contrast to the abundance of available products, studies on sentiment-based indices remain scarce. While many works focus on individual components, the construction of sentiment-based performance indices is virtually unexplored. In the following paper, we aim to bridge this gap and present our approach for building such an index.

¹<https://www.ravenpack.com/>

A Framework for the Construction of a Sentiment-Driven Performance Index: The Case of DAX40

Fabian Billert^{1,2}, Stefan Conrad¹,

¹ Heinrich-Heine University of Düsseldorf

² GET Capital AG

Fabian.Billert@uni-duesseldorf.de, Stefan.Conrad@uni-duesseldorf.de

Abstract

We extract the sentiment from german and english news articles on companies in the DAX40 stock market index and use it to create a sentiment-powered pendant. Comparing it to existing products which adjust their weights at pre-defined dates once per month, we show that our index is able to react more swiftly to sentiment information mined from online news. Over the nearly 6 years we considered, the sentiment index manages to create an annualized return of 7.51% compared to the 2.13% of the DAX40, while taking transaction costs into account. In this work, we present the framework we employed to develop this sentiment index.

Introduction

Advances in natural language processing in the last few years have led to an increasing amount of research on extracting information from news, social media and other sources (Min et al. 2021). This allowed researchers to start explaining existing phenomena from a new point of view. In a financial context, one of the most promising and now more easily available features has been the investor sentiment, which expresses the opinion of investors towards different financial instruments (Ferreira, Gandomi, and Cardoso 2021), (Li et al. 02.2018). Recent work has found that the sentiment is an important factor for predicting the future market development ((Wan et al. 2021), (Sidogi, Mbuva, and Marwala 2021), (Ranco et al. 2015), (Koratamaddi et al. 2021), (Allen, McAleer, and Singh 2019), (Feng et al. 2021)), although the extent of its impact is not clear. As a result, researchers are still investigating how to optimally use this new information.

Some works use the sentiment as a feature along with other, more conventional features such as historical returns and volatility to improve the prediction of the return of a stock on a specific day. This is often done by using complex neural networks in a time-series prediction approach ((Ko and Chang 2021), (Sai and Rida 2020), (Sidogi, Mbuva, and Marwala 2021), (Sen and Mehtab 2021)).

Another branch of works uses the sentiment of different instruments in a portfolio optimization approach to find the best allocation (Ferreira, Gandomi, and Cardoso 2021).

Here, approaches often use reinforcement learning in a combination with neural networks to determine the optimal allocation for each day. Koratamaddi *et al.* use an adaptive deep deterministic policy gradient approach and combine it with sentiment information, showing that the result outperforms the baseline without sentiment (Koratamaddi et al. 2021). Malandri *et al.* use the sentiment in combination with an LSTM for five different portfolios and show that the sentiment information consistently improves the performance (Malandri et al. 2018).

In this work, we use a new approach to create a sentiment-powered performance index, using the universe of the german DAX40 as our proof-of-concept. This creates the possibility for investors to include the public opinion in their investment decisions in a convenient way.

The paper is organized as follows: First, we take a look at existing sentiment indices and explain their approach. We continue by giving a rough overview of the workflow in this report, from the raw text to the calculation of the sentiment index. Here, we introduce the data we used and expand upon each of the different steps in the workflow. Finally, we show the results for our sentiment index and compare it with the performance of the DAX40.

Existing Products

In recent years, the sentiment information has been used more and more to create products which take the mood of investors into account. A few companies have created automatic solutions in the form of sentiment indices, which have related ETFs or other vehicles the public can invest into.

At the end of 2021, S&P Dow Jones Indices, in collaboration with Twitter, launched a set of sentiment indices based on the S&P 500 which mine the sentiment of Tweets concerning the companies in the original index and use it to optimize the constituent's weights. The S&P 500 Twitter Sentiment Index gathers daily sentiments and picks the 200 companies with the best sentiment once a month, weighting them by their market cap (S&P 2021a). Their second index, the S&P 500 Twitter Sentiment Select Equal Weight Index, instead selects only the best 50 companies and weights them equally (S&P 2021b). The indices have performed well thus far - creating a slow but steady outperformance of the S&P

500, although their track record is not very long.

A different example of an existing sentiment index is the BUZZ NextGen AI US Sentiment Leaders Index, which takes into account all large cap companies based in the US (BUZZ 2021b). This Index has a longer track-record, back to the end of 2015, although the related ETF has only been incepted in 2021. According to the ETF provider, the sentiment is determined using content "from online sources including social media, news articles, blog posts and other alternative datasets" (BUZZ 2021a). The 75 companies with the highest sentiment are included in the index and weighted according to their sentiment, with a maximum weight of 3%. This index is also adjusted once a month.

The downside of these products is that they are rebalanced only on a monthly basis, so the strategies cannot react to sentiment signals immediately. Big, impactful events are usually absorbed by the markets much faster, on the timescale of days or even hours (Wan et al. 2021). This means that the existing strategies miss out on most of the sentiment signals. In the next chapters, we investigate a different approach and try to create an improved product which is able to adapt more swiftly to opinionated events.

Workflow

In this section we expand on the different steps in our workflow. Starting with the news articles, we first preprocess them, before determining the sentiment. Since we usually have multiple articles for a single company on a given day, we then aggregate the sentiment to get a single value per company. From there, we calculate a weight value for each company using only the acquired sentiment value.

Data

The news article data we use consists of more than 2.85 million articles for the forty companies in the german DAX40 as of May 2022. The data spans the time from 01.01.2017 to 04.09.2022 and was scraped from german and english media websites. We received this data in a batch. To get the relevant articles for each company, we searched for articles which use the company name in their headline. For some companies, we adjust the search filter to exclude results that do not refer to the company, but a different entity. An example for this is the german sports clothing company Puma, where we found a lot of articles referring to the animal puma. For other companies, such as the automotive companies like BMW, we excluded entries reporting cases of theft or singular accidents, since small-scale events like this are unlikely to influence the stock price in any way but emit a strong sentiment. Those articles were removed by checking for the relevant keywords (like "zoo" for puma) in the headline and body of the article. After those filters, we are left with a total of around 630k documents.

Preprocessing

Before preprocessing the data, we remove a subset of articles which we deemed counterproductive to a potential forecast. A few websites regularly publish an article summarizing the previous trading day, addressing the top performers

in its headline. The sentiment contained in those articles obviously refers to the performance of the past, which we deemed detrimental to determining an indication for the future performance. These articles were removed by scanning the text for key phrases such as "dieser artikel wurde automatisch erstellt" - meaning "this article was created automatically".

We further remove duplicate articles, as well as articles with a headline longer than 1000 tokens. Finally, we move forward using only the headlines of the articles, which are transformed to lowercase.

Sentiment Classification

To extract the sentiments from the headlines, we use a BERT-based model architecture. The model is based on the distilbert model in (Sanh et al. 2019) and was fine-tuned first with the Financial Phrasebank dataset from (Malo et al. 2014) and then again with manually annotated financial news headlines. Since the model was trained on english data, but a large number of our news articles are in german, we had to find a solution to include the german-language articles. Rather than using two separate models or retraining the model as a multilanguage model, we translated the german articles to english before classifying their sentiment. For this, we use the de-to-en translation model from (Ng et al. 2019). To confirm that the translation does not lead to a steep decline in accuracy regarding the sentiment classification, we tested the model on around 1500 manually annotated headlines of our dataset (for both languages). We found that for the english articles, we achieve a balanced accuracy of 67%, while for the german-to-english translated articles we still achieve 65%.

The model returns probabilities for three classes, negative, neutral and positive. Instead of simply taking the determined class, we use the probability and multiply it with -1 for the negative class and 0 for the neutral class. This yields a value in the interval $[-1, 1]$ for each news article.

Sentiment Aggregation

The news articles are published at different times throughout the day. Since the german stock exchanges close at 5:30 pm local time, we can not consider articles published later in the day. To make sure we do not include any forward bias, we push all articles published after 5 pm to the next day. Following that, we calculate a single sentiment score for each day by taking the mean of all sentiment values. On days where no articles for a company are available, we set its sentiment to zero.

To stabilize the sentiment on days where a low amount of articles exhibits a very strong sentiment, we further apply a scaling to the sentiment values. More specifically, for a single company, we look at the amount of different sources that publish an article on a given day and multiply the sentiment with an adjustment factor that we calculate like so:

$$adj_i^t = \frac{u_i^t}{\text{mean}(u_i^{0,\dots,t-1})} \quad \text{if } u_i^t < \text{mean}(u_i^{0,\dots,t-1}) \quad (1)$$

$$adj_i^t = 1 \quad \text{else,} \quad (2)$$



Figure 1: Left axis: Performance of the Sentiment index and the DAX40. Right axis: Number of transactions.

where u_t is the number of unique sources on day t for company i . This means that the sentiment of a company is scaled towards 0 if the number of unique sources writing about this company is lower than the previous mean.

Weight Determination

Using the daily sentiments for the different companies, we want to determine the individual weights that maximize the total sentiment of the universe in the next step. Obviously, this can be achieved by simply picking the company with the highest sentiment and setting its weight to one. However, due to different constraints, this is unrealistic. Firstly, different market regulations prohibit such instruments in many countries. Secondly, this would rack up huge transaction costs since 100% of the index has to be churned on most days.

We try to solve this problem with a convex optimization approach using the python library cvxpy (Agrawal et al. 2018), (Diamond and Boyd 2016). This library allows the user to maximize an objective function under a set of constraints. The objective function maximized in our approach is as follows:

$$S = \sum_i w_i^t s_i^t - \delta |w_i^{t-1} - w_i^t| \quad (3)$$

where w_i^t and s_i^t are the weight and sentiment of instrument i at time t respectively. δ is a penalty factor which reflects the transaction costs necessary to change the weights of an instrument. In classical approaches where an expected return is optimized, this factor would be equal to the transaction costs. In our case, the factor has the scale of the sentiment that needs to change for the transaction to be worth it. In order to trade only for significant sentiment changes, we set the penalty factor to 1, which is one polarity.

As for constraints, we decide to limit the total investment in a single company to 10%. In addition, we require the sum of weights to be equal to one, so it is not possible to be invested in cash¹.

¹In reality, the optimizer cannot fulfill such a stringent constraint, since the weights of the single stocks change every day. We instead require the sum of weights to be between 99% and 99.9%

Sentiment Index

Having determined a way to calculate the weights for the different companies in the index, it is straightforward to determine the returns of the index:

$$r_{ind}^t = \sum_i w_i^t \cdot r_i^t - TC_i^t, \quad (4)$$

where TC_i^t are the transaction costs, and r_i^t is the return of stock i on day t , which we calculate as follows:

$$r^t = \frac{P^t - P^{t-1}}{P^{t-1}} \quad (5)$$

with P^t being the end-of-day price on day t . We set the trading costs to 0.05% of the price of the stock, which is multiplied with the amount of stocks traded to get the total cost of a transaction. Consequently, this part of the equation is zero for a stock if we do not trade it.

Results

In Figure 1, we show the performance of the sentiment index (in blue) and compare it to the DAX40 (in red). On the second y-axis, the number of trades is depicted. We can see that the sentiment index shows a similar performance in 2017 and 2018, but manages to outpace the DAX in the following three years. In the current year, the index lags slightly behind again. Especially just before and shortly after the corona-crisis, the index pulls well ahead of the DAX. In total, the index was able to generate an annualized return of 7.51% compared to the 2.13% of the DAX over the more than 5 years presented here.

In that time, the sentiment index performs 309 trades, 153 of which are single trades. Single trades are usually performed when an instrument goes above the upper bound of 10% we set for the optimization. Other than that, on 128 days the index trades twice and on 18 days it trades 3 times. Disregarding the initial investment, there are never more than 3 transactions in a single day. In conclusion, this approach of a sentiment index performs trades more frequently, but only for cases with very strong sentiments.

Conclusion

In this work, we present a framework for determining sentiment values from news headlines and transforming them into weights for a performance sentiment index, achieving a significant outperformance compared to the benchmark index over the five year time span we observed. We compare our approach to existing products, which have pre-defined re-balancing dates in contrast to our more flexible approach which is able to change the constituent weights at any point in time without amassing large amounts of transaction costs. We apply the framework to the German stock market, the DAX40, as an example. However, it is straightforward to apply it to any other universe of stocks.

In future works, we would like to adapt this framework for portfolios, where it is possible to reduce the investment ratio and hold cash. This opens a new dimension as the general market sentiment is important, as opposed to just the sentiment of the constituents. In addition, we plan on continually improving different areas of the framework and compare new results to the here presented baseline approach.

One such area is the sentiment model, which could be refined further, for example by using a larger amount of annotated data to finetune it. In addition, the use of the penalty factor can be investigated and refined further, for example by using different values for different market situations.

Acknowledgements

This work was partially funded by GET Capital AG. The authors would further like to thank pressrelations GmbH for the support in infrastructure they provided.

