JOIM
www.joim.com

Journal Of Investment Management, Vol. 23, No. 3, (2025), pp. 49–71

OPTIMIZING LARGE LANGUAGE MODELS FOR SUSTAINABLE INVESTORS*

Andrew Chin[a,†], *Che Guan*[b], *Promod Rajaguru*[c,d], *Qifeng Sun*[c,e] *and Yuning Wu*[c,f]

We use large language models (LLMs) and natural language processing (NLP) to extract environmental, social and governance (ESG) insights from real-time news, creating an expert-annotated dataset to evaluate ESG classification, firm relevance, and sentiment. Our fine-tuned models outperform pre-trained ones in ESG detection, firm impact, and sentiment analysis. Furthermore, in-context learning does not improve performance, indicating optimal tuning. Event studies and backtests show that our sentiment signals predict underperforming stocks, with higher model confidence in negative sentiment correlating to worse outcomes. These findings emphasize the value of fine-tuning models with expert-annotated data, and leveraging ESG sentiment signals to generate investment insights from qualitative data to enhance alpha generation.

With the increasing focus on sustainable investing globally, asset managers are striving to gain an edge by uncovering impactful environmental, social, and governance (ESG) insights from a trove of new information sources. Specifically, portfolio managers are scouring a wider variety of data sources to shed light on pertinent ESG issues that may impact investment returns. These investors are also leveraging a broader set of tools to extract investment signals from these data sources.

One important source of real-time insights is the global news cycle, covering a broad spectrum

[a]Chief AI Officer, AllianceBernstein Holding L.P., New York, NY, USA.
E-mail: andrew.chin@alliancebernstein.com
[b]Principal Data Scientist at AllianceBernstein L.P., Nashville, TN, USA.
E-mail: che.guan@alliancebernstein.com
[c]Master's Student in Data Science, Vanderbilt University, Nashville, TN, USA.
[d]E-mail: promod.k.rajaguru@vanderbilt.edu
[e]E-mail: qifeng.sun@vanderbilt.edu
[f]E-mail: yuning.wu@vanderbilt.edu

*This research was a collaborative effort between AllianceBernstein LP and students from Vanderbilt University's Data Science Institute.
†Corresponding author.

of information outlets and potential impactful events. Investment analysts are monitoring and sifting through the news to discern information and events that may impact the financial performance of their investments. Corporate actions, product launches, competitive positioning, and regulatory guidance are examples of events that may impact the financial well-being of a company. In spite of these potential benefits in covering the news, the abundance of information and the unstructured nature of news content present significant hurdles in effectively extracting relevant insights. In our study, we leverage various pre-trained and fine-tuned natural language processing (NLP) models and techniques to classify news articles, conduct sentiment analysis, perform causal inference on the impact of sentiment on company performance, and derive actionable ESG signals for investment analysis. Our findings demonstrate the value of fine-tuning language models for ESG analysis and utilizing the derived ESG sentiment signals for alpha generation, thus greatly improving the capabilities for tech-savvy sustainable investors.

Our study makes several key contributions to the existing literature. First, we develop an expert-annotated dataset with expert labels on ESG relevance, company impact, and sentiment. We then assess the performance of various language models, both small and large, in classifying news articles by their ESG content, revealing that fine-tuning enhances accuracy significantly. Notably, the Llama models demonstrate strong learning capabilities from annotated samples, offering distinct advantages for sustainable investors. We also find that few-shot learning does not improve performance, particularly for the fine-tuned models. Furthermore, we show that language models effectively determine the impact of news on specific companies mentioned in the articles—a critical factor in understanding the causal effects of news events on company performance.

In sentiment classification, our results indicate that language models perform exceptionally well, especially with fine-tuning. For sustainable investors, these insights provide valuable signals for stock selection. Specifically, our event studies and backtests highlight the effectiveness of negative sentiment signals. A notable finding is that extremely negative ESG news is strongly associated with significant underperformance.

We structure the paper as follows. We first provide a literature review and then describe our data sources and data wrangling. We then give an overview of our expert annotations to create a dataset to benchmark various models in the paper. With this dataset acting as the ground truth, we compare different language models, along with fine-tuning and few-shot learning, for different classification tasks. We then conduct event studies to demonstrate the value of our derived sentiment signals. Finally, we show a simple backtest to leverage the proposed investment signals to identify underperformers.

1 Literature Review

The application of language models in analyzing ESG data is an emerging field, yet its potential for providing actionable insights for investors remains underexplored. NLP, a branch of artificial intelligence that enables computers to understand, interpret, and generate human language, has been used to analyze textual data. Prior research in NLP has shown promising applications of machine learning in categorizing and sentiment-analyzing news data, especially within financial markets (Chen *et al.*, 2020). However, few studies have integrated these techniques to specifically examine ESG-relevant news and its impact on corporate performance. This review outlines existing contributions in ESG analysis, sentiment

classification, and the use of language models, emphasizing gaps addressed by our study.

1.1 Sentiment analysis for financial news

Sentiment analysis leveraging NLP techniques has long been employed in finance to extract information from textual data, with numerous studies underscoring its predictive power for asset pricing and stock returns. While initial studies used pre-defined dictionaries to classify sentiment (Loughran and McDonald, 2011), language models like BERT (Bidirectional Encoder Representations from Transformers) have been used in recent studies to classify sentiment in financial news (Hu *et al*., 2018; Dorfleitner and Zhang, 2024). BERT-based models significantly improved the understanding of context and language nuances, and have been adapted for sentiment analysis in the financial domain. However, few studies have focused on fine-tuning these models to capture sentiment specifically within the ESG context, where language around sustainability and impact may differ. Our study addresses this by fine-tuning models to classify ESG-specific sentiment and assessing their impact on stock performance.