References

- Agrawal, A.; Verschueren, R.; Diamond, S.; and Boyd, S. 2018. A rewriting system for convex optimization problems. *Journal of Control and Decision*, 5(1): 42–60.
- Allen, D. E.; McAleer, M.; and Singh, A. K. 2019. Daily market news sentiment and stock prices. *Applied Economics*, 51(30): 1–24.
- BUZZ. 2021a. BUZZ - VanEck Social Sentiment ETF. <https://www.vaneck.com/us/en/investments/social-sentiment-etf-buzz/index/>. Accessed: 2022-12-21.
- BUZZ. 2021b. It's a New Era of Investing. Welcome to BUZZ. - BUZZ Indexes. <https://www.investwithbuzz.com/>. Accessed: 2022-12-20.
- Diamond, S.; and Boyd, S. 2016. CVXPY: A Python-Embedded Modeling Language for Convex Optimization. arXiv:1603.00943.
- Feng, F.; Wang, X.; He, X.; Ng, R.; and Chua, T.-S. 2021. Time horizon-aware modeling of financial texts for stock price prediction. *Proceedings of the Second ACM International Conference on AI in Finance*, 1–8.
- Ferreira, F. G. D. C.; Gandomi, A. H.; and Cardoso, R. T. N. 2021. Artificial Intelligence Applied to Stock Market Trading: A Review. *IEEE Access*, 9: 30898–30917.
- Ko, C.-R.; and Chang, H.-T. 2021. LSTM-based sentiment analysis for stock price forecast. *PeerJ Computer Science*, 7: e408.
- Koratamaddi, P.; Wadhwani, K.; Gupta, M.; and Sanjeevi, S. G. 2021. Market sentiment-aware deep reinforcement learning approach for stock portfolio allocation. *Engineering Science and Technology, an International Journal*, 24(4): 848–859.
- Li, Q.; Chen, Y.; Wang, J.; Chen, Y.; and Chen, H. 2018. Web Media and Stock Markets : A Survey and Future Directions from a Big Data Perspective. *IEEE Transactions on Knowledge and Data Engineering*, 30(2): 381–399.
- Malandri, L.; Xing, F. Z.; Orsenigo, C.; Vercellis, C.; and Cambria, E. 2018. Public Mood-Driven Asset Allocation: the Importance of Financial Sentiment in Portfolio Management. *Cognitive Computation*, 10(6): 1167–1176.
- Malo, P.; Sinha, A.; Korhonen, P.; Wallenius, J.; and Takala, P. 2014. Good debt or bad debt: Detecting semantic orientations in economic texts. *Journal of the Association for Information Science and Technology*, 65(4): 782–796.
- Min, B.; Ross, H.; Sulem, E.; Veyseh, A. P. B.; Nguyen, T. H.; Sainz, O.; Agirre, E.; Heinz, I.; and Roth, D. 2021. Recent Advances in Natural Language Processing via Large Pre-Trained Language Models: A Survey. arXiv:2111.01243.
- Ng, N.; Yee, K.; Baeovski, A.; Ott, M.; Auli, M.; and Edunov, S. 2019. Facebook FAIR's WMT19 News Translation Task Submission. arXiv:1907.06616.
- Ranco, G.; Aleksovski, D.; Caldarelli, G.; Grčar, M.; and Mozetič, I. 2015. The Effects of Twitter Sentiment on Stock Price Returns. *PLoS ONE*, 10(9): e0138441.
- Sai, V. K.; and Rida, A. 2020. Predicting Stock Movement Using Sentiment Analysis of Twitter Feed with Neural Networks. *Journal of Data Analysis and Information Processing*, 08(04): 309–319.
- Sanh, V.; Debut, L.; Chaumond, J.; and Wolf, T. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv:1910.01108.
- Sen, J.; and Mehtab, S. 2021. A Robust Predictive Model for Stock Price Prediction Using Deep Learning and Natural Language Processing.
- Sidogi, T.; Mbuvha, R.; and Marwala, T. 2021. Stock Price Prediction Using Sentiment Analysis. *2021 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 00: 46–51.
- S&P. 2021a. S&P 500 Twitter Sentiment Index — S&P Dow Jones Indices. <https://www.spglobal.com/spdji/en/indices/strategy/sp-500-twitter-sentiment-index/#overview>. Accessed: 2022-12-20.
- S&P. 2021b. S&P 500 Twitter Sentiment Select Equal Weight Index — S&P Dow Jones Indices. <https://www.spglobal.com/spdji/en/indices/strategy/sp-500-twitter-sentiment-select-equal-weight-index/#overview>. Accessed: 2022-12-20.
- Wan, X.; Yang, J.; Marinov, S.; Calliess, J.-P.; Zohren, S.; and Dong, X. 2021. Sentiment correlation in financial news networks and associated market movements. *Scientific Reports*, 11(1): 3062.

6.2 ESG-Investing through Aspect Regimes with Dynamic Time Warping

Paper: F. Billert and S. Conrad. Identifying Aspect-Regimes for Enhanced ESG-Investing Through News Data. In *2025 IEEE Symposium on Computational Intelligence for Financial Engineering and Economics (CiFer)*, Trondheim, Norway, 2025, pp. 1-7

Personal Contribution: Fabian Billert developed the research idea and implemented it, including optimizing the hyperparameters and creating the different simulations. Fabian Billert wrote all parts of the paper. Stefan Conrad supervised the work.

Status: Published

Contrary to the previous section, in this paper, we pursue a portfolio-approach based on the dataset we created in [section 5](#). As we discussed in [subsection 2.3.3](#), positive ESG-practices of companies can be beneficial for their value [\[21\]](#). However, how can ESG-information be included in portfolio optimization processes?

Henriksson et al. [\[51\]](#) argue that including only ESG-information relevant to a company’s industry is preferable to simply using all available ESG-information. In addition, they develop a new variable they call “Good Minus Bad (GMB)” factor in order to deal with data sparsity that occurs when firms do not sufficiently disclose their ESG-data. Xidonas et al. [\[164\]](#) use ESG data from Sustainalytics and follow a minimax-based optimization-model in order to maximize the different ESG metrics in the portfolio. They find that their approach achieves higher returns with lower volatility.

As we noted previously, conventional ESG-ratings have various downsides. Several other works pursue a news-based approach and extract ESG-related sentiment themselves. Feng et al. [\[40\]](#) collect online news articles concerning companies in the S&P500 and STOXX600 and extract different types of sentiment (including ESG-specific sentiment) using a lexicon of tonal words. They then explore different approaches for portfolio optimization while taking this sentiment into account. Goutte et al. [\[47\]](#) follow a similar approach but use news data from the Global Database of Events, Languages, and Tone (GDELT) project² and find that ESG-related indicators improve the performance of the designed portfolios.

In the following, we attempt to use the previously mined ESG-sentiments in order to optimize portfolio allocation. Following what Henriksson et al. [\[51\]](#) find, we take into account that different ESG-aspects have varying degrees of importance for different companies. For this reason, we apply the Dynamic Time Warping (DTW) algorithm to identify “aspect-regimes” where sentiments for certain ESG-aspects seem to be more important than others.

²<https://www.gdeltproject.org/data.html>

Identifying Aspect-Regimes for Enhanced ESG-Investing Through News Data

1st Fabian Billert

Institute for Informatics
Heinrich-Heine University of Düsseldorf
Düsseldorf, Germany
fabian.billert@uni-duesseldorf.de

2nd Stefan Conrad

Institute for Informatics
Heinrich-Heine University of Düsseldorf
Düsseldorf, Germany
stefan.conrad@uni-duesseldorf.de

Abstract—ESG (Environmental, Social, Governance)-thematics play an important role for investing as they can represent additional risk for a company and subsequently its stock price. These ESG-thematics can be extracted from alternative data sources such as news data. However, the extend to which an ESG-event influences a company can vary greatly depending on the ESG-aspect of the event and the current expectations of the affected company.

In this paper, we present an approach where we leverage Dynamic Time Warping (DTW) in order to identify periods in which certain ESG-aspects are closely aligned with the stock-price movement of a company. We demonstrate this approach using more than 50,000 news articles mined for 42 german companies between January 2023 and September 2024, (i) answering the question of which aspects are important for a company in what time-frames and (ii) showing that this approach increases the performance of an ESG-sensitive portfolio by a large amount.

Index Terms—ESG-investing, portfolio optimization, dynamic time warping, financial news, alternative data

I. INTRODUCTION

Environmental, social and governance (ESG) thematics have become an important factor for investors and companies alike in the last few years [1], [2]. The ESG-impact of a company is often quantified by a so-called ESG-score which is calculated by established rating agencies and aims to summarize the sustainability practices of the company gathered from many sources in a single value [3]. However, these ratings have recently been criticized as different raters determine different scores for the same company [4], and it is not always clear how the final value is calculated [5]. For this reason, researchers have made efforts to extract corporate sustainability behavior from alternative data sources - such as news articles [6] and company reports [7].

Advances in natural language processing (NLP) capabilities in the last few years have enhanced the quality of information extracted from such alternative data sources massively as well as enabled researchers of different fields to access this data [8]–[11]. [12] for example has rallied NLP researchers to develop models to determine ESG aspects of news articles while [13] encouraged novel ESG-sentiment classification models. Using this ESG-information for investing has become an increasingly interesting topic, as bad sustainability practices

and associated events pose an additional risk for investors which can be detected and accounted for [14].

However, it is not always clear which ESG-events are relevant for a company and which are expected for their industry and thus do not present a new risk. [15] for example finds that ESG-risks mainly influence returns of S&P 500 companies if they are driven by the Social and Governance ESG-aspects while Environmental events are mostly ignored. In general, though, not a lot of research has assessed the influence of different ESG-aspects on companies.

In this work, we investigate this correlation between specific ESG-aspects and associated stock returns further. We present an approach which applies the dynamic time warping (DTW) algorithm to ESG-sentiment data extracted from news articles in order to identify aspect regimes in which news of one specific ESG-aspect (e.g. E, S or G), or any combination of ESG-aspects are predicting future price movement. At the same time, by applying this technique in a realistic portfolio optimization approach, we show that it can improve the portfolio performance compared to simply using all ESG-data. The underlying news articles stem from [6], which extracts ESG-information for german companies in the financial DAX-40 index. This dataset contains ESG-sentiments as well as aspects for over 50,000 news articles from the start of 2023 until September 2024.

We share parts of our code on GitHub¹. The repository contains a notebook which processes the ESG-data and applies the DTW algorithm to determine which sentiments to use as predictors at what times. Note that the repository uses prices from yfinance², while we use prices from a different source in this paper (which we are not allowed to share). For this reason, the results might vary slightly. At the same time, we unfortunately are not permitted to share the code which performs the portfolio optimization.

II. RELATED WORK

A. Impact of ESG-Events on Stock Prices

Recent works have found evidence to support the point that ESG-related events have an impact on the stock price.

¹<https://github.com/Bailefan/ESG-Investing>

²<https://github.com/ranaroussi/yfinance>

[16] extracts sentiments from ESG-related news articles and creates ESG-sentiment indices for each constituent of the Dow Jones Industrial Average index. They find that the constituent's returns are generally affected by the ESG-sentiments, but note that the relation between the two can differ greatly between different kinds of companies. [17] investigates the interdependence of ESG ratings, ESG-news and the stock performance of more than 3000 companies and comes to the conclusion that while certain ESG news do correlate with the future market performance of stocks, this correlation drops in cases of high disagreement between different rating agencies for the ESG-rating of a company. [18] describes similar findings, but notes that the historical ESG-record of a company affects the influence of the news sentiments as companies with an overall better ESG-record are more resilient with regards to negative ESG-news while companies with a poor ESG-record have a more pronounced reaction to positive ESG-news. [19] investigates if news-based ESG-sentiment influences the crash-risk of stock-prices of Chinese companies. They find a negative correlation between the two, meaning an increased ESG-sentiment leads to a lower crash-risk of stock-prices.

B. Investing with Alternative Data

There have been many different approaches of including alternative data into investment decisions. [20] constructs an investor sentiment index from different macroeconomic factors and uses it to optimize a portfolio of major stock indices which outperforms all benchmarks. [21] uses a Long Short Term Memory (LSTM) based model to optimize a portfolio while taking sentiment information from news articles and twitter data into account. They find that their approach outperforms traditional portfolio optimization methods but is more unstable in general. [22], too, use an LSTM model to include sentiment information into portfolio optimization for different universes - in all cases, the approaches including the sentiment outperform a baseline without this information. [23] uses Reinforcement-Learning (RL) to train an automated trader which acts based on both historical stock price data and sentiment data from twitter and google news. The authors show that adding the sentiment data to the trading agent improves the portfolio optimization results. [24] constructs a sentiment index based on news articles of german companies and shows that this index achieves an outperformance compared to its benchmark.

Using specifically ESG-sentiment data in order to improve portfolio optimization has also been approached by several research works. [25] use a Multi Criteria Decision-Making (MCDM) model named TODIM to construct portfolios which take the ESG scores of equities into account and find that this increases the performance significantly compared to an ESG-blind baseline. [26] introduces an adjusted Mean-Variance approach which manages to increase ESG-scores in portfolios while still achieving performance gains - although at reduced Sharpe-ratios.

C. DTW for Portfolio-Optimization

Dynamic time warping (DTW) has emerged as a widely used technique in finance and, more specifically, portfolio optimization approaches. Originally mainly used for audio processing [27], [28], it has become a popular tool to work with time-series data of all kinds - such as genomic signals [29] and gait data [30]. [31] uses a clustering technique called DTW Baycenter Averaging (DBA) in order to increase portfolio diversification and investigated the resulting performance during the 2008 global financial crisis, in 2018 and during the 2020 Covid-19 crisis, showing that their approach increased the performance compared to popular portfolio optimization benchmarks. [32] proposes improvements to the Anticor algorithm by including the DTW distance as an improved similarity measure and achieves performance gains in all investigated markets. [33] use DTW in order to distinguish leading and lagging time series out of more than 700 assets and show that using the trend of the leading companies to trade with the lagging companies results in performance gains.

III. NANO-ESG DATASET

Nano-ESG [6] contains ESG-information for 51,067 news articles concerning 42 german companies which either are current constituents of the DAX-40 or were only recently removed from the financial index. The information is extracted from german and english news articles by leveraging Large Language Models (LLMs). All articles were published between 2023-01-01 and 2024-09-17. In the end, the dataset contains short summaries, ESG-sentiments (negative, neutral, positive), ESG-aspects (environmental, social, governance) as well as the timestamp of each article.

A. Dataset bias

Nano-ESG exhibits some bias with regards to both sentiments and aspects. Regarding aspects, there are significantly more articles with the environmental aspect - nearly twice as many as for the other two aspects. However, this seems to be mainly focused on a few companies with many articles as, for each company, the average ratio of articles per aspect is approximately similar for the three aspects. Looking at the polarity, the positive sentiment outweighs the other two sentiments by a large amount. On average, 54% of a company's news articles are positive, 33% are negative and 13% are neutral.

B. Relevancy

The dataset also includes a measure for how relevant an article is with regards to the sentiment for the target company. This relevance score was obtained by prompting the LLM for its assessment of the relevancy of an article on a scale from 1 to 10. [6] investigates if the relevance scores the LLM determines are meaningful or not by manually evaluating a subset of the articles. They find that the percentage of articles that were deemed irrelevant by at least one of the annotators is generally higher for articles where the LLM assigns a low relevance score. We conclude from this that the relevance score

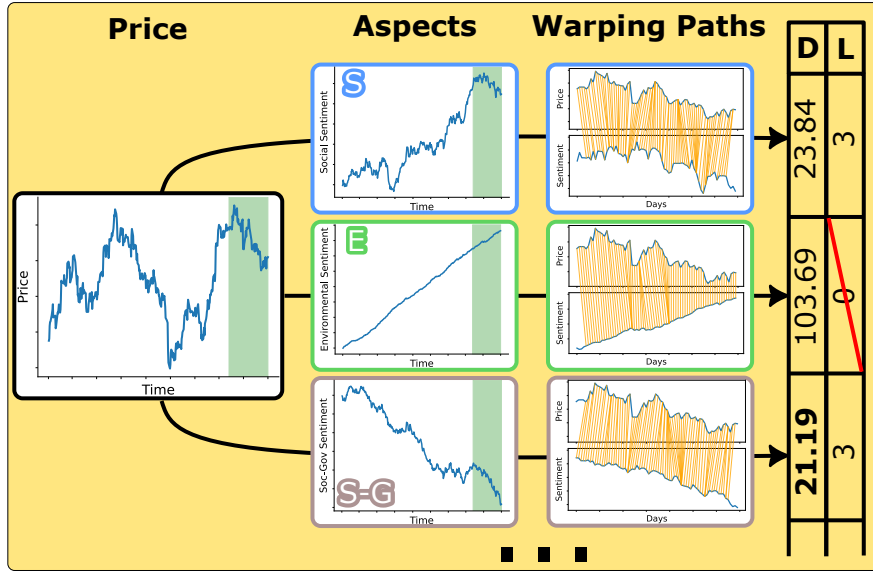


Fig. 1: Approach to determine which aspect is currently relevant for the company Daimler as an example. Each aspect is used together with the price to calculate the warping paths. Finally, depending on the (D)istance and (L)ag one aspect is identified as the current aspect regime.

determined by the LLM is a valuable metric which indicates which articles are less likely to contain relevant information for the target company and thus provide additional noise with regards to the company's current ESG-sentiment-state.

IV. ESG-INVESTING WITH DTW

In this section, we first explain how we transform the article-level sentiment information to a sentiment graph for each company. By using dynamic time warping to find similarities between the different ESG-sentiments and the stock prices, we can then identify live aspect-regimes without using future information. Finally, we explain how we use the identified aspect regimes to simulate an active portfolio.

A. ESG-Sentiment Preprocessing

In this subsection we describe the pipeline we use to determine continuous ESG-sentiment graphs from the Nano-ESG dataset, which we can compare with the prices of the companies. First of all, to avoid overfitting, we shift all articles which were published between the last 30 minutes before the stock exchange closes and the end of the day to the next day. Articles published after the trading day ends cannot influence the price, and we leave an additional 30 minutes window so that it is realistic to act on the new information. To process the data, we start by converting the sentiments to numerical values, meaning a negative sentiment is set to -1 , a neutral sentiment to 0 and a positive sentiment to 1 . Next, we calculate the mean sentiment of all articles of every trading day. Here, we include the relevance score of each news article:

Relevance-Scaling - As we discuss in Subsection III-B, the relevance score is related to the likelihood of an article being relevant for the company or not. In order to account for this,

we weight each article by its own relevance score. Afterwards we divide by the maximum relevance score of the articles so that the sentiment is constrained to lie between -1 and 1 . We end up with the following formula for the sentiment on day t for one company:

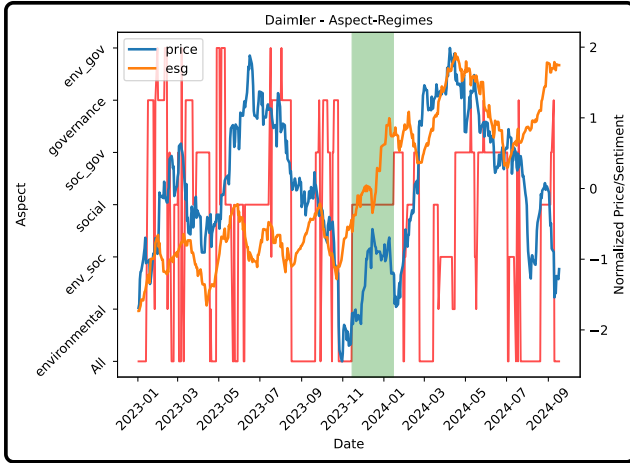
$$S^t = \frac{\text{mean}_i(s_i^t \cdot \text{Rel}(s_i^t))}{\text{max}_i(\text{Rel}(s_i^t))}, \quad (1)$$

where s_i^t is the sentiment of a single article i and $\text{Rel}(s_i^t)$ is its Relevance Score. This allows us to enhance sentiments of high relevance articles while not completely losing out on information from low relevance articles in cases where the overall relevancy is low. On days where we have no sentiment information for a company, we set its sentiment to 0 . Finally, we calculate the cumulative sum of the daily ESG-sentiments. In this step, we also consider the different aspects:

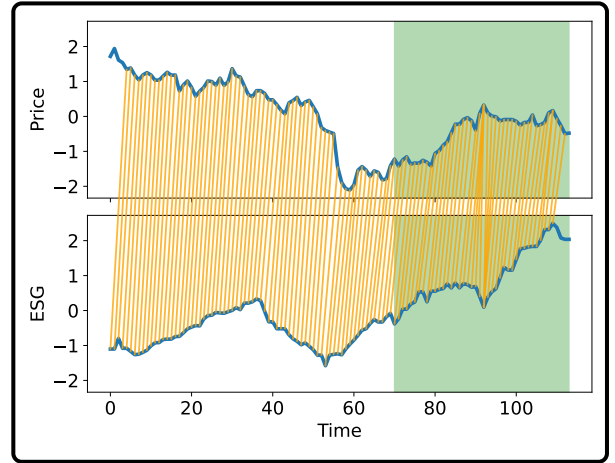
Aspect-Data - By filtering the original articles with regards to different ESG-aspects, we can create time-series for all possible combinations, for example environmental, or environmental and social (E-S). This means we can define a total of seven distinct aspect combinations. Certain aspects might be more or less important for a company depending on different internal and external factors. Because of this, we explore an approach based on DTW in the next subsection in order to find out which ESG-aspects are important for a company at what times. Examples for three different aspect-sentiment graphs are displayed in the middle row of Fig. 1 for the company Daimler.

B. Dynamic Time Warping

Dynamic Time Warping (DTW) is a technique for measuring similarity between two time-series which may vary in



(a) Price, composed ESG-sentiment and detected aspect-regimes of Daimler. We focus on the Social-Regime marked in green in plot (b).



(b) Warping paths for the Social-Regime marked in green in plot (a). The 70 days before the regime starts correspond to the `data_window` used to identify the regime.

Fig. 2: An example of all detected aspect regimes of Daimler. Each aspect regime was detected at runtime. The ESG-sentiments in (a) are composed, meaning they always show the sentiments for the relevant aspect.

speed or duration. It aims to find an optimal alignment between the two sequences while allowing for non-linear "warping" of the time axis. In order to determine how similar the price is to the different aspects, we use the DTW implementation from [34] called `dtwdistance`. We calculate the warping paths between the two time-series, which allows us to find the minimum distance as well as the best alignment. All time-series are z-score normalized before using the DTW algorithm. Since we are most interested in the current correlation (at time t) of the two time series, we aim to focus on the alignment of the more recent data. Because of this, we only provide a subset of the data $t - \text{data_window}$ until t instead of the whole time series to the algorithm. Another important parameter is the `dtw\_window`, which determines how far apart points of the two time-series can be and still be matched with one another. A third parameter, the ψ -relaxation parameter, is set to the same value as the `dtw\_window` value. ψ indicates how far ahead one time series is allowed to be of the other time series at their edges - by default ($\psi = 0$) all data points are forcibly mapped between the two graphs. Setting ψ to a non-zero values allows us to observe which time series lags behind at time t , and which one is ahead and thus acts as a predictor for the other time series.

C. Identifying Live ESG-Aspect Regimes

On every day t we apply the DTW algorithm for every price-aspect combination. A schema which describes this process is depicted in Fig. 1. The distance returned by the algorithm indicates how similar the aspect-sentiment is to the price. By iterating over all aspects, we can find the combination with the lowest distance. At the same time, we calculate the current lag through the best identified path. We then ignore cases where the sentiment does not act as a predictor for the price

- independent of the distance. We use the sentiment of the aspect we find as an estimator in our simulation for time t . If we cannot find any aspects which predict the price, we simply set the estimator for that day and company to 0.

Fig. 2a shows an example of the price, ESG-sentiment and regime for the automotive company Daimler (also called Mercedes-Benz Group). Here, the ESG-sentiment graph is a composed graph containing the sentiment data of the determined aspect regime at any point. The aspect regime marked in green is shown in more detail in Fig. 2b along with the warping paths in yellow. Note that while the green part in this plot corresponds to the same identified aspect regime, the entirety of the displayed data from $t = 0$ to $t = 70$ (days) is used to apply DTW and identify this regime starting at $t = 71$.

D. Simulation Setup

We create a portfolio with all companies we have sentiments for and aim to optimize the estimators identified in Subsection IV-C on each day. For this, we use a convex optimization approach via the python library `cvxpy` [35], [36]. In this library, it is possible to define an objective function which will be maximized while taking a set of constraints into account. Our approach is similar to [24] and we use a variation of the objective function used there which maximizes the overall sentiment S at time t :

$$S^t = \sum_i w_i^t \hat{s}_i^t - \delta |w_i^{t-1} - w_i^t|, \quad (2)$$

where w_i^t represents the weight of stock i and δ is a penalty factor which reflects the transaction costs necessary to adjust the weight of the stock. Since we aim to optimize sentiments, δ should be a value on a similar scale. We choose to set $\delta = 1$, which corresponds to a step of one polarity of the

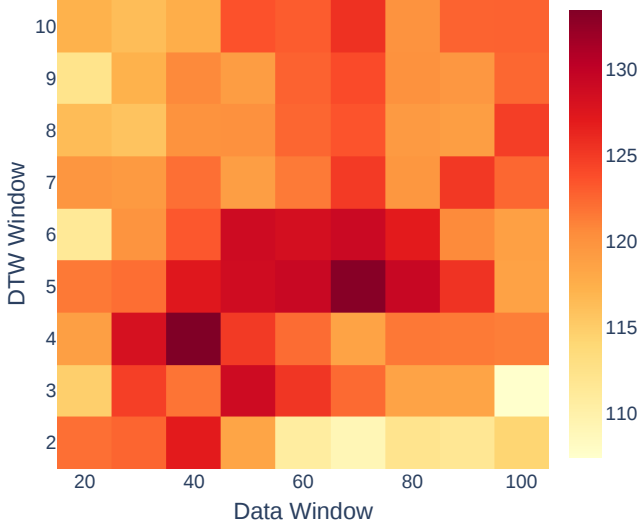


Fig. 3: Optimization of the parameters Data Window (x-axis) and DTW Window (y-axis) between January 2023 and the end of February 2024. Darker colors indicate a higher performance in this time-range.

sentiment. Finally, $\hat{s}_i^t$ is the ESG-sentiment of the aspect with the smallest DTW distance and positive lag as discussed in Subsection IV-B. At the same time, we necessitate some constraints to make the simulation more realistic: every stock has an upper bound of 10%. In addition, we decided to focus on an approach without cash for now, so we require the sum of all weights to add up to 1. If the weights of a stock have to be changed, the change has to be bigger than 0.5%. Lastly, we also apply a transaction cost of 0.05% to every change of weights.

E. Hyperparameter Optimization

The two main hyperparameters (HP) which can influence the simulations are the *dtw_window*, which determines how far apart two points from two time series can be and still be matched, and the *data_window* we use to select how much of each time series we want to apply DTW to at each step. In order to find stable values for these two parameters, we first do a grid search for one part of the data, where we simulate for each combination of parameters and then look for clusters where the performance of the simulation stays relatively stable when the hyperparameters vary. Since the amount of available sentiment information we have is constrained to a relatively small time-range, we have to be careful with the selection of the data used for HP-optimization. In our experiments we found that using too little data to validate leads to instable hyperparameters. In the end, we decided to use roughly two thirds of the data, or from January 2023 until the end of February 2024 to optimize our hyperparameters.

The results for this are displayed in Fig. 3. We can identify two maxima in performance - one at $(x=40, y=4)$ and one at $(x=70, y=5)$. In order to determine which maximum shows

TABLE I: Final Performance and Sharpe Ratio of our Simulation and Baselines.

	DAX-40	ESG-Sentiments	Aspect Regimes
Sharpe Ratio	0.269	0.032	0.902
Performance	105.6	103.6	113.7

more stability, we look at the mean performance and Sharpe ratios of the (5×5) square surrounding it. We find that the HPs around the $(70, 5)$ maximum have a higher average Sharpe ratio and a higher average performance, both with lower standard deviations. As a result, we select the parameters *dtw_window* = 5 and *data_window* = 70 as the optimal hyperparameters.

F. Baselines

We compare our simulation to two different baselines in order to judge how effective our approach is. The first baseline is the german stock index DAX-40, which consists of mostly the same company included in the Nano-ESG dataset. Two companies are missing: Merck as well as Porsche Automobil Holding, while the dataset includes the recently departed companies Delivery Hero, Hellofresh, Puma and Fresenius Medical Care. The weight of the missing companies in the DAX-40 is lower than 3%, so the DAX should still be a reasonable comparison to our approach. The second baseline is created by simply using all available sentiments from the Nano-ESG dataset without applying DTW and identifying relevant aspects.

V. RESULTS

We present our simulation for the validation-time-range in 2024 in graph Fig. 4, along with the three baselines described above. We can see that over the depicted time-range, the approach investing via aspect regimes achieves the highest performance. At the same time, using only the ESG-sentiments without processing them using DTW performs the worst. The DAX shows a slightly higher performance than this baseline but is also outperformed by our presented approach.

In Table I we list the final performances and the Sharpe ratios of the simulation and different baselines. We used a risk-free rate of 3.7%, which is the average of the euribor-1m³ in the relevant time range. Although the Sharpe Ratio of the aspect regimes approach is slightly below one, it is still significantly higher than the values we measure for the other two baselines. In order to confirm the stability of the hyperparameters, we assess the performance and Sharpe-ratios near our selected HPs as above. We find an average Sharpe ratio of 0.6577 ± 0.2737 and an average final performance of 110.49 ± 3.16 , significantly higher than both of the baselines we compare our approach to.