1.2 Advances in language models for ESG relevance

Recent advancements in language models have opened new avenues for ESG classification and relevance analysis, with models like BERT and GPT-3 (Generative Pre-trained Transformer) demonstrating enhanced performance in various text classification tasks (Brown *et al*., 2020; Derrick, 2024). While BERT and GPT models are both based on the transformer architecture, decoder-only models like the GPT series have excelled at understanding text as well as in generating text, making them more useful across a range of applications. In addition, research comparing different language models for financial text classification, such as FinBERT (Yang *et al*., 2020; Huang *et al*., 2023), has shown that fine-tuning can significantly improve accuracy in domain-specific contexts. Our study compares multiple language models, including smaller models and the advanced Mistral, Llama, and GPT models, to assess their effectiveness in ESG relevance classification, demonstrating that fine-tuning improves accuracy and that Llama, in particular, offers distinct advantages in learning from expert-annotated data.

1.3 Contribution to sustainable investing

The application of NLP in sustainable investing remains nascent, with studies focusing largely on quantitative ESG scores rather than qualitative insights derived from real-time news data. Previous research by Berg *et al*. (2019) highlights the challenges of using ESG metrics for predicting financial performance, often due to the static nature of most ESG data. By providing a dynamic, news-based approach, our study advances sustainable investing literature, showing that language models can differentiate between high- and low-performing stocks based on ESG sentiment, and that highly negative ESG news signal underperformance. This finding aligns with the limited literature suggesting that negative sentiment has a stronger impact on stock returns than positive sentiment (Tetlock *et al*., 2008), especially within ESG contexts.

Our study contributes to the growing body of research in applying NLP techniques in ESG investing. We address key gaps by using advanced language models to provide real-time, actionable insights for investors, contributing novel methodologies in ESG relevance classification, causal inference for firm impacts, and sentiment analysis tailored to sustainable investing strategies. The results offer new pathways for

integrating qualitative ESG insights into financial decision-making.

2 Data

We use a sample of news articles from the fintech company Perigon. This news provider delivers real-time, AI-enriched global news data, and provides access to a vast and diverse range of news sources, encompassing over 150,000 sources worldwide.

For our study, we extract a total of 120,520 unique news articles from September 13, 2022 to September 8, 2023. This dataset contains all the news for the 508 distinct companies that were constituents of the S&P 500 index over the period. Given that articles can be linked to multiple companies, we expand our dataset by allowing articles to be associated with up to five companies. This results in 213,994 observations in our final research dataset.

Exhibit 1 shows the number of unique news articles per day. There are approximately 330 articles per day for the companies in our dataset but the weekends have significantly less news, as exhibited by the periodicity of the chart. Given that we are focused on the constituents of the S&P 500 index, the daily average implies that we have news for about 60% of the companies on a daily basis.

Exhibit 2 shows the number of companies in different cohorts of article mentions over the 1-year period. Note that many companies do not have frequent news events, with about 40% of the companies having less than 50 articles during this period. On the other hand, there are about 100 companies with wide news coverage, averaging about 1 article per day.

Exhibit 3 shows the number of news articles within the top 20 industries (sorted by number of articles). Note that stocks in the Media, Interactive Media & Services, and Automobiles industries dominate in the news and some industries garner very little attention in the media. News focus and volume is idiosyncratic, driven by macro-economic forces and company-specific events, and we expect the data to look different depending on the study period.

Exhibit 1 : Number of news articles per day.

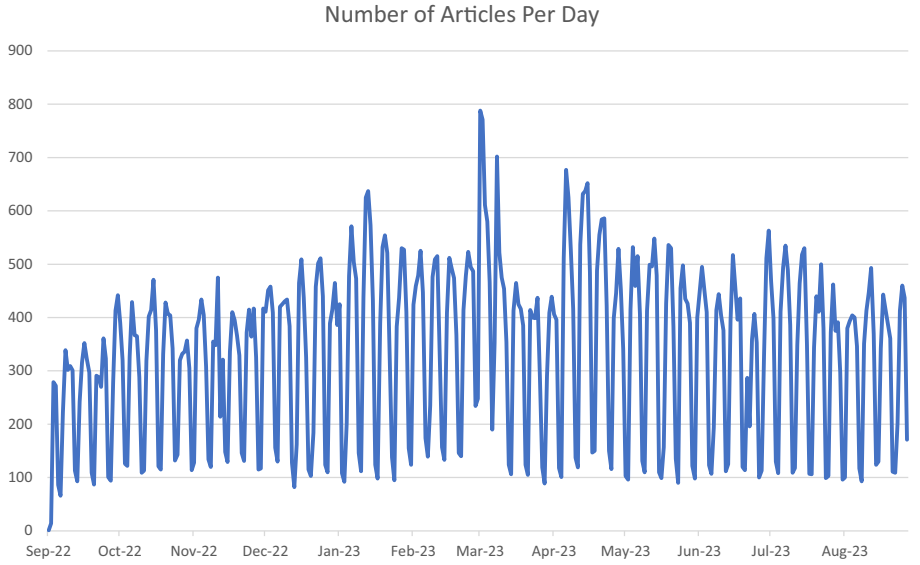


Exhibit 2: Number of companies in each article count cohort.

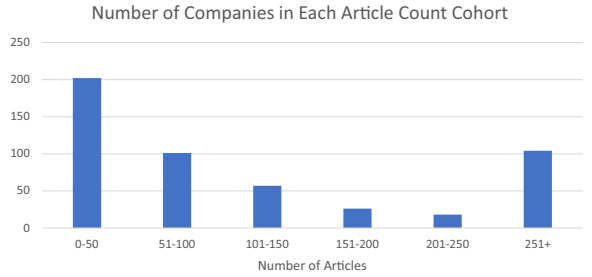
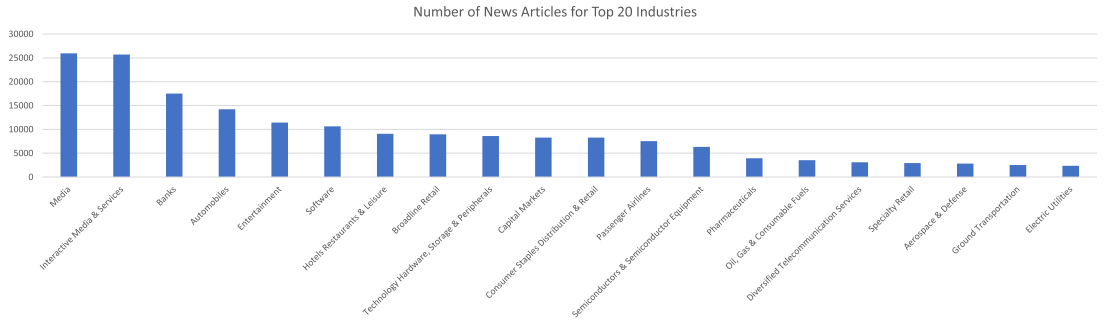


Exhibit 3: Number of news articles for top 20 industries.