Finally, we analyze the aspect regimes we detect using the presented approach more in detail. In Table II we compare statistics of all detected aspect regimes with various subsets of companies which represent different industries. The top row

³<https://www.euribor-rates.eu/en/>

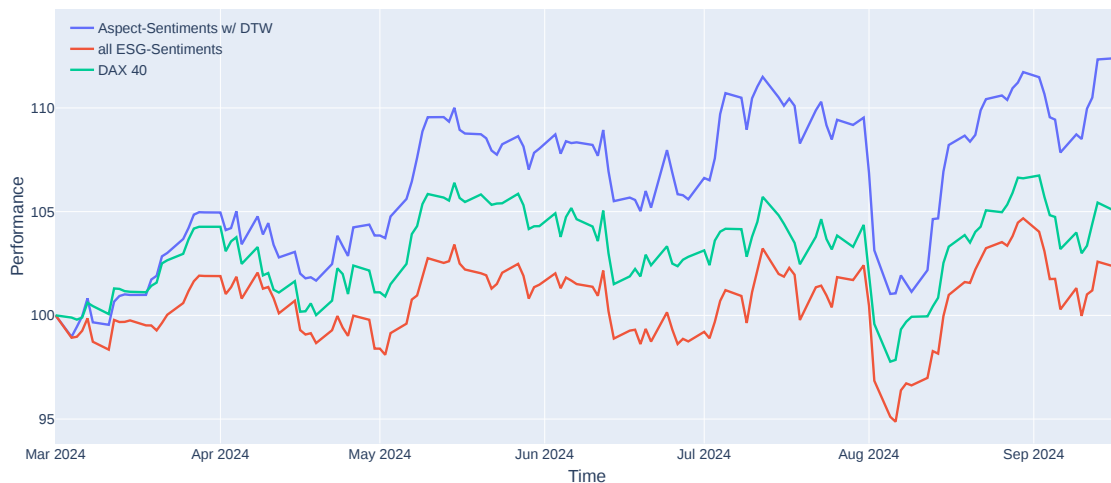


Fig. 4: Performance of our strategy using DTW to identify aspect-regimes (blue), in comparison to simply using all ESG-sentiments (red), using all ESG-sentiments but filtering by lag (green) and the DAX-40 (purple).

shows the percentage of regimes we found for each aspect-combination. Below that, we display how much higher or lower this percentage is for different company groups. The automobile industry (BMW, VW, Daimler, Porsche) interestingly exhibits less regimes for environmentally-adjacent aspects in favor of more social regimes. Important social topics for car manufacturers are for example human rights in the supply chain and suspected forced-labor plants in other countries. Electronic vehicles are also an important topic but don't seem to correlate with the price as much - maybe because these companies are lagging behind expectations.

Looking at financial companies (Deutsche Bank, Commerzbank and Deutsche Börse) we can see the opposite picture: environmental issues seem to be more important here. Major topics here are carbon-neutral portfolios and monetary support for sustainable projects. We also investigated the chemical sector (BASF, Bayer, Covestro, Brenntag) but found no significant diversion from the amount of typically detected regimes.

VI. CONCLUSION AND FUTURE WORK

In this study, we presented a novel approach to identify live ESG-aspect regimes using dynamic time warping and showed that using aspect-specific ESG-sentiments achieves an outperformance in portfolio simulations compared to an aspect-blind baseline. Although the resulting simulation only covers a fairly short amount of time, we attempted to increase

the confidence of our results by testing a wide area of values around our determined hyperparameters, showing that they exhibit similar results. However, once more ESG-data is available, we aim to further test our presented approach.

In addition, we analyzed the regime-differences of various industries and showed two cases which exhibit interesting outliers with regards to aspects. Future works could investigate this point in more detail, for example by examining the underlying news articles of different aspect regimes and identifying which specific topics resulted in correlated price movements.

VII. LIMITATIONS

In addition to the previously discussed limitation related to data volume, which led to a relatively short simulation time-frame, there are a few other potential challenges to consider. Firstly, some companies within the Nano-ESG dataset have limited underlying ESG-sentiment data. While in those cases not a lot of aspect regimes will be detected in the first place, the regimes that are identified for such companies may be less reliable compared to those with more comprehensive data coverage. Secondly, the source dataset only includes one type of media - news articles. This presents another consideration. If an ESG-related event is disseminated through channels not covered by the dataset (e.g., social media) and impacts stock prices before being reported in news articles, it may go undetected by the approach used in this study.

ACKNOWLEDGMENT

We want to thank GET Capital AG for the support in infrastructure they provided.

REFERENCES

- [1] H. Schramm-Klein, J. Zentes, S. Steinmann, B. Swoboda, and D. Morschett, "Retailer Corporate Social Responsibility Is Relevant to Consumer Behavior," *Business & Society*, vol. 55, no. 4, pp. 550–575, 9 2013.

TABLE II: Ratios of detected Regimes for different aspects for all companies (top), and diversion from these values for company-subsets of different industries (below).

All in (%)	E	E-S	S	S-G	G	G-E
Overall	12.8	11.2	20.0	18.8	21.9	15.3
Automobile	-7.8	-4.9	+15.7	+7.4	-0.6	-9.8
Finance	+11.8	-8.5	-0.1	-4.8	-14.4	+8.4
Chemicals	+2.3	-1.6	-1.1	-0.2	-2.6	+3.1

- [2] T. Islam, R. Islam, A. H. Pitafi, L. Xiaobei, M. Rehmani, M. Irfan, and M. S. Mubarak, "The impact of corporate social responsibility on customer loyalty: The mediating role of corporate reputation, customer satisfaction, and trust," *Sustainable Production and Consumption*, vol. 25, pp. 123–135, 8 2020.
- [3] M. Research, "ESG Ratings Methodology," 2024. [Online]. Available: <https://www.msci.com/documents/1296102/34424357/MSCI+ESG+Ratings+Methodology.pdf>
- [4] F. Berg, J. F. Kölbl, and R. Rigobon, "Aggregate Confusion: The Divergence of ESG Ratings," *SSRN Electronic Journal*, 2019.
- [5] A. Clément, E. Robinot, and L. Trespeuch, "The use of ESG scores in academic literature: a systematic literature review," *Journal of Enterprising Communities: People and Places in the Global Economy*, vol. ahead-of-print, no. ahead-of-print, 2023.
- [6] F. Billert and S. Conrad, "Nano-ESG: Extracting Corporate Sustainability Information from News Articles," *arXiv*, 12 2024.
- [7] J. Ni, J. Bingler, C. Colesanti-Senni, M. Kraus, G. Gostlow, T. Schimanski, D. Stambach, S. A. Vaghefi, Q. Wang, N. Webersinke, T. Wekhof, T. Yu, and M. Leippold, "CHATREPORT: Democratizing Sustainability Disclosure Analysis through LLM-based Tools," *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 21–51, 2023.
- [8] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *arXiv*, 2018.
- [9] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, and G. Lample, "LLaMA: Open and Efficient Foundation Language Models," *arXiv*, 2023.
- [10] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, D. Bikel, L. Blecher, C. C. Ferrer, M. Chen, G. Cucurull, D. Esiobu, J. Fernandes, J. Fu, W. Fu, B. Fuller, C. Gao, V. Goswami, N. Goyal, A. Hartshorn, S. Hosseini, R. Hou, H. Inan, M. Kardas, V. Kerkez, M. Khabsa, I. Kloumann, A. Korenev, P. S. Koura, M.-A. Lachaux, T. Lavril, J. Lee, D. Liskovich, Y. Lu, Y. Mao, X. Martinet, T. Mihaylov, P. Mishra, I. Molybog, Y. Nie, A. Poulton, J. Reizenstein, R. Rungta, K. Saladi, A. Schelten, R. Silva, E. M. Smith, R. Subramanian, X. E. Tan, B. Tang, R. Taylor, A. Williams, J. X. Kuan, P. Xu, Z. Yan, I. Zarov, Y. Zhang, A. Fan, M. Kambadur, S. Narang, A. Rodriguez, R. Stojnic, S. Edunov, and T. Scialom, "Llama 2: Open Foundation and Fine-Tuned Chat Models," *arXiv*, 2023.
- [11] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language Models are Few-Shot Learners," *arXiv*, 2020.
- [12] C.-C. Chen, Y.-M. Tseng, J. Kang, A. Lhuissier, M.-Y. Day, T.-T. Tu, and H.-H. Chen, "Multi-Lingual ESG Issue Identification," in *Proceedings of the Fifth Workshop on Financial Technology and Natural Language Processing (FinNLP) and the Second Multimodal AI For Financial Forecasting (Muffin)*, Macao, 8 2023, pp. 111–115. [Online]. Available: <https://aclanthology.org/2023.finnlp-1.11>
- [13] C.-C. Chen, Y.-M. Tseng, J. Kang, A. Lhuissier, Y. Seki, M.-Y. Day, T.-T. Tu, and H.-H. Chen, "Multi-Lingual ESG Impact Type Identification," *Proceedings of the Sixth Workshop on Financial Technology and Natural Language Processing*, pp. 46–50, 2023.
- [14] P. Capelli, F. Ielasi, and A. Russo, "Measuring ESG risks in multi-asset portfolios: Decomposing VaRESG into CVaRESG," *Finance Research Letters*, vol. 66, p. 105692, 2024.
- [15] M. L. D. Nicolas, A. Desroziers, F. Caccioli, and T. Aste, "ESG Reputation Risk Matters: An Event Study Based on Social Media Data," *arXiv*, 2023.
- [16] A. Schmidt, "Sustainable News – A Sentiment Analysis of the Effect of ESG Information on Stock Prices," *SSRN Electronic Journal*, 2019.
- [17] G. Serafeim and A. Yoon, "Stock price reactions to ESG news: the role of ESG ratings and disagreement," *Review of Accounting Studies*, vol. 28, no. 3, pp. 1500–1530, 2022.
- [18] G. Dorfleitner and R. Zhang, "ESG News Sentiment and Stock Price Reactions: A Comprehensive Investigation via BERT," *Schmalenbach Journal of Business Research*, vol. 76, no. 2, pp. 197–244, 2024.
- [19] H. Yu, C. Liang, Z. Liu, and H. Wang, "News-based ESG sentiment and stock price crash risk," *International Review of Financial Analysis*, vol. 88, p. 102646, 2023.
- [20] N. Banholzer, S. Heiden, and D. Schneller, "Exploiting investor sentiment for portfolio optimization," *Business Research*, vol. 12, no. 2, pp. 671–702, 2019.
- [21] M. Muthivhi and T. L. v. Zyl, "Fusion of Sentiment and Asset Price Predictions for Portfolio Optimization," *arXiv*, 2022.
- [22] L. Malandri, F. Z. Xing, C. Orsenigo, C. Vercellis, and E. Cambria, "Public Mood-Driven Asset Allocation: the Importance of Financial Sentiment in Portfolio Management," *Cognitive Computation*, vol. 10, no. 6, pp. 1167–1176, 2018.
- [23] P. Koratamaddi, K. Wadhvani, M. Gupta, and S. G. Sanjeevi, "Market sentiment-aware deep reinforcement learning approach for stock portfolio allocation," *Engineering Science and Technology, an International Journal*, vol. 24, no. 4, pp. 848–859, 2021.
- [24] F. Billert and S. Conrad, "A Framework for the Construction of a Sentiment-Driven Performance Index: The Case of DAX40," *The AAAI-2023 Workshop On Multimodal AI For Financial Forecasting*, 2 2023.
- [25] S. Davoodi, A. Fereydooni, and M. A. Rastegar, "Can portfolio construction considering ESG still gain high profits?" *Research in International Business and Finance*, vol. 67, p. 102126, 2024.
- [26] T. Nundlall and T. L. V. Zyl, "Machine Learning for Socially Responsible Portfolio Optimisation," *Proceedings of the 2023 7th International Conference on Intelligent Systems, Metaheuristics & Swarm Intelligence*, pp. 1–6, 2023.
- [27] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 26, no. 1, pp. 43–49, 1978.
- [28] T. K. Vintsyuk, "Speech discrimination by dynamic programming," *Cybernetics*, vol. 4, no. 1, pp. 52–57, 1968.
- [29] H. Skutkova, M. Vitek, P. Babula, R. Kizek, and I. Provaznik, "Classification of genomic signals using dynamic time warping," *BMC Bioinformatics*, vol. 14, no. Suppl 10, p. S1, 2013.
- [30] H.-S. Lee, "Application of dynamic time warping algorithm for pattern similarity of gait," *Journal of Exercise Rehabilitation*, vol. 15, no. 4, pp. 526–530, 2019.
- [31] T. Lim and H. Ng, "Financial Portfolio Management Based on Shaped-Based Unsupervised Machine Learning: A Dynamic Time Warping Baycenter Averaging Approach to International Markets and Periods of Downside Event Risks," *The Journal of Investing*, vol. 32, no. 2, pp. 74–96, 2022.
- [32] H. He and H. Li, "A New Boosting Algorithm for Online Portfolio Selection Based on dynamic Time Warping and Anti-correlation," *Computational Economics*, vol. 63, no. 5, pp. 1777–1803, 2024.
- [33] Y. Zhang, M. Cucuringu, A. Shestopaloff, and S. Zohren, "Dynamic Time Warping for Lead-Lag Relationship Detection in Lagged Multi-Factor Models," *4th ACM International Conference on AI in Finance*, pp. 454–462, 2023.
- [34] W. Meert, K. Hendrickx, A. Yurtman, P. Robberechts, D. Vohl, E. Ma, G. Verbruggen, M. Rossi, M. Shaikh, M. Yasirroni, Todd, W. Zieliński, T. V. Craenendonck, and S. Wu, "DTAIDistance," 8 2020. [Online]. Available: <https://doi.org/10.5281/zenodo.7158824>
- [35] S. Diamond and S. Boyd, "CVXPY: A Python-Embedded Modeling Language for Convex Optimization," *arXiv*, 2016.
- [36] A. Agrawal, R. Verschuere, S. Diamond, and S. Boyd, "A rewriting system for convex optimization problems," *Journal of Control and Decision*, vol. 5, no. 1, pp. 42–60, 2018.

7

CONCLUSION

In this thesis, we presented our contributions to the research field of financial information extraction, both in general and for the case of financial sustainability. In the following, we will summarize our findings, put them into context and outline further research directions.

7.1 Summary

In this thesis, we investigated several different approaches of extracting financial information from textual data, with a heavy focus on news articles specifically.

In [chapter 2](#), we introduce several foundational topics that underpin the work presented in this thesis. We begin by summarizing the key model paradigms that have transformed the field of natural language processing (NLP) over the four years during which this thesis was developed. In particular, we discuss the strengths and limitations of encoder-based representational models and decoder-based generative models. We then delve into modular variants of transformer-based architectures, with a focus on bottleneck adapters, which are employed in several of the subsequent chapters. Finally, we turn to the field of financial NLP, first reviewing prominent work in financial sentiment analysis, and then examining how sentiment signals can inform financial decision-making.

One of the realizations of this last section in particular was that financial datasets are still quite sparse, as most models rely on just a few popular datasets. To work on improving this, we introduced a novel library for financial data acquisition in [chapter 3](#) called `edgar-crawler`, which is able to download and parse important financial reports from the SEC’s EDGAR database [\[7\]](#).

In [chapter 4](#) we explored the representational model approach and applied the modular bottleneck adapter architecture introduced in [section 2.2](#) to train several adapters which can extract different kinds of financial information from news articles [\[3, 4, 2\]](#). Using Adapter Fusion [\[112\]](#), we showed that we can efficiently compose the adapter-specific knowledge for better predictions.

We continued by exploring the second model paradigm we discussed in [chapter 5](#) using generative Large Language Models (LLMs) to extract sustainability related financial data from news articles. We create a complex dataset called Nano-ESG with more than 50.000 data points containing time-dependent ESG-information for more than 40 German companies [\[6\]](#).