To determine whether the content of the articles are focused on ESG topics and thus relevant for sustainable investors, we use the framework from SASB (Sustainability Accounting Standards Board). Specifically, SASB identifies ESG issues that are likely to affect the financial or operating conditions of companies across 79 industries and we use these insights to determine the likelihood that a news article covers ESG-related topics.

Finally, we use Yahoo Finance's historical stock price data to calculate absolute and excess (relative to broad market) returns.

3 Expert Dataset Using Expert Annotations

We create an expert-annotated dataset as the ground truth for our various experiments by randomly sampling approximately 5,250 news articles across the various industries from our research dataset and then having experts annotate each article for ESG relevance, firm relevance and sentiment assessment using the labels in Exhibit 4. While the annotations were painstakingly time-consuming, we wanted to create a variety of labels and datasets to serve as the ground truth to run our experiments.

4 Target Definition and Analysis

We are ultimately interested in evaluating the ability of language models to assess the sentiment of a company resulting from ESG-related news articles. To do this, we create two datasets that are derived from the labels in Exhibit 4.

As shown in Exhibit 5, the dataset "Sentiment for all ESG News with Firm Impact" captures the sentiment of all articles that are ESG-related and there is company impact while Exhibit 6, the dataset "Sentiment for *Core* ESG News with Firm Impact," focuses solely on the articles where the ESG news and the company impact are considered to be major. We hypothesize that investors may react differently to news that have major impacts versus news that are

Exhibit 4: ESG/firm relevance and sentiment labels.

ESG Relevance

Major	The article has a major focus on at least one of the SASB ESG topics
Minor	The article has a minor focus on a SASB ESG topic
No	The article does not focus on any of the SASB ESG topics

Firm Relevance

Major	The article has a major focus on the firm indicated by the vendor
Minor	The article has a minor focus on the firm indicated by the vendor
No	The article does not focus on the firm indicated by the vendor
Irrelevant	The article is not related to an ESG topic (ESG Relevance labeled "No")

Sentiment

Positive	The sentiment on the firm resulting from the ESG discussion in the article is positive
Neutral	The sentiment on the firm resulting from the ESG discussion in the article is neutral
Negative	The sentiment on the firm resulting from the ESG discussion in the article is negative
Irrelevant	The article is not related to an ESG topic (ESG Relevance labeled "No")

Exhibit 5: Sentiment for all ESG News with firm impact.

Positive	ESG Relevance is Major or Minor, Firm Relevance is Major or Minor, Sentiment is Positive
Neutral	ESG Relevance is Major or Minor, Firm Relevance is Major or Minor, Sentiment is Neutral
Negative	ESG Relevance is Major or Minor, Firm Relevance is Major or Minor, Sentiment is Negative
Irrelevant	None of the above three cases

Exhibit 6: Sentiment for core ESG news with firm impact.

Positive	ESG Relevance is Major, Firm Relevance is Major, Sentiment is Positive
Neutral	ESG Relevance is Major, Firm Relevance is Major, Sentiment is Neutral
Negative	ESG Relevance is Major, Firm Relevance is Major, Sentiment is Negative
Irrelevant	None of the above three cases

only tangentially related. These two datasets will allow us to test these ideas and tease out the potential impacts from the highly charged ESG articles.

The sentiment distributions of the top 20 industries in the two expert annotated datasets, as sorted by the number of articles, are shown in Exhibits 7 and 8. We note that these distributions are very similar to Exhibit 3, confirming that the expert dataset is representative of the full dataset.

We summarize the sentiment annotations for the two datasets in Exhibit 9. The chart on the left of the exhibit shows the distribution of the labels in Exhibit 5. Note that about 69% of the articles in the full annotated dataset either do not have a focus on ESG or a specific company.

Exhibit 7: Expert annotated dataset—industry distribution for all ESG News with firm impact.

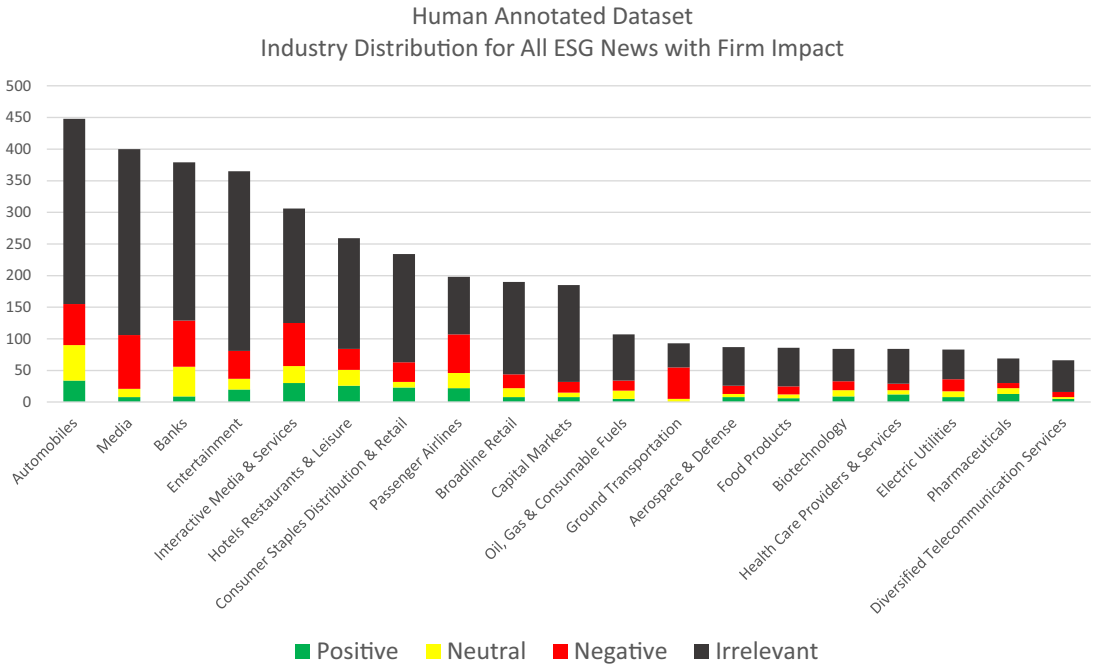


Exhibit 8: Expert annotated dataset—Industry distribution for core ESG news with firm impact.

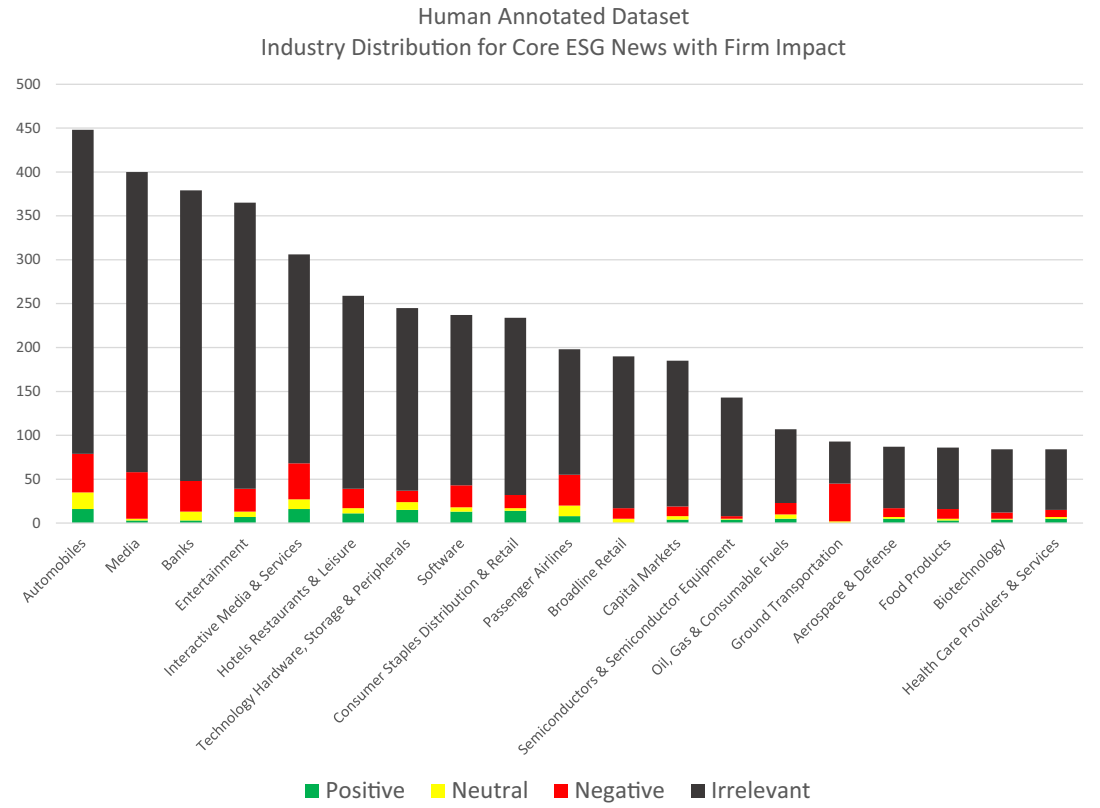
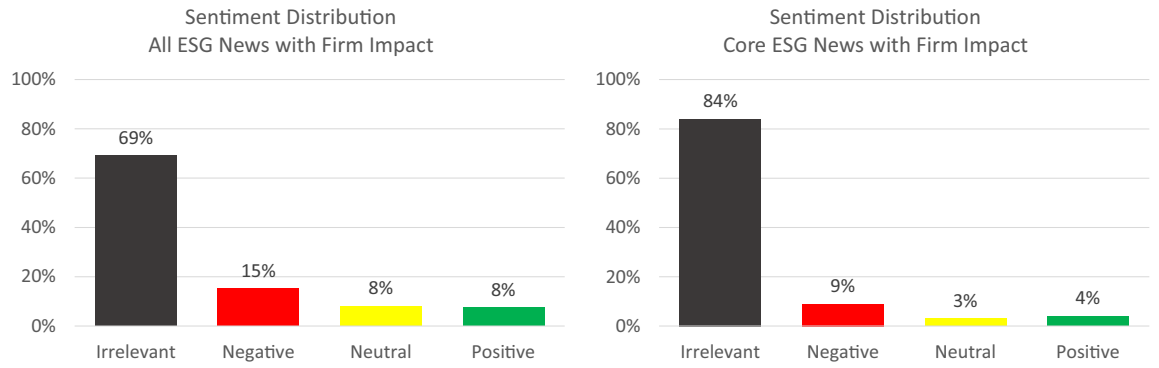


Exhibit 9: Expert annotated dataset; summary sentiment distribution.



Since we are only interested in ESG-related articles and their impacts on specific companies, we classify these articles as "Irrelevant" and do not assess their sentiment. In addition, Negative ESG-related articles appear much more frequently than other types of articles, about twice the rate of Neutral and Positive news articles. This may be due to media bias focusing on bad news to garner attention, especially on ESG issues. Additionally, positive news may be described generically (i.e., company performing well, product has more sales), while negative news may be more specific (i.e., focus on specific issue), and thus, easier to discern.

The chart on the right shows the same summary for the subset of articles with highly ESG-relevant content and a clear focus on specific companies, using the labels from Exhibit 6. In this smaller ESG dataset, only about 16% of the articles have a sentiment label. Note that as a percentage of articles with sentiment labels, Negative articles in this dataset appear even more frequently when compared to the left-hand chart on the larger dataset, implying that ESG and company links are clearer and more direct for negatively charged news.