In the last chapter, we leveraged the learnings of the previous sections and used the extracted sentiment information for concrete investment decisions. First, we investigated creating a sentiment-portfolio to the market-capitalization weighted DAX-40 using general sentiments mined from news article headlines [\[1\]](#). Afterwards, we built upon the Nano-ESG dataset introduced in [section 5](#) and created a portfolio approach which detects ESG-aspect-regimes in order to optimize the stock allocation [\[5\]](#).

The works presented in this thesis contribute to various areas within the field of financial NLP. Reflecting the broader evolution of NLP paradigms - from representational to generative models - we systematically explored and evaluated these approaches for financial information extraction. Finally, we bridge the gap between academic research and real-world application by building financial instruments that leverage the results of our work.

7.2 Applicability Beyond the Financial Domain

Although the presented works focus entirely on finance-related textual data, as we have touched upon in [section 7.2](#), the different techniques we utilized can be transferred to other domains.

Pre-trained Representational Models In [section 2.3](#), we introduced two models pre-trained on the financial domain, both called FinBERT. Researchers have similarly pre-trained representational models in order to adapt them to specific domains. Chalkidis et al. [\[20\]](#) for example conduct a thorough study of how to adapt BERT to the legal domain and releases several adapted models. In the Biomedical domain, Lee et al. [\[76\]](#) train BioBERT, which is based on BERT but further pre-trained on biomedical corpora. Similarly to models used in the financial domain, these domain-specific models can be fine-tuned for specific downstream tasks.

Adapters in different Domains We presented different approaches to employing specifically bottleneck adapters on problems in the financial domain. However, adapters themselves are domain-agnostic. Several works have successfully used them in other domains. Xie et al. [\[165\]](#) for example successfully infuse knowledge represented in the PICO (Population, Intervention, Comparison and Outcome) framework into knowledge adapters and then apply Adapter Fusion to improve results on extractive summarization. They find that this approach generally outperforms simply finetuning the base model. Lai et al. [\[75\]](#) more generally train different adapters in the biomedical domain and combine them through Adapter Fusion, showing that this improves the overall performance. Chalkidis et al. [\[19\]](#) create a multilingual dataset in the legal domain and test it on different fine-tuning strategies. Similar to our findings in [chapter 4](#), they find that adapters provide superior results when transferring knowledge to different languages, and also perform best when trained on all available languages at once. Fi-

nally, adapters and Adapter Fusion have been applied in different works, for example reducing bias [63, 73] and improving fact-checking [155].

Generative Models in Zero-Shot Scenarios In chapter 5, we used generative Language Models in order to extract information from text in a zero-shot manner. Recently, LLMs have increasingly been used as annotators, and their annotation capabilities have been evaluated in different domains, as we did for the subject of corporate sustainability. Previous works found that such models can already outperform human crowd workers in many scenarios [153, 79, 50]. In the financial domain in general, Li et al. [79] study different tasks and finds that zero-shot GPT-4 exceeds fine-tuned models in several of them.

In other domains the results are not as convincing: Savelka et al. [130] investigate the zero-shot performance of GPT-based models on different tasks in the legal domain and finds that GPT-4 especially shows promising results, although it still lags behind fine-tuned models in some of the cases. At the same time, in the biomedical domain, Chen et al. [24] find that LLMs like GPT-4 are still outperformed by fine-tuned BERT-based models in most tasks. However, both of these studies were conducted in 2023 and thus only include results from somewhat outdated models. More recently, Brodeur et al. [15] found that for medical tasks, both GPT-4o as well as the more recent o1 reasoning model¹ outperform humans. Yu et al. [173] similarly observe improved performance of reasoning models on legal tasks, although they still struggle on several tasks. In the chemical domain, Sharlin et al. [132] find that while GPT-4o struggles on Nuclear Magnetic Resonance (NMR) structure determination tasks, reasoning models (and specifically o1) outperform top-performing undergraduate organic chemistry students.

7.3 Future Work

With the general advancement of the field of NLP, new possibilities for the different areas of financial NLP specifically emerge continuously.

In chapter 3, we introduced `edgar-crawler`, which enables users to download and preprocess financial reports from the EDGAR database. A logical next step would be to advance the tool and include a way to automatically answer questions users have for a given report. As LLMs still struggle with processing long documents, this poses a substantial challenge. Specifically, for the SEC reports we work with here, Reddy et al. [124] find that financial QA over the typically long 10-K reports is difficult for LLMs even when first retrieving the portions of the report important for the question. Investigating effective ways to improve financial QA would thus be a first necessary step for the proposed application.

At the time of writing the works in chapter 4, most representational models were still based on BERT [31] or RoBERTa [83]. However, as we discussed in subsection 2.1.1, newer models such as ModernBERT [161] were published recently, with general improvements compared to the older base models. As the popular Adapters library [113, 116] now also supports ModernBERT (as of May 2025), it might be possible to improve upon our reported performances by implementing the more novel base

¹<https://openai.com/de-DE/index/introducing-openai-o1-preview/>

model.

In addition, while the older models were limited to 512 tokens, models such as ModernBERT can handle up to 8192 tokens. At the time, the sentiment-based index we constructed in [section 6.1](#) did not include the complete content of the news articles and solely relied on the headlines because of this limitation. After we created the Nano-ESG dataset, we counted how often a company is mentioned in the headline in all articles deemed relevant for it. Surprisingly, in merely 20% of the articles, this was the case. While Nano-ESG concerns only sustainability-related topics where the main information might not be mentioned in the headline directly, it is safe to assume that our work in [\[1\]](#) can be improved if more relevant articles are considered by also scanning the article content for sentiment information. Here, another possibility would be to use a similar approach as we did in [chapter 5](#) and use generative models to extract general sentiment information of companies.

Other research directions could investigate how different GLMs perform when extracting general, or sustainability-specific information from news articles - in [\[6\]](#), we only investigated models in the GPT-family. With more powerful models published after we conducted our work, it might be interesting to study the variance in sentiment-determination between them.

More specifically, we were interested by the Relevance-Score extracted by GPT-4o. In the paper we note that for articles with a lower score, the evaluators thought they were irrelevant more frequently than for articles with a higher score. What we did not include in our paper because of the limitations of space, is that GPT-4o-mini assigned scores in a much more narrow segment compared to the more powerful GPT-4o. The smaller model only determined scores between 6 and 9, while the bigger model's scores were more diverse. At the same time, the oldest model we used - GPT-3.5-turbo - only assigned scores of 7 or 8, continuing the trend. Besides experimenting with more different models, in future approaches, it might be interesting to study how GLMs determine these more subjective properties in other domains.

We mentioned one of the major motivations for creating the Nano-ESG dataset in [subsection 2.3.3](#) - determining ESG-impact from news articles is much more transparent than traditional ESG-scores. In order to provide the general public with continued, transparent ESG-data, a web-service which displays up-to-date information would be an interesting application to build in the future. It would even be possible to include the topic-detection approach we showcase in Billert et al. [\[6\]](#) so that the most relevant topics are accessible at a glance.

In conclusion, the possibilities for future research are very open. One of the biggest advantages of the modern world of Artificial Intelligence is that it is easier than ever before to explore creative ideas, experiment with novel approaches, and develop impactful solutions.



PUBLICATIONS

Contributing Publications

- [1] Fabian Billert and Stefan Conrad. A Framework for the Construction of a Sentiment-Driven Performance Index: The Case of DAX40. *The AAAI-2023 Workshop On Multimodal AI For Financial Forecasting*, 2 2023. [arXiv:2409.20397](https://arxiv.org/abs/2409.20397).
- [2] Fabian Billert and Stefan Conrad. Exploring Knowledge Composition for ESG Impact Type Determination. *Proceedings of the Sixth Workshop on Financial Technology and Natural Language Processing*, pages 79–83, 11 2023. URL: <https://aclanthology.org/2023.finnlp-2.12/>, doi:10.18653/v1/2023.finnlp-2.12.
- [3] Fabian Billert and Stefan Conrad. HHU at SemEval-2023 Task 3: An Adapter-based Approach for News Genre Classification. *Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 1166–1171, 2023. doi:10.18653/v1/2023.semeval-1.162.
- [4] Fabian Billert and Stefan Conrad. Team HHU at the FinNLP-2023 ML-ESG Task: A Multi-Model Approach to ESG-Key-Issue Classification. *Proceedings of the Fifth Workshop on Financial Technology and Natural Language Processing and the Second Multimodal AI For Financial Forecasting*, pages 146–150, 8 2023. URL: <https://aclanthology.org/2023.finnlp-1.17>.
- [5] Fabian Billert and Stefan Conrad. Identifying Aspect-Regimes for Enhanced ESG-Investing Through News Data. *2025 IEEE Symposium on Computational Intelligence for Financial Engineering and Economics (CiFer)*, 00:1–7, 2025. doi:10.1109/cifer64978.2025.10975736.
- [6] Fabian Billert and Stefan Conrad. Nano-ESG: Extracting Corporate Sustainability Information from News Articles. *Advances in Information Retrieval, 47th European Conference on Information Retrieval, ECIR 2025, Lucca, Italy, April 6–10, 2025*,

- Proceedings, Part IV*, pages 324–338, 2025. [doi:10.1007/978-3-031-88717-8_24](https://doi.org/10.1007/978-3-031-88717-8_24).
- [7] Lefteris Loukas, Fabian Billert, Manos Fergadiotis, Prodromos Malakasiotis, and Ion Androutsopoulos. EDGAR-CRAWLER: From Raw Web Documents to Structured Financial NLP Datasets. *Companion Proceedings of the ACM on Web Conference 2025*, pages 761–764, 2025. [doi:10.1145/3701716.3715289](https://doi.org/10.1145/3701716.3715289).

BIBLIOGRAPHY

- [8] Toyin Aguda, Suchetha Siddagangappa, Elena Kochkina, Simerjot Kaur, Dongsheng Wang, Charese Smiley, and Sameena Shah. Large Language Models as Financial Data Annotators: A Study on Effectiveness and Efficiency. *arXiv*, 2024. [arXiv:2403.18152](#), [doi:10.48550/arxiv.2403.18152](#).
- [9] Amir Amel-Zadeh and Jonathan Faasse. The Information Content of 10-K Narratives: Comparing MD&A and Footnotes Disclosures. *SSRN Electronic Journal*, 2016. [doi:10.2139/ssrn.2807546](#).
- [10] Dogu Araci. FinBERT: Financial Sentiment Analysis with Pre-trained Language Models. *arXiv*, 2019. [arXiv:1908.10063](#).
- [11] Maria Cristina Arcuri, Gino Gandolfi, and Ivan Russo. Does fake news impact stock returns? Evidence from US and EU stock markets. *Journal of Economics and Business*, 125:106130, 2023. [doi:10.1016/j.jeconbus.2023.106130](#).
- [12] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. Qwen Technical Report. *arXiv*, 2023. [arXiv:2309.16609](#), [doi:10.48550/arxiv.2309.16609](#).
- [13] Florian Berg, Julian F Kölbel, and Roberto Rigobon. Aggregate Confusion: The Divergence of ESG Ratings. *SSRN Electronic Journal*, 2019. [doi:10.2139/ssrn.3438533](#).
- [14] Zvi Bodie, Alex Kane, and Alan J. Marcus. *Investments*. Mc Graw Hill Education, 2014.
- [15] Peter G Brodeur, Thomas A Buckley, Zahir Kanjee, Ethan Goh, Evelyn Bin Ling, Priyank Jain, Stephanie Cabral, Raja-Elie Abdunour, Adrian D Haimovich, Jason A Freed, Andrew Olson, Daniel J Morgan, Jason Hom, Robert Gallo, Liam G McCoy, Haadi Mombini, Christopher Lucas, Misha Fotoohi, Matthew

- Gwiazdon, Daniele Restifo, Daniel Restrepo, Eric Horvitz, Jonathan Chen, Arjun K Manrai, and Adam Rodman. Superhuman performance of a large language model on the reasoning tasks of a physician. *arXiv*, 2024. [arXiv:2412.10849](#), [doi:10.48550/arxiv.2412.10849](#).
- [16] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language Models are Few-Shot Learners. *arXiv*, 2020. [arXiv:2005.14165](#), [doi:10.48550/arxiv.2005.14165](#).
- [17] Patrice B  chard and Orlando Marquez Ayala. Reducing hallucination in structured outputs via Retrieval-Augmented Generation. *arXiv*, 2024. [arXiv:2404.08189](#), [doi:10.48550/arxiv.2404.08189](#).
- [18] Daniel Celeny and Lo  c Mar  chal. Cyber risk and the cross-section of stock returns. *arXiv*, 2024. [arXiv:2402.04775](#), [doi:10.48550/arxiv.2402.04775](#).
- [19] Ilias Chalkidis, Manos Fergadiotis, and Ion Androutsopoulos. MultiEURLEX - A multi-lingual and multi-label legal document classification dataset for zero-shot cross-lingual transfer. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6974–6996, 2021. [doi:10.18653/v1/2021.emnlp-main.559](#).
- [20] Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. LEGAL-BERT: The Muppets straight out of Law School. *arXiv*, 2020. [arXiv:2010.02559](#), [doi:10.48550/arxiv.2010.02559](#).
- [21] Le Chau, Le Anh, and Vo Duc. Valuing ESG: How financial markets respond to corporate sustainability. *International Business Review*, 34(3):102418, 2025. [doi:10.1016/j.ibusrev.2025.102418](#).
- [22] Chung-Chi Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. NLP in FinTech Applications: Past, Present and Future. *arXiv*, 2020. [arXiv:2005.01320](#), [doi:10.48550/arxiv.2005.01320](#).
- [23] Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. BGE M3-Embedding: Multi-Lingual, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation. *arXiv*, 2024. [arXiv:2402.03216](#), [doi:10.48550/arxiv.2402.03216](#).
- [24] Qingyu Chen, Yan Hu, Xueqing Peng, Qianqian Xie, Qiao Jin, Aidan Gilson, Maxwell B. Singer, Xuguang Ai, Po-Ting Lai, Zhizheng Wang, Vipina K. Keloth, Kalpana Raja, Jimin Huang, Huan He, Fongci Lin, Jingcheng Du, Rui Zhang, W. Jim Zheng, Ron A. Adelman, Zhiyong Lu, and Hua Xu. Benchmarking large language models for biomedical natural language processing applications and recommendations. *Nature Communications*, 16(1):3280, 2025. [arXiv:2305.16326](#), [doi:10.1038/s41467-025-56989-2](#).