For robustness, we show the sentiment classification for the Top 20 industries (as sorted by number of articles) for the two annotated datasets in Exhibit A.1 in the appendix. We note here that the sentiment distribution and characteristics are similar to the broader annotated dataset.

We now use these two datasets to conduct various experiments on classification techniques, causal inference, and sentiment analysis. The data is split 80% (∼4,200), 10% (∼525), and 10% (∼525) for training, validation, and testing purposes, respectively.

5 Classification

Given a news article, we want to determine (1) whether the article contains ESG-related content, (2) the relevance of the news on the identified company, and (3) the sentiment of the article. For these classification tasks, we test various open-source and closed-source language models, including FinBERT, GPT, Llama, and Mistral. BERT-based models are open-source, and are designed to understand contextual meaning in text, making them particularly effective for our tasks. LLMs like GPT, Llama, and Mistral, are significantly more complex and leverage transformer-based architectures to interpret text with greater depth and adaptability. Descriptions of the models are shown in Exhibit A.2 in the appendix.

While we start with pre-trained open-source models, we also fine-tune and use in-context learning (few-shot learning) to improve their performance.

Many of the pre-trained models perform well across different domains, surpassing human capabilities in some tasks—see Maslej *et al*., 2024 for representative studies. However, we believe that sustainable investing, an investment approach that aims to generate long-term financial returns while integrating ESG factors, is specialized and further model tuning is required to optimize their performance on our tasks. The pre-trained models are calibrated to perform well across a wide range of tasks but in a niche area with limited data, adding more expertly annotated data may improve performance.

To fine-tune the models, we use the ∼4,200 observations in the annotated training set. Exhibit 10 shows representative prompts for fine-tuning and inference. In the first example, the input of the *fine-tuning* prompt consists of "company" and "text", where "company" is the company ticker and "text" contains the article title and body. The binary target variable "firm_or_not" is either encoded "yes" or "no" depending on the expert annotations in the training set. During the fine-tuning process, a model's parameters are changed from the underlying base model to adjust to our specific tasks using the annotated dataset.

The second example in Exhibit 10 is an *inference* prompt—the model is supplied with inputs containing "company" and "text" and its objective is to generate "yes" or "no" by determining whether the text is related to the company. Our tabular results in the subsequent sections are based on inference prompts using approximately 525 articles in the test dataset.

Note that during the inference stage, we can also include examples in the inference prompt

Exhibit 10 : Representative prompts for fine-tuning and inference.

Prompt Example for Fine-tuning

[INST] Classify whether a given news article is related to the specified company [{data_point["company"]}] using the provided news headline and content enclosed in square brackets. The classifications are as follows:

1. Relevant: The news article has an ESG context related to the company [{data_point["company"]}].
2. Not Relevant: The news article has an ESG context but is not related to the company [{data_point["company"]}].

Provide the corresponding label "yes" for relevant company and "no" for not relevant company. [/INST] [{data_point["text"]}] = {data_point["firm_or_not"]}

Prompt Example for Inference

[INST] Classify whether a given news article is related to the specified company [{data_point["company"]}] using the provided news headline and content enclosed in square brackets. The classifications are as follows:

1. Relevant: The news article has an ESG context related to the company [{data_point["company"]}].
2. Not Relevant: The news article has an ESG context but is not related to the company [{data_point["company"]}].

Provide the corresponding label "yes" for relevant company and "no" for not relevant company. [/INST] [{data_point["text"]}] =

to guide the model. This procedure, called in-context learning, helps the model recognize patterns and respond accordingly, without changing the model's underlying parameters.

6 ESG Classification

6.1 ESG classification for all news

In this section, we compare the models in classifying ESG articles and evaluate the impact of fine-tuning and in-context learning on top of the pre-trained open-source models. Exhibit 11 shows the results of the ESG classification task for all news articles that have "major" or "minor" references to ESG topics.

We use F1-scores for each of the classes as well as the overall accuracy to compare the models. The F1-score balances precision (e.g., how often the model is right when it classifies a particular article) and recall (e.g., how well does the model find the ESG-related articles) to measure a model's efficacy. Higher F1-scores across both the "Yes" and "No" classes suggest better model performance. We also highlight the number of observations for each of the classes under the "Support" column.

We make a few observations. First, all the large language models perform significantly better than FinBERT-ESG, suggesting that the current class of "large" models excel at interpretation and classification tasks, areas where BERT-based models have previously done well. Indeed, across all of our experiments, we find that while BERT-based models are small, easy to train, and customizable, its performance pales in comparison to the robust capabilities of LLMs. Also, OpenAI's GPT4 outperforms all other pre-trained models, matching its impressive performance on various benchmarks in tests across other domains.

However, this advantage disappears when the open-source models are fine-tuned with relevant observations. Specifically, the fine-tuned models outperform the pre-trained models across the board, highlighting the importance of expert annotated datasets. In addition, while the pre-trained Mistral models perform similarly to the pre-trained Llama models, both fine-tuned Llama models perform better than their counterpart Mistral models after fine-tuning. This may suggest that the Llama models "learn" faster during the fine-tuning process.

Finally, the fine-tuned Llama-3-8B-Instruct model has the best overall performance, with the highest F1 scores (0.86 for the "No" class and 0.65 for the "Yes" class) and accuracy (0.80). We highlight this model throughout the paper as a point of comparison across all the experiments.

Next, we explore the impact of in-context learning on top of the pre-trained and fine-tuned LLMs by providing up to four examples in the inference prompt. The objective is to enable the LLMs to enhance performance in classifying ESG articles using a limited set of labeled examples. Findings from Exhibit 12 indicate that few-shot learning does not improve performance. For the pre-trained models, including examples in the prompt has no impact on either the F1-scores or accuracy rates.

Exhibit 11 : ESG classification for all ESG News.