-
- [25] Thomson Reuters Text Research Collection. Reuters-TRC2, 2004. URL: <https://trec.nist.gov/data/reuters/reuters.html>.
 - [26] Keith Cortis, André Freitas, Tobias Daudert, Manuela Huerlimann, Manel Zarrouk, Siegfried Handschuh, and Brian Davis. SemEval-2017 Task 5: Fine-Grained Sentiment Analysis on Financial Microblogs and News. *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 519–535, 2017. doi:10.18653/v1/s17-2089.
 - [27] Dan Cosley, Andrea Forte, Luigina Ciolfi, David McDonald, Gang Wang, Tianyi Wang, Bolun Wang, Divya Sambasivan, Zengbin Zhang, Haitao Zheng, and Ben Y Zhao. Crowds on Wall Street. *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, pages 17–30, 2015. doi:10.1145/2675133.2675144.
 - [28] Leland D Crane, Emily Green, Molly Harnish, Will McClennan, Paul E Soto, Betsy Vrankovich, and Jacob Williams. Tracking Real Time Layoffs with SEC Filings: A Preliminary Investigation. *Finance and Economics Discussion Series*, (2024-020):1–30, 2024. doi:10.17016/feds.2024.020.
 - [29] Matthew Dahl, Varun Magesh, Mirac Suzgun, and Daniel E Ho. Large Legal Fictions: Profiling Legal Hallucinations in Large Language Models. *arXiv*, 2024. arXiv:2401.01301, doi:10.48550/arxiv.2401.01301.
 - [30] John R. Deller and John Hansen. The Electrical Engineering Handbook. VII: *Signal Processing*, (Inequalities31972):861–890, 2005. doi:10.1016/b978-012170960-0/50063-3.
 - [31] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv*, 2018. arXiv:1810.04805.
 - [32] Bosheng Ding, Chengwei Qin, Linlin Liu, Yew Ken Chia, Boyang Li, Shafiq Joty, and Lidong Bing. Is GPT-3 a Good Data Annotator? *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11173–11195, 2023. doi:10.18653/v1/2023.acl-long.626.
 - [33] Xiao Ding, Yue Zhang, Ting Liu, and Junwen Duan. Deep Learning for Event-Driven Stock Prediction. *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence (IJCAI 2015)*, 2015.
 - [34] Kelvin Du, Frank Xing, Rui Mao, and Erik Cambria. Financial Sentiment Analysis: Techniques and Applications. *ACM Computing Surveys*, 56(9):1–42, 2024. doi:10.1145/3649451.
 - [35] Kelvin Du, Yazhi Zhao, Rui Mao, Frank Xing, and Erik Cambria. Natural language processing in finance: A survey. *Information Fusion*, 115:102755, 2025. doi:10.1016/j.inffus.2024.102755.

- [36] Ersan Ersoy, Beata Swiecka, Simon Grima, Ercan Özen, and Inna Romanova. The Impact of ESG Scores on Bank Market Value? Evidence from the U.S. Banking Industry. *Sustainability*, 14(15):9527, 2022. doi:[10.3390/su14159527](https://doi.org/10.3390/su14159527).
- [37] Elena Escrig-Olmedo, María Ángeles Fernández-Izquierdo, Idoia Ferrero-Ferrero, Juana María Rivera-Lirio, and María Jesús Muñoz-Torres. Rating the Raters: Evaluating how ESG Rating Agencies Integrate Sustainability Principles. *Sustainability*, 11(3):915, 2019. doi:[10.3390/su11030915](https://doi.org/10.3390/su11030915).
- [38] Eugene F. Fama and Kenneth R. French. Luck versus Skill in the Cross-Section of Mutual Fund Returns. *The Journal of Finance*, 65(5):1915–1947, 2010. doi:[10.1111/j.1540-6261.2010.01598.x](https://doi.org/10.1111/j.1540-6261.2010.01598.x).
- [39] Kezia Farnham. ESG scores and ratings: What they are, why they matter, 2023. URL: <https://www.diligent.com/resources/blog/esg-risk-scores>.
- [40] Xin Feng, Hans-Jörg von Mettenheim, Georgios Sermpinis, and Charalampos Stasinakis. Sustainable Portfolio Construction via Machine Learning: ESG, SDG and Sentiment. *European Financial Management*, 31(3):1148–1169, 2025. doi:[10.1111/eufm.12531](https://doi.org/10.1111/eufm.12531).
- [41] Katja Filippova. Controlled Hallucinations: Learning to Generate Faithfully from Noisy Data. *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 864–870, 2020. doi:[10.18653/v1/2020.findings-emnlp.76](https://doi.org/10.18653/v1/2020.findings-emnlp.76).
- [42] Gunnar Friede, Timo Busch, and Alexander Bassen. ESG and financial performance: aggregated evidence from more than 2000 empirical studies. *Journal of Sustainable Finance & Investment*, 5(4):210–233, 2015. doi:[10.1080/20430795.2015.1118917](https://doi.org/10.1080/20430795.2015.1118917).
- [43] Gary L. Gastineau. *The Exchange-Traded Funds Manual*. John Wiley & Sons, Ltd, 2011. URL: <https://onlinelibrary.wiley.com/doi/book/10.1002/9781118266946>, doi:[10.1002/9781118266946](https://doi.org/10.1002/9781118266946).
- [44] Carlos-Emiliano González-Gallardo, Emanuela Boros, Nancy Girdhar, Ahmed Hamdi, Jose G Moreno, and Antoine Doucet. Yes but.. Can ChatGPT Identify Entities in Historical Documents? *arXiv*, 2023. arXiv:[2303.17322](https://arxiv.org/abs/2303.17322), doi:[10.48550/arxiv.2303.17322](https://doi.org/10.48550/arxiv.2303.17322).
- [45] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. URL: <http://www.deeplearningbook.org>.
- [46] Ian J Goodfellow, Mehdi Mirza, Da Xiao, Aaron Courville, and Yoshua Bengio. An Empirical Investigation of Catastrophic Forgetting in Gradient-Based Neural Networks. *arXiv*, 2013. arXiv:[1312.6211](https://arxiv.org/abs/1312.6211), doi:[10.48550/arxiv.1312.6211](https://doi.org/10.48550/arxiv.1312.6211).
- [47] Stephane Goutte, Fei Liu, Hoang Viet Le, and Hans-Jorg von Mettenheim. ESG Investing: A Sentiment Analysis Approach. *SSRN Electronic Journal*, 2023. doi:[10.2139/ssrn.4316107](https://doi.org/10.2139/ssrn.4316107).

-
- [48] Manuel Gozzi and Federico Di Maio. Comparative Analysis of Prompt Strategies for Large Language Models: Single-Task vs. Multitask Prompts. *Electronics*, 13(23):4712, 2024. doi:[10.3390/electronics13234712](https://doi.org/10.3390/electronics13234712).
 - [49] Vasileios Hatzivassiloglou and Kathleen R. McKeown. Predicting the semantic orientation of adjectives. *Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics -*, pages 174–181, 1997. doi:[10.3115/979617.979640](https://doi.org/10.3115/979617.979640).
 - [50] Zeyu He, Chieh-Yang Huang, Chien-Kuang Cornelia Ding, Shaurya Rohatgi, and Ting-Hao 'Kenneth' Huang. If in a Crowdsourced Data Annotation Pipeline, a GPT-4. *arXiv*, 2024. arXiv:[2402.16795](https://arxiv.org/abs/2402.16795), doi:[10.1145/3613904.3642834](https://doi.org/10.1145/3613904.3642834).
 - [51] Roy Henriksson, Joshua Livnat, Patrick Pfeifer, and Margaret Stumpp. Integrating ESG in Portfolio Construction. *The Journal of Portfolio Management*, 45(4):67–81, 2019. doi:[10.3905/jpm.2019.45.4.067](https://doi.org/10.3905/jpm.2019.45.4.067).
 - [52] Trang-Thi Ho and Yennun Huang. Stock Price Movement Prediction Using Sentiment Analysis and CandleStick Chart Representation. *Sensors (Basel, Switzerland)*, 21(23):7957, 2021. doi:[10.3390/s21237957](https://doi.org/10.3390/s21237957).
 - [53] Sepp Hochreiter and Jrgen Schmidhuber. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780, 1997. doi:[10.1162/neco.1997.9.8.1735](https://doi.org/10.1162/neco.1997.9.8.1735).
 - [54] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W Rae, Oriol Vinyals, and Laurent Sifre. Training Compute-Optimal Large Language Models. *arXiv*, 2022. arXiv:[2203.15556](https://arxiv.org/abs/2203.15556), doi:[10.48550/arxiv.2203.15556](https://doi.org/10.48550/arxiv.2203.15556).
 - [55] Zhi Hong, Logan Ward, Kyle Chard, Ben Blaiszik, and Ian Foster. Challenges and Advances in Information Extraction from Scientific Literature: a Review. *JOM*, 73(11):3383–3400, 2021. doi:[10.1007/s11837-021-04902-9](https://doi.org/10.1007/s11837-021-04902-9).
 - [56] Benjamin D Horne and Sibel Adali. This Just In: Fake News Packs a Lot in Title, Uses Simpler, Repetitive Content in Text Body, More Similar to Satire than Real News. *arXiv*, 2017. arXiv:[1703.09398](https://arxiv.org/abs/1703.09398), doi:[10.48550/arxiv.1703.09398](https://doi.org/10.48550/arxiv.1703.09398).
 - [57] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-Efficient Transfer Learning for NLP. *arXiv*, 2019. arXiv:[1902.00751](https://arxiv.org/abs/1902.00751).
 - [58] Bo Hu, Zhendong Mao, and Yongdong Zhang. An overview of fake news detection: From a new perspective. *Fundamental Research*, 5(1):332–346, 2025. doi:[10.1016/j.fmre.2024.01.017](https://doi.org/10.1016/j.fmre.2024.01.017).
 - [59] Ke-Wei Huang. Exploring the Information Contents of Risk Factors in SEC Form 10-K: A Multi-Label Text Classification Application. *SSRN Electronic Journal*, 2010. doi:[10.2139/ssrn.1784527](https://doi.org/10.2139/ssrn.1784527).

- [60] EunJeong Hwang and Yong-Hyuk Kim. Interdependency between the Stock Market and Financial News. *2019 IEEE International Conference on Big Data (Big Data)*, 00:6074–6076, 2019. doi:[10.1109/bigdata47090.2019.9006533](https://doi.org/10.1109/bigdata47090.2019.9006533).
- [61] Konstantin Ignatov. When ESG talks: ESG tone of 10-K reports and its significance to stock markets. *International Review of Financial Analysis*, 89:102745, 2023. doi:[10.1016/j.irfa.2023.102745](https://doi.org/10.1016/j.irfa.2023.102745).
- [62] S&P Dow Jones Indices. SPIVA® U.S. Scorecard. Technical report, 2024. URL: <https://www.spglobal.com/spdji/en/documents/spiva/spiva-us-year-end-2024.pdf>.
- [63] Prachi Jain, Ashutosh Sathe, Varun Gumma, Kabir Ahuja, and Sunayana Sitaram. MAFIA: Multi-Adapter Fused Inclusive Language Models. *arXiv*, 2024. arXiv:[2402.07519](https://arxiv.org/abs/2402.07519), doi:[10.48550/arxiv.2402.07519](https://doi.org/10.48550/arxiv.2402.07519).
- [64] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Delong Chen, Wenliang Dai, Ho Shu Chan, Andrea Madotto, and Pascale Fung. Survey of Hallucination in Natural Language Generation. *arXiv*, 2022. arXiv:[2202.03629](https://arxiv.org/abs/2202.03629), doi:[10.48550/arxiv.2202.03629](https://doi.org/10.48550/arxiv.2202.03629).
- [65] 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. *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 888–893, 2017. doi:[10.18653/v1/s17-2152](https://doi.org/10.18653/v1/s17-2152).
- [66] Harsimrat Kaeley, Ye Qiao, and Nader Bagherzadeh. Support for Stock Trend Prediction Using Transformers and Sentiment Analysis. *arXiv*, 2023. arXiv:[2305.14368](https://arxiv.org/abs/2305.14368), doi:[10.48550/arxiv.2305.14368](https://doi.org/10.48550/arxiv.2305.14368).
- [67] Haoqiang Kang and Xiao-Yang Liu. Deficiency of Large Language Models in Finance: An Empirical Examination of Hallucination. *arXiv*, 2023. arXiv:[2311.15548](https://arxiv.org/abs/2311.15548), doi:[10.48550/arxiv.2311.15548](https://doi.org/10.48550/arxiv.2311.15548).
- [68] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling Laws for Neural Language Models. *arXiv*, 2020. arXiv:[2001.08361](https://arxiv.org/abs/2001.08361), doi:[10.48550/arxiv.2001.08361](https://doi.org/10.48550/arxiv.2001.08361).
- [69] Colm Kearney and Sha Liu. Textual sentiment in finance: A survey of methods and models. *International Review of Financial Analysis*, 33:171–185, 2014. doi:[10.1016/j.irfa.2014.02.006](https://doi.org/10.1016/j.irfa.2014.02.006).
- [70] Mozaffar Khan, George Serafeim, and Aaron Yoon. Corporate Sustainability: First Evidence on Materiality. *SSRN Electronic Journal*, 2016. doi:[10.2139/ssrn.2575912](https://doi.org/10.2139/ssrn.2575912).
- [71] Yubin Kim, Hyewon Jeong, Shan Chen, Shuyue Stella Li, Mingyu Lu, Kmail Alhamoud, Jimin Mun, Cristina Grau, Minseok Jung, Rodrigo Gameiro, Lizhou Fan, Eugene Park, Tristan Lin, Joonsik Yoon, Wonjin Yoon, Maarten

- Sap, Yulia Tsvetkov, Paul Liang, Xuhai Xu, Xin Liu, Daniel McDuff, Hyeonhoon Lee, Hae Won Park, Samir Tulebaev, and Cynthia Breazeal. Medical Hallucinations in Foundation Models and Their Impact on Healthcare. *arXiv*, 2025. [arXiv:2503.05777](https://arxiv.org/abs/2503.05777), [doi:10.48550/arxiv.2503.05777](https://doi.org/10.48550/arxiv.2503.05777).
- [72] Prahlad Koratamaddi, Karan Wadhwani, Mridul Gupta, and Sriram G. Sanjeevi. Market sentiment-aware deep reinforcement learning approach for stock portfolio allocation. *Engineering Science and Technology, an International Journal*, 24(4):848–859, 2021. [doi:10.1016/j.jestch.2021.01.007](https://doi.org/10.1016/j.jestch.2021.01.007).
- [73] Deepak Kumar, Oleg Lesota, George Zerveas, Daniel Cohen, Carsten Eickhoff, Markus Schedl, and Navid Rekabsaz. Parameter-efficient Modularised Bias Mitigation via AdapterFusion. *arXiv*, 2023. [arXiv:2302.06321](https://arxiv.org/abs/2302.06321), [doi:10.48550/arxiv.2302.06321](https://doi.org/10.48550/arxiv.2302.06321).
- [74] Aditya Kuppa, Nikon Rasumov-Rahe, and Marc Voses. Chain Of Reference prompting helps LLM to think like a lawyer. *Generative AI and Law (GenLaw) Workshop Proceedings*, 2021.
- [75] Tuan Manh Lai, ChengXiang Zhai, and Heng Ji. KEBLM: Knowledge-Enhanced Biomedical Language Models. *Journal of Biomedical Informatics*, 143:104392, 2023. [doi:10.1016/j.jbi.2023.104392](https://doi.org/10.1016/j.jbi.2023.104392).
- [76] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2019. [arXiv:1901.08746](https://arxiv.org/abs/1901.08746), [doi:10.1093/bioinformatics/btz682](https://doi.org/10.1093/bioinformatics/btz682).
- [77] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *arXiv*, 2020. [arXiv:2005.11401](https://arxiv.org/abs/2005.11401), [doi:10.48550/arxiv.2005.11401](https://doi.org/10.48550/arxiv.2005.11401).
- [78] Junyi Li, Jie Chen, Ruiyang Ren, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. The Dawn After the Dark: An Empirical Study on Factuality Hallucination in Large Language Models. *arXiv*, 2024. [arXiv:2401.03205](https://arxiv.org/abs/2401.03205), [doi:10.48550/arxiv.2401.03205](https://doi.org/10.48550/arxiv.2401.03205).
- [79] Xianzhi Li, Samuel Chan, Xiaodan Zhu, Yulong Pei, Zhiqiang Ma, Xiaomo Liu, and Sameena Shah. Are ChatGPT and GPT-4 General-Purpose Solvers for Financial Text Analytics? A Study on Several Typical Tasks. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 408–422, 2023. [doi:10.18653/v1/2023.emnlp-industry.39](https://doi.org/10.18653/v1/2023.emnlp-industry.39).
- [80] Yinheng Li, Shaofei Wang, Han Ding, and Hang Chen. Large Language Models in Finance: A Survey. *arXiv*, 2023. [arXiv:2311.10723](https://arxiv.org/abs/2311.10723), [doi:10.48550/arxiv.2311.10723](https://doi.org/10.48550/arxiv.2311.10723).

- [81] Xiaoqiang Lin, Zhongxiang Dai, Arun Verma, See-Kiong Ng, Patrick Jaillet, and Bryan Kian Hsiang Low. Prompt Optimization with Human Feedback. *arXiv*, 2024. [arXiv:2405.17346](#), [doi:10.48550/arxiv.2405.17346](#).
- [82] Jian Liu, Jianxin Yao, and Gengfeng Wu. Advances in Intelligent Data Analysis VI, 6th International Symposium on Intelligent Data Analysis, IDA 2005, Madrid, Spain, September 8-10, 2005. Proceedings. *Lecture Notes in Computer Science*, pages 216–227, 2005. [doi:10.1007/11552253_20](#).
- [83] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv*, 2019. [arXiv:1907.11692](#), [doi:10.48550/arxiv.1907.11692](#).
- [84] Yuchen Liu, Daniil Mikriukov, Owen Christopher Tjahyadi, Gangmin Li, Terry R. Payne, Yong Yue, Kamran Siddique, and Ka Lok Man. Revolutionising Financial Portfolio Management: The Non-Stationary Transformer’s Fusion of Macroeconomic Indicators and Sentiment Analysis in a Deep Reinforcement Learning Framework. *Applied Sciences*, 14(1):274, 2023. [doi:10.3390/app14010274](#).
- [85] Zhuang Liu, Degen Huang, Kaiyu Huang, Zhuang Li, and Jun Zhao. FinBERT: A Pre-trained Financial Language Representation Model for Financial Text Mining. *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, pages 4513–4519, 2020. [doi:10.24963/ijcai.2020/622](#).
- [86] MSCI ESG RESEARCH LLC. ESG Ratings Methodology, 2024. URL: <https://www.msci.com/documents/1296102/34424357/MSCI+ESG+Ratings+Methodology.pdf>.
- [87] Tim Loughran and Bill McDonald. When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks. *The Journal of Finance*, 66(1):35–65, 2011. [doi:10.1111/j.1540-6261.2010.01625.x](#).
- [88] Nicolas Madison and Eduardo Schiehl. The Effect of Financial Materiality on ESG Performance Assessment. *Sustainability*, 13(7):3652, 2021. [doi:10.3390/su13073652](#).
- [89] Varun Magesh, Faiz Surani, Matthew Dahl, Mirac Suzgun, Christopher D Manning, and Daniel E Ho. Hallucination-Free? Assessing the Reliability of Leading AI Legal Research Tools. *arXiv*, 2024. [arXiv:2405.20362](#), [doi:10.48550/arxiv.2405.20362](#).
- [90] Macedo Maia, Siegfried Handschuh, André Freitas, Brian Davis, Ross McDermott, Manel Zarrouk, and Alexandra Balahur. WWW’18 Open Challenge. *Companion of the The Web Conference 2018 on The Web Conference 2018 - WWW ’18*, pages 1941–1942, 2018. [doi:10.1145/3184558.3192301](#).
- [91] Pekka Malo, Ankur Sinha, Pekka Korhonen, Jyrki Wallenius, and Pyry Takala. Good debt or bad debt: Detecting semantic orientations in economic texts. *Journal of the Association for Information Science and Technology*, 65(4):782–796, 2014. [doi:10.1002/asi.23062](#).

-
- [92] Anqi Mao, Mehryar Mohri, and Yutao Zhong. Cross-Entropy Loss Functions: Theoretical Analysis and Applications. *arXiv*, 2023. [arXiv:2304.07288](https://arxiv.org/abs/2304.07288), doi: [10.48550/arxiv.2304.07288](https://doi.org/10.48550/arxiv.2304.07288).
 - [93] Harry Markowitz. Portfolio Selection. *The Journal of Finance*, 7(1):77, 1952. doi: [10.2307/2975974](https://doi.org/10.2307/2975974).
 - [94] McKinsey. Does ESG really matter—and why?, 2022. URL: <https://www.mckinsey.com/capabilities/sustainability/our-insights/does-esg-really-matter-and-why>.
 - [95] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient Estimation of Word Representations in Vector Space. *arXiv*, 2013. [arXiv:1301.3781](https://arxiv.org/abs/1301.3781), doi: [10.48550/arxiv.1301.3781](https://doi.org/10.48550/arxiv.1301.3781).
 - [96] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed Representations of Words and Phrases and their Compositionality. *arXiv*, 2013. [arXiv:1310.4546](https://arxiv.org/abs/1310.4546), doi: [10.48550/arxiv.1310.4546](https://doi.org/10.48550/arxiv.1310.4546).
 - [97] Morningstar and Sustainalytics. Global Sustainable Fund Flows: Q1 2025 in Review. 2025. URL: <https://www.morningstar.com/business/insights/research/global-esg-flows>.
 - [98] MSCI. MSCI Analyst Sentiment Indexes Methodology, 2025. URL: <https://www.msci.com/documents/10199/a925c038-cf5e-9701-fe7a-e4414a06b1ca>.
 - [99] MSCI. MSCI World Analyst Sentiment Index, 2025. URL: <https://www.msci.com/documents/10199/8cd51f74-1924-f93d-4353-8f15ab5e9379>.
 - [100] Mufhumudzi Muthivhi and Terence L van Zyl. Fusion of Sentiment and Asset Price Predictions for Portfolio Optimization. *arXiv*, 2022. [arXiv:2203.05673](https://arxiv.org/abs/2203.05673).
 - [101] Sebastião Vieira de Freitas Netto, Marcos Felipe Falcão Sobral, Ana Regina Bezerra Ribeiro, and Gleibson Robert da Luz Soares. Concepts and forms of greenwashing: a systematic review. *Environmental Sciences Europe*, 32(1):19, 2020. doi: [10.1186/s12302-020-0300-3](https://doi.org/10.1186/s12302-020-0300-3).
 - [102] Jingwei Ni, Julia Bingler, Chiara Colesanti-Senni, Mathias Kraus, Glen Gostlow, Tobias Schimanski, Dominik Stambach, Saeid Ashraf Vaghefi, Qian Wang, Nicolas Webersinke, Tobias Wekhof, Tingyu Yu, and Markus Leippold. CHA-TREPORT: Democratizing Sustainability Disclosure Analysis through LLM-based Tools. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 21–51, 2023. doi: [10.18653/v1/2023.emnlp-demo.3](https://doi.org/10.18653/v1/2023.emnlp-demo.3).
 - [103] Matteo Olivato, Luca Putelli, Nicola Arici, Alfonso Emilio Gerevini, Alberto Lavelli, and Ivan Serina. Language Models for Hierarchical Classification of Radiology Reports With Attention Mechanisms, BERT, and GPT-4. *IEEE Access*, 12:69710–69727, 2024. doi: [10.1109/access.2024.3402066](https://doi.org/10.1109/access.2024.3402066).

- [104] Nuno Oliveira, Paulo Cortez, and Nelson Areal. Stock market sentiment lexicon acquisition using microblogging data and statistical measures. *Decision Support Systems*, 85:62–73, 2016. doi:[10.1016/j.dss.2016.02.013](https://doi.org/10.1016/j.dss.2016.02.013).
- [105] OpenAI. Introducing ChatGPT | OpenAI, 2022. URL: <https://openai.com/index/chatgpt/>.
- [106] OpenAI. GPT-4 Technical Report. *arXiv*, 2023. arXiv:[2303.08774](https://arxiv.org/abs/2303.08774).
- [107] OpenAI. Introducing GPT-4.5, 2025. URL: <https://openai.com/index/introducing-gpt-4-5/>.
- [108] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. *arXiv*, 2022. arXiv:[2203.02155](https://arxiv.org/abs/2203.02155), doi:[10.48550/arxiv.2203.02155](https://doi.org/10.48550/arxiv.2203.02155).
- [109] Daniel P. Palomar. *Portfolio Optimization: Theory and Application*. Cambridge University Press, 2025. URL: <https://portfoliooptimizationbook.com/>.
- [110] Jigar Patel, Sahil Shah, Priyank Thakkar, and K Kotecha. Predicting stock and stock price index movement using Trend Deterministic Data Preparation and machine learning techniques. *Expert Systems with Applications*, 42(1):259–268, 2015. doi:[10.1016/j.eswa.2014.07.040](https://doi.org/10.1016/j.eswa.2014.07.040).
- [111] Jonas Pfeiffer, Aishwarya Kamath, Andreas Rücklé, Kyunghyun Cho, and Iryna Gurevych. AdapterFusion: Non-Destructive Task Composition for Transfer Learning. *arXiv*, 2020. arXiv:[2005.00247](https://arxiv.org/abs/2005.00247), doi:[10.48550/arxiv.2005.00247](https://doi.org/10.48550/arxiv.2005.00247).
- [112] Jonas Pfeiffer, Aishwarya Kamath, Andreas Rücklé, Kyunghyun Cho, and Iryna Gurevych. AdapterFusion: Non-Destructive Task Composition for Transfer Learning. *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 487–503, 2021. doi:[10.18653/v1/2021.eacl-main.39](https://doi.org/10.18653/v1/2021.eacl-main.39).
- [113] Jonas Pfeiffer, Andreas Rücklé, Clifton Poth, Aishwarya Kamath, Ivan Vulić, Sebastian Ruder, Kyunghyun Cho, and Iryna Gurevych. AdapterHub: A Framework for Adapting Transformers. *arXiv*, 2020. arXiv:[2007.07779](https://arxiv.org/abs/2007.07779), doi:[10.48550/arxiv.2007.07779](https://doi.org/10.48550/arxiv.2007.07779).
- [114] Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. MAD-X: An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7654–7673, 2020. doi:[10.18653/v1/2020.emnlp-main.617](https://doi.org/10.18653/v1/2020.emnlp-main.617).
- [115] Jan Pfänder and Sacha Altay. Spotting false news and doubting true news: a systematic review and meta-analysis of news judgements. *Nature Human Behaviour*, 9(4):688–699, 2025. doi:[10.1038/s41562-024-02086-1](https://doi.org/10.1038/s41562-024-02086-1).

-
- [116] Clifton Poth, Hannah Sterz, Indraneil Paul, Sukannya Purkayastha, Leon Engländer, Timo Imhof, Ivan Vulić, Sebastian Ruder, Iryna Gurevych, and Jonas Pfeiffer. Adapters: A Unified Library for Parameter-Efficient and Modular Transfer Learning. *arXiv*, 2023. [arXiv:2311.11077](#).
 - [117] Luboš Pástor, Robert F. Stambaugh, and Lucian A. Taylor. Sustainable investing in equilibrium. *Journal of Financial Economics*, 142(2):550–571, 2021. [doi:10.1016/j.jfineco.2020.12.011](#).
 - [118] Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 Technical Report. *arXiv*, 2024. [arXiv:2412.15115](#), [doi:10.48550/arxiv.2412.15115](#).
 - [119] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving Language Understanding by Generative Pre-Training. *OpenAI*, 2018.
 - [120] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language Models are Unsupervised Multitask Learners. *OpenAI Blog*, 2019.
 - [121] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *arXiv*, 2019. [arXiv:1910.10683](#), [doi:10.48550/arxiv.1910.10683](#).
 - [122] Gabriele Ranco, Darko Aleksovski, Guido Caldarelli, Miha Grčar, and Igor Mozetič. The Effects of Twitter Sentiment on Stock Price Returns. *PLoS ONE*, 10(9):e0138441, 2015. [arXiv:1506.02431](#), [doi:10.1371/journal.pone.0138441](#).
 - [123] Amirhossein Razavi, Mina Soltangheis, Negar Arabzadeh, Sara Salamat, Morteza Zihayat, and Ebrahim Bagheri. Advances in Information Retrieval, 47th European Conference on Information Retrieval, ECIR 2025, Lucca, Italy, April 6–10, 2025, Proceedings, Part III. *Lecture Notes in Computer Science*, pages 303–313, 2025. [doi:10.1007/978-3-031-88714-7_29](#).
 - [124] Varshini Reddy, Rik Koncel-Kedziorski, Viet Dac Lai, Michael Krumbick, Charles Lovering, and Chris Tanner. DocFinQA: A Long-Context Financial Reasoning Dataset. *arXiv*, 2024. [arXiv:2401.06915](#), [doi:10.48550/arxiv.2401.06915](#).
 - [125] Sohini Roychowdhury, Marko Krema, Brian Moore, Xingjian Lai, Dike Effedua, and Bharat Jethwani. FiSTECH: Financial Style Transfer to Enhance Creativity without Hallucinations in LLMs. *arXiv*, 2024. [arXiv:2408.05365](#), [doi:10.48550/arxiv.2408.05365](#).