	Mistral-7B-Instruct-v0.2		Mistral-7B-v0.1		Llama-3-8B-Instruct		Llama-3-8B-hf		FinBert-ESG	GPT-35-turbo	GPT4o_mini	GPT4	
F1-Score	Pre-trained	Fine-tuned	Pre-trained	Fine-tuned	Pre-trained	Fine-tuned	Pre-trained	Fine-tuned	Pre-trained	Pre-trained	Pre-trained	Pre-trained	Support
No	0.79	0.82	0.79	0.84	0.80	**0.86**	0.82	0.82	0.60	0.79	0.83	0.83	348
Yes	0.44	0.28	0.02	0.64	0.03	**0.65**	0.00	0.66	0.54	0.30	0.39	0.54	177
Accuracy	0.70	0.71	0.66	0.78	0.67	**0.80**	0.69	0.77	0.57	0.67	0.73	0.74	525

Exhibit 12: ESG classification for all ESG News—Impact of few-shot learning.

Few-Shot learning		Mistral-7B-Instruct-v0.2		Mistral-7B-v0.1		Llama-3-8B-Instruct		Llama-3-8B-hf		
	F1-Score	Pre-trained	Fine-tuned	Pre-trained	Fine-tuned	Pre-trained	Fine-tuned	Pre-trained	Fine-tuned	Support
Zero-shot Learning	No	0.79	0.82	0.79	0.84	0.80	**0.86**	0.82	0.82	348
	Yes	0.44	0.28	0.02	0.64	0.03	**0.65**	0.00	0.66	177
	Accuracy	0.70	0.71	0.66	0.78	0.67	**0.80**	0.69	0.77	525
Two-shot learning	No	0.79	0.73	0.79	0.82	0.80	**0.64**	0.82	0.74	348
	Yes	0.44	0.57	0.02	0.41	0.03	**0.61**	0.00	0.63	177
	Accuracy	0.70	0.67	0.66	0.73	0.67	**0.63**	0.69	0.69	525
Four-shot learning	No	0.79	0.61	0.79	0.81	0.80	**0.58**	0.82	0.78	348
	Yes	0.44	0.56	0.02	0.19	0.03	**0.59**	0.00	0.62	177
	Accuracy	0.70	0.59	0.66	0.69	0.67	**0.58**	0.69	0.72	525

Interestingly, the performance of the fine-tuned models actually degrades with few-shot learning, suggesting that adding examples in the prompting stage negatively impacts the benefits gained through the fine-tuning stage. Using the Llama-3-8B-Instruct model as an example, the pre-trained model has an accuracy rate of 0.67 regardless of the number of examples given in the prompt. The accuracy rate goes up to 0.80 upon fine-tuning but deteriorates significantly to 0.58 once four examples are given during the prompting stage. Taken together, our results suggest that fine-tuning, with a reasonable number of annotated observations, is sufficient for ESG classification.

6.2 Classification for all ESG News with firm impact

We now assess the performance of the models in classifying articles based on the labels in Exhibit 4—we are focused on determining if an article has a "major" or "minor" ESG focus and a "major" or "minor" impact on specific companies. The results in Exhibit 13 show that all the pre-trained models have similar performance, with GPT4 once again outperforming the others. In addition, fine-tuning has no impact on the Mistral models—in fact, the performance of the Mistral models degrades slightly after fine-tuning. On the other hand, fine-tuning the Llama models improves the F1-score of the "Yes" class and the overall accuracy rate, highlighting the importance of fine-tuning for our specialized tasks in classifying firms and ESG issues.

We now compare the results in Exhibit 11 with those in Exhibit 13. The key difference between the two classification tasks is that Exhibit 11 is solely focused on predicting if an article is ESG-related while Exhibit 13 is trying to determine if an article has an ESG focus *and* is relevant to the firm. Across all the models, the results in Exhibit 13 deteriorate, suggesting that all the models struggle with the more complex task of determining ESG as well as firm relevance.

To understand this problem better, we analyze a subset of our ESG-related articles. We hypothesize that the models struggle with determining

Exhibit 13: ESG classification for all ESG News with firm impact.

	Mistral-7B-Instruct-v0.2		Mistral-7B-v0.1		Llama-3-8B-Instruct		Llama-3-8B-hf		FinBert-ESG	GPT-35-turbo	GPT4o_mini	GPT4	
F1-Score	Pre-trained	Fine-tuned	Pre-trained	Fine-tuned	Pre-trained	Fine-tuned	Pre-trained	Fine-tuned	Pre-trained	Pre-trained	Pre-trained	Pre-trained	Support
No	0.80	0.80	0.82	0.82	0.82	**0.83**	0.81	0.82	0.58	0.78	0.84	0.81	363
Yes	0.41	0.00	0.00	0.00	0.01	**0.52**	0.04	0.56	0.52	0.27	0.37	0.53	162
Accuracy	0.70	0.69	0.69	0.69	0.69	**0.75**	0.68	0.74	0.55	0.66	0.74	0.73	525

whether ESG references in articles are connected with the companies mentioned in those same articles. Finding those connections requires nuanced thinking and more context, areas where current language models may still struggle with. To test this, we take all the ESG-related articles in our annotated dataset and remove those articles that our experts have deemed to have no firm relevance—these are the articles that discuss ESG-related topics but they are not related to the firms mentioned in the articles. The new test dataset contains 423 annotated articles, down from the 525 articles in the prior experiments.

After removing the articles without firm relevance, we collect our performance metrics in Exhibit 14. Comparing this table to Exhibit 13, the performance of the pre-trained models is similar, with the Mistral and Llama models performing slightly worse and the other models performing in-line on the smaller dataset. On the other hand, the fine-tuned Mistral and Llama models perform significantly better on the smaller dataset in Exhibit 14. In addition, these fine-tuned results have also improved from Exhibit 11, reinforcing the importance of classifying firm relevance accurately.

These findings confirm our hypothesis that the models struggle with the noise that's introduced from including irrelevant firms in ESG articles. In practice, sustainable investors should take steps to mitigate this issue to optimize model performance. One possibility is to separate the two classification tasks in the process—that is, prompt an LLM twice, once for firm relevance and another time for ESG relevance. However, we find that the success of this procedure is dependent on the quality of the sub-models—they need to be optimized and calibrated to excel on the individual tasks. In other research we have done, we have found that LLMs perform better with simple and direct prompts, rather than nested and more complicated prompts.

Similar to our experiments on classifying ESG articles, we also explore the impact of in-context

Exhibit 14: ESG classification for all ESG News with firm impact (removed articles where mentioned firms were deemed to be irrelevant).

	Mistral-7B-Instruct-v0.2		Mistral-7B-v0.1		Llama-3-8B-Instruct		Llama-3-8B-hf		FinBert-ESG	GPT-35-turbo	GPT4o_mini	GPT4	
F1-Score	Pre-trained	Fine-tuned	Pre-trained	Fine-tuned	Pre-trained	Fine-tuned	Pre-trained	Fine-tuned	Pre-trained	Pre-trained	Pre-trained	Pre-trained	Support
No	0.81	0.90	0.76	0.87	0.76	**0.87**	0.76	0.87	0.57	0.78	0.81	0.80	261
Yes	0.58	0.81	0.00	0.77	0.00	**0.77**	0.00	0.86	0.60	0.39	0.42	0.55	162
Accuracy	0.74	0.87	0.62	0.84	0.62	**0.83**	0.61	0.87	0.58	0.68	0.72	0.73	423

Exhibit 15: ESG classification for all ESG News with firm impact—Impact of few-shot learning.

Few-Shot learning	Metric	Mistral-7B-Instruct-v0.2		Mistral-7B-v0.1		Llama-3-8B-Instruct		Llama-3-8B-hf		
	F1-Score	Pre-trained	Fine-tuned	Pre-trained	Fine-tuned	Pre-trained	Fine-tuned	Pre-trained	Fine-tuned	Support
Zero-shot learning	No	0.80	0.80	0.82	0.82	0.82	**0.83**	0.81	0.82	363
	Yes	0.41	0.00	0.00	0.00	0.01	**0.52**	0.04	0.56	162
	Accuracy	0.70	0.69	0.69	0.69	0.69	**0.75**	0.68	0.74	525
Two-shot learning	No	0.80	0.77	0.82	0.82	0.82	**0.69**	0.81	0.82	363
	Yes	0.41	0.19	0.00	0.00	0.01	**0.55**	0.04	0.52	162
	Accuracy	0.70	0.65	0.69	0.69	0.69	**0.63**	0.68	0.74	525
Four-shot learning	No	0.80	0.72	0.82	0.82	0.82	**0.61**	0.81	0.84	363
	Yes	0.41	0.42	0.00	0.00	0.01	**0.55**	0.04	0.42	162
	Accuracy	0.70	0.62	0.69	0.69	0.69	**0.58**	0.68	0.75	525

learning on top of the pre-trained and fine-tuned models for the combined task of ESG and firm relevance. Our findings in Exhibit 15 indicate performance generally degrades with few-shot learning, consistent with results on general ESG classification. As before, fine-tuning is sufficient in creating effective LLMs for sustainable investing.

7 Sentiment Classification

7.1 Sentiment classification on public financial news dataset

Sentiment analysis is a common task for investors because it provides insights into the emotional and psychological factors that influence financial markets. Language models have been shown to be effective in capturing sentiment of textual data (Dorfleitner and Zhang, 2024).

Before we assess the performance of our suite of models on articles in our ESG news dataset, we test their performance on a public dataset. We aim to establish the overall efficacy of our models on more generic text, not ESG focused. Specifically, we test our models on the Kaggle dataset containing financial news headlines. This dataset has 4,838 observations with annotated sentiment labels. After training various models on this dataset using the same 80% training/10% validation/10% testing split, the results in Exhibit 16 suggest that some of the off-the-shelf pre-trained models are reasonable but our fine-tuned models seem more promising for sentiment analysis. After fine-tuning, accuracy rates range between 85% and 90% and the F1-scores of the three classes improve significantly.

7.2 Sentiment classification for all ESG news with firm impact

While the Kaggle results are encouraging, we want to focus on ESG-related news, not all types of news. In our annotated dataset, only about 31% of the articles (see Exhibit 9) are ESG-related so we want to assess the performance of the models in scoring sentiment on this subset. Exhibit 17 summarizes the results. Note that the fine-tuned Llama models perform significantly better than their pre-trained counterparts, as evidenced by the F1-scores and accuracy rates. Fine-tuned Llama-3-8B-Instruct, in particular, has an accuracy rate of 0.77, significantly outperforming the performance of the other models.

7.3 Sentiment classification for core ESG news with firm impact

We now focus on the more narrow dataset where firm relevance and ESG relevance have both been tagged as "Major." Recall that the prior section had focused on all articles where firm relevance and ESG relevance were either "Major" or "Minor." Our more focused results, as indicated in Exhibit 18, are promising. Similar to the results in Exhibit 17, the fine-tuned models outperform their pre-trained counterparts. However, when compared to Exhibit 17, the fine-tuned models in Exhibit 18 have significantly higher accuracy rates but more erratic F1-scores on the various classes. We think the uneven F1-scores are a result of the small support samples for relevant ESG articles. As a result, we test the robustness of our findings by allocating more observations

Exhibit 16: Sentiment classification for Kaggle financial news headlines.

	Mistral-7B-Instruct-v0.2		Llama-3-8B-Instruct		FinBert-Sentiment	GPT-35-turbo	GPT4o_mini	GPT4	
F1-Score	Pre-trained	Fine-tuned	Pre-trained	Fine-tuned	Pre-trained	Pre-trained	Pre-trained	Pre-trained	Support
Negative	0.80	0.90	0.00	**0.92**	0.86	0.15	0.83	0.87	61
Neutral	0.78	0.88	0.73	**0.91**	0.88	0.42	0.86	0.84	288
Positive	0.62	0.79	0.03	**0.85**	0.83	0.43	0.69	0.63	136
Accuracy	0.74	0.86	0.58	**0.90**	0.86	0.41	0.81	0.80	485

Exhibit 17 : Sentiment classification for all ESG News with firm impact.