- [126] Jinlong Ruan, Wei Wu, and Jiebo Luo. Stock Price Prediction Under Anomalous Circumstances. *2020 IEEE International Conference on Big Data (Big Data)*, 00:4787–4794, 2020. [arXiv:2109.15059](#), [doi:10.1109/bigdata50022.2020.9378030](#).
- [127] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv*, 2019. [arXiv:1910.01108](#).
- [128] Sunita Sarawagi. Information Extraction. *Foundations and Trends in Databases*, 1(3):261–377, 2007. [doi:10.1561/19000000003](#).
- [129] Bhaskarjit Sarmah, Tianjie Zhu, Dhagash Mehta, and Stefano Pasquali. Towards reducing hallucination in extracting information from financial reports using Large Language Models. *arXiv*, 2023. [arXiv:2310.10760](#), [doi:10.48550/arxiv.2310.10760](#).
- [130] Jaromir Savelka and Kevin D. Ashley. The unreasonable effectiveness of large language models in zero-shot semantic annotation of legal texts. *Frontiers in Artificial Intelligence*, 6:1279794, 2023. [doi:10.3389/frai.2023.1279794](#).
- [131] Jürgen Schmidhuber. Deep learning in neural networks: An overview. *Neural Networks*, 61:85–117, 2015. [arXiv:1404.7828](#), [doi:10.1016/j.neunet.2014.09.003](#).
- [132] Samiha Sharlin, Fariha Agbere, Kevin Ishimwe, Zuzana Osifová, Ondřej Socha, Martin Dračinský, and Tyler Josephson. NMR-Challenge for LLMs: Evaluating Chemical Reasoning in Humans and AI. 2025. [doi:10.26434/chemrxiv-2025-x8h36-v2](#).
- [133] Soumya Sharma, Subhendu Khatuya, Manjunath Hegde, Afreen Shaikh, Koustuv Dasgupta, Pawan Goyal, and Niloy Ganguly. Financial Numeric Extreme Labelling: A dataset and benchmarking. *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3550–3561, 2023. [doi:10.18653/v1/2023.findings-acl.219](#).
- [134] William F Sharpe. The Sharpe Ratio. *The Journal of Portfolio Management*, 21(1):49–58, 1994. [doi:10.3905/jpm.1994.409501](#).
- [135] Bowen Shi, Jichang Zhao, and Ke Xu. A Word2vec Model for Sentiment Analysis of Weibo. *2019 16th International Conference on Service Systems and Service Management (ICSSSM)*, 00:1–6, 2019. [doi:10.1109/icsssm.2019.8887652](#).
- [136] Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. Retrieval Augmentation Reduces Hallucination in Conversation. *arXiv*, 2021. [arXiv:2104.07567](#), [doi:10.48550/arxiv.2104.07567](#).
- [137] Thendo Sidogi, Rendani Mbuvha, and Tshilidzi Marwala. Stock Price Prediction Using Sentiment Analysis. *2021 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 00:46–51, 2021. [doi:10.1109/smc52423.2021.9659283](#).

-
- [138] Isaac Slaughter, Axel Peytavin, Johan Ugander, and Martin Saveski. Community Notes Moderate Engagement With and Diffusion of False Information Online. *arXiv*, 2025. [arXiv:2502.13322](#), [doi:10.48550/arxiv.2502.13322](#).
 - [139] S&P. S&P 500 RavenPack AI Index Brochure, 2024. URL: <https://www.spglobal.com/spdji/en/documents/brochure/brochure-sp-500-ravenpack-ai-index.pdf>.
 - [140] S&P. S&P 500 RavenPack AI Indices Methodology, 2024. URL: <https://www.spglobal.com/spdji/en/documents/methodologies/methodology-sp-500-rvnpck-ai-indices.pdf>.
 - [141] Mark Stencel, Erica Ryan, and Joel Luther. With half the planet going to the polls in 2024, fact-checking sputters - Poynter, 2024. URL: <https://www.poynter.org/fact-checking/2024/duke-reporters-lab-fact-checking-cencus-2024/>.
 - [142] Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. RoFormer: Enhanced Transformer with Rotary Position Embedding. *arXiv*, 2021. [arXiv:2104.09864](#), [doi:10.48550/arxiv.2104.09864](#).
 - [143] R.S. Sutton and A.G. Barto. Reinforcement Learning: An Introduction. *IEEE Transactions on Neural Networks*, 9(5):1054–1054, 1998. [doi:10.1109/tnn.1998.712192](#).
 - [144] Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. Lexicon-Based Methods for Sentiment Analysis. *Computational Linguistics*, 37(2):267–307, 2011. [doi:10.1162/coli_a_00049](#).
 - [145] Kaiming Tao, Zachary A. Osman, Philip L. Tzou, Soo-Yon Rhee, Vineet Ahluwalia, and Robert W. Shafer. GPT-4 performance on querying scientific publications: reproducibility, accuracy, and impact of an instruction sheet. *BMC Medical Research Methodology*, 24(1):139, 2024. [doi:10.1186/s12874-024-02253-y](#).
 - [146] Gemini Team. Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities. *arXiv*, 2025. URL: <https://arxiv.org/abs/2507.06261>.
 - [147] Llama Team and AI @ Meta. The Llama 3 Herd of Models. *arXiv*, 2024. [arXiv:2407.21783](#), [doi:10.48550/arxiv.2407.21783](#).
 - [148] Qwen Team. Qwen3 Technical Report. *arXiv*, 2025. [arXiv:2505.09388](#), [doi:10.48550/arxiv.2505.09388](#).
 - [149] Paul C. Tetlock, Maytal Saar-Tsechansky, and Sofus Macskassy. More Than Words: Quantifying Language to Measure Firms’ Fundamentals. *The Journal of Finance*, 63(3):1437–1467, 2008. [doi:10.1111/j.1540-6261.2008.01362.x](#).
 - [150] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro,

- Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. LLaMA: Open and Efficient Foundation Language Models. *arXiv*, 2023. [arXiv:2302.13971](#), [doi:10.48550/arxiv.2302.13971](#).
- [151] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open Foundation and Fine-Tuned Chat Models. *arXiv*, 2023. [arXiv:2307.09288](#), [doi:10.48550/arxiv.2307.09288](#).
- [152] Peter D. Turney. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02*, pages 417–424, 2002. [doi:10.3115/1073083.1073153](#).
- [153] Petter Törnberg. ChatGPT-4 Outperforms Experts and Crowd Workers in Annotating Political Twitter Messages with Zero-Shot Learning. *arXiv*, 2023. [arXiv:2304.06588](#), [doi:10.48550/arxiv.2304.06588](#).
- [154] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention Is All You Need. *arXiv*, 2017. [arXiv:1706.03762](#).
- [155] Inna Vogel, Pauline Möhle, Meghana Meghana, and Martin Steinebach. Adapter fusion for check-worthiness detection - combining a task adapter with a NER adapter. *ROMCIR 2024: The 4th Workshop on Reducing Online Misinformation through Credible Information Retrieval*, 2024. URL: <https://ceur-ws.org/Vol-3677/paper4.pdf>.
- [156] Xingchen Wan, Ruoxi Sun, Hootan Nakhost, and Sercan O Arik. Teach Better or Show Smarter? On Instructions and Exemplars in Automatic Prompt Optimization. *arXiv*, 2024. [arXiv:2406.15708](#), [doi:10.48550/arxiv.2406.15708](#).
- [157] Xingchen Wan, Jie Yang, Slavi Marinov, Jan-Peter Calliess, Stefan Zohren, and Xiaowen Dong. Sentiment correlation in financial news networks and associated market movements. *Scientific Reports*, 11(1):3062, 2021. [arXiv:2011.06430](#), [doi:10.1038/s41598-021-82338-6](#).

-
- [158] Chuan-Ju Wang, Ming-Feng Tsai, Tse Liu, and Chin-Ting Chang. Financial Sentiment Analysis for Risk Prediction. *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, 2013. URL: <https://aclanthology.org/I13-1097/>.
 - [159] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-Consistency Improves Chain of Thought Reasoning in Language Models. *arXiv*, 2022. [arXiv:2203.11171](https://arxiv.org/abs/2203.11171), doi: [10.48550/arxiv.2203.11171](https://doi.org/10.48550/arxiv.2203.11171).
 - [160] Yan Wang and Xueke Wang. From ratings to action: The impact of ESG performance on corporate innovation. *Heliyon*, 10(4):e26683, 2024. doi: [10.1016/j.heliyon.2024.e26683](https://doi.org/10.1016/j.heliyon.2024.e26683).
 - [161] Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Nathan Cooper, Griffin Adams, Jeremy Howard, and Iacopo Poli. Smarter, Better, Faster, Longer: A Modern Bidirectional Encoder for Fast, Memory Efficient, and Long Context Finetuning and Inference. *arXiv*, 2024. [arXiv:2412.13663](https://arxiv.org/abs/2412.13663), doi: [10.48550/arxiv.2412.13663](https://doi.org/10.48550/arxiv.2412.13663).
 - [162] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *arXiv*, 2022. [arXiv:2201.11903](https://arxiv.org/abs/2201.11903), doi: [10.48550/arxiv.2201.11903](https://doi.org/10.48550/arxiv.2201.11903).
 - [163] Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. BloombergGPT: A Large Language Model for Finance. *arXiv*, 2023. [arXiv:2303.17564](https://arxiv.org/abs/2303.17564).
 - [164] Panos Xidonas and Eric Essner. On ESG Portfolio Construction: A Multi-Objective Optimization Approach. *Computational Economics*, 63(1):21–45, 2024. doi: [10.1007/s10614-022-10327-6](https://doi.org/10.1007/s10614-022-10327-6).
 - [165] Qianqian Xie, Jennifer Amy Bishop, Prayag Tiwari, and Sophia Ananiadou. Pre-trained language models with domain knowledge for biomedical extractive summarization. *Knowledge-Based Systems*, 252:109460, 2022. doi: [10.1016/j.knosys.2022.109460](https://doi.org/10.1016/j.knosys.2022.109460).
 - [166] Yong Xie, Karan Aggarwal, and Aitzaz Ahmad. Efficient Continual Pre-training for Building Domain Specific Large Language Models. *Findings of the Association for Computational Linguistics ACL 2024*, pages 10184–10201, 2024. doi: [10.18653/v1/2024.findings-acl.606](https://doi.org/10.18653/v1/2024.findings-acl.606).
 - [167] Derong Xu, Wei Chen, Wenjun Peng, Chao Zhang, Tong Xu, Xiangyu Zhao, Xian Wu, Yefeng Zheng, Yang Wang, and Enhong Chen. Large language models for generative information extraction: a survey. *Frontiers of Computer Science*, 18(6):186357, 2024. doi: [10.1007/s11704-024-40555-y](https://doi.org/10.1007/s11704-024-40555-y).

- [168] Lingling Xu, Haoran Xie, Si-Zhao Joe Qin, Xiaohui Tao, and Fu Lee Wang. Parameter-Efficient Fine-Tuning Methods for Pretrained Language Models: A Critical Review and Assessment. *arXiv*, 2023. [arXiv:2312.12148](#), [doi:10.48550/arxiv.2312.12148](#).
- [169] An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. Qwen2 Technical Report. *arXiv*, 2024. [arXiv:2407.10671](#), [doi:10.48550/arxiv.2407.10671](#).
- [170] Hongyang Yang, Xiao-Yang Liu, and Christina Dan Wang. FinGPT: Open-Source Financial Large Language Models. *arXiv*, 2023. [arXiv:2306.06031](#), [doi:10.48550/arxiv.2306.06031](#).
- [171] Hongyang Yang, Boyu Zhang, Neng Wang, Cheng Guo, Xiaoli Zhang, Likun Lin, Junlin Wang, Tianyu Zhou, Mao Guan, Runjia Zhang, and Christina Dan Wang. FinRobot: An Open-Source AI Agent Platform for Financial Applications using Large Language Models. *arXiv*, 2024. [arXiv:2405.14767](#), [doi:10.48550/arxiv.2405.14767](#).
- [172] Yang Yang, Zhilei Wu, Yuexiang Yang, Shuangshuang Lian, Fengjie Guo, and Zhiwei Wang. A Survey of Information Extraction Based on Deep Learning. *Applied Sciences*, 12(19):9691, 2022. [doi:10.3390/app12199691](#).
- [173] Yaoyao Yu, Leilei Gan, Yinghao Hu, Bin Wei, Kun Kuang, and Fei Wu. Evaluating Test-Time Scaling LLMs for Legal Reasoning: OpenAI o1, DeepSeek-R1, and Beyond. *arXiv*, 2025. [arXiv:2503.16040](#), [doi:10.48550/arxiv.2503.16040](#).
- [174] Dan Zhang, Tao Feng, Lilong Xue, Yuandong Wang, Yuxiao Dong, and Jie Tang. Parameter-Efficient Fine-Tuning for Foundation Models. *arXiv*, 2025. [arXiv:2501.13787](#), [doi:10.48550/arxiv.2501.13787](#).
- [175] Xichen Zhang and Ali A. Ghorbani. An overview of online fake news: Characterization, detection, and discussion. *Information Processing & Management*, 57(2):102025, 2020. [doi:10.1016/j.ipm.2019.03.004](#).
- [176] Xin Zhang, Yanzhao Zhang, Dingkun Long, Wen Xie, Ziqi Dai, Jialong Tang, Huan Lin, Baosong Yang, Pengjun Xie, Fei Huang, Meishan Zhang, Wenjie Li, and Min Zhang. mGTE: Generalized Long-Context Text Representation and Reranking Models for Multilingual Text Retrieval. *arXiv*, 2024. [arXiv:2407.19669](#), [doi:10.48550/arxiv.2407.19669](#).

-
- [177] Zikang Zhang, Wangjie Wu, Tianci Wu, Xinrui Wang, Juntao Li, and Min Zhang. A Survey of Generative Information Extraction. pages 4840–4870, 2025. URL: <https://aclanthology.org/2025.coling-main.324/>.
 - [178] Xinyi Zhou and Reza Zafarani. A Survey of Fake News. *ACM Computing Surveys (CSUR)*, 53(5):1–40, 2020. [arXiv:1812.00315](#), [doi:10.1145/3395046](#).
 - [179] Jingming Zhuo, Songyang Zhang, Xinyu Fang, Haodong Duan, Dahua Lin, and Kai Chen. ProSA: Assessing and Understanding the Prompt Sensitivity of LLMs. *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 1950–1976, 2024. [doi:10.18653/v1/2024.findings-emnlp.108](#).
 - [180] Arkaitz Zubiaga, Ahmet Aker, Kalina Bontcheva, Maria Liakata, and Rob Procter. Detection and Resolution of Rumours in Social Media. *ACM Computing Surveys (CSUR)*, 51(2):1–36, 2018. [arXiv:1704.00656](#), [doi:10.1145/3161603](#).