	Mistral-7B-Instruct-v0.2		Mistral-7B-v0.1		Llama-3-8B-Instruct		Llama-3-8B-hf		GPT-35-turbo	GPT4o_mini	GPT4	
F1-Score	Pre-trained	Fine-tuned	Pre-trained	Fine-tuned	Pre-trained	Fine-tuned	Pre-trained	Fine-tuned	Pre-trained	Pre-trained	Pre-trained	Support
Negative	0.47	0.47	0.00	0.43	0.00	**0.59**	0.00	0.56	0.29	0.47	0.53	79
Neutral	0.14	0.22	0.00	0.17	0.00	**0.44**	0.00	0.19	0.12	0.16	0.29	43
Positive	0.36	0.27	0.08	0.36	0.17	**0.54**	0.17	0.39	0.25	0.31	0.33	40
Irrelevant	0.76	0.48	0.81	0.82	0.81	**0.86**	0.75	0.83	0.61	0.61	0.71	363
Accuracy	0.64	0.41	0.68	0.70	0.68	**0.77**	0.58	0.71	0.46	0.50	0.60	525

Exhibit 18 : Sentiment classification for core ESG news with firm impact.

	Mistral-7B-Instruct-v0.2		Mistral-7B-v0.1		Llama-3-8B-Instruct		Llama-3-8B-hf		GPT-35-turbo	GPT4o_mini	GPT4	
F1-Score	Pre-trained	Fine-tuned	Pre-trained	Fine-tuned	Pre-trained	Fine-tuned	Pre-trained	Fine-tuned	Pre-trained	Pre-trained	Pre-trained	Support
Negative	0.39	0.48	0.00	0.57	0.00	**0.52**	0.00	0.45	0.38	0.51	0.54	50
Neutral	0.05	0.33	0.00	0.10	0.00	**0.00**	0.00	0.00	0.15	0.20	0.27	13
Positive	0.17	0.41	0.00	0.43	0.07	**0.50**	0.06	0.35	0.31	0.37	0.27	20
Irrelevant	0.70	0.92	0.90	0.93	0.91	**0.92**	0.50	0.93	0.47	0.66	0.61	442
Accuracy	0.55	0.86	0.82	0.87	0.83	**0.86**	0.33	0.86	0.39	0.55	0.52	525

to the test dataset in the training, validation and testing split. We share an example of our robustness tests in Exhibit A.3 in the Appendix using a 70% training/15% validation/15% testing split and find that the results are similar.

Revisiting our three key tasks—determining ESG relevance, firm impact, and sentiment—we find that LLM performance is optimized by combining all three tasks into a single inference prompt, rather than addressing them separately. In general, this result is dependent on the dataset, the model, the complexity of the tasks, and the quality of fine-tuning so we advise readers to experiment on their specific tasks.

8 Event Study

Our results so far have shown that language models can be effective in filtering ESG content, firm relevance, and sentiment from news articles. We will now establish the importance of these insights for sustainable investors.

To quantify the impact, we examine the subsequent one-day stock returns conditioned on different scenarios—the presence of ESG-related content and different cohorts of sentiment (positive, neutral, negative, and extreme negative) in the new articles. We focus on excess returns—returns above the market benchmark, the S&P 500 index return—because excess returns exclude broader market influences that may not be directly tied to company-specific issues.

For each stock, we aggregate all the relevant news articles per day and take the majority sentiment signal. In other words, if a stock has more negatively charged articles versus positive sentiment articles on a given day, we label that day's overall sentiment as negative.

For illustrative purposes, we also construct a new cohort of stocks which have "Extreme Negative" sentiment. Our motivation stems from the fact that while we leverage the language models to classify sentiment as either positive or negative, we have not assessed the polarity of that sentiment.

Exhibit 19: Summary statistics for different sentiment cohorts.

	Sentiment for All ESG News with Firm Impact					Sentiment for Core ESG News with Firm Impact				
	All ESG	Positive	Neutral	Negative	Extreme Negative	Core ESG	Positive	Neutral	Negative	Extreme Negative
Average 1-Day Excess Return (%)	−0.03	−0.03	−0.03	−0.03	−0.08	−0.06	−0.02	−0.16	−0.06	−0.09
Median 1-Day Excess Return (%)	−0.01	−0.01	0.04	−0.02	−0.07	−0.08	−0.07	−0.18	−0.07	−0.06
Total Number of Observations	10,807	2,053	1,212	7,542	3,943	5,282	995	303	3,984	2,301
P-value for One-sample *t*-test with Mean Greater than Zero	0.90	0.76	0.68	0.85	0.98	0.97	0.61	0.91	0.95	0.97
P-value for One-sample *t*-test with Mean Less than Zero	0.10	0.24	0.32	0.15	**0.02**	**0.03**	0.39	0.09	**0.05**	**0.03**

In interpreting news content, investors care about the degree of sentiment to determine its impact on a company's prospects.

To do this, we use the probabilities that accompany the sentiment labels from FinBERT-Sentiment as a proxy for polarity—larger probabilities would suggest higher confidence in the classification results. Specifically, we define $ChargedSentiment$ as

$$\begin{aligned}&\text{ChargedSentiment}\\&\quad = (P(\text{positive}) - P(\text{negative}))/\\&\qquad (1 + P(\text{neutral})),\end{aligned}$$

where $P(\text{positive})$, $P(\text{negative})$, and $P(\text{neutral})$ are the probabilities of the text having positive, negative, and neutral sentiment, respectively.

With this as backdrop, we aim to assess whether articles with more extreme sentiment exert a greater influence on stock performance. Given the small number of articles in the Positive cohort (see next section), we focus our attention on the extreme readings of negative sentiment. For this exercise, we used a ChargedSentiment threshold of −0.50, resulting in a cutoff that filters out the bottom ∼50% of the observations. For robustness, we tested alternative threshold values and obtained consistent results. However, we note that very extreme thresholds produce unstable results because of the limited number of stocks in the cohorts.

Looking at the results in Exhibit 19 for all ESG News on the left-hand side, we note that on average, companies with any ESG-related news underperformed the market by −0.03% per day over our 1-year study period. While this is economically meaningful (−0.03% per day equates to approximately −7.5% per year), the *p*-value for a one-sample *t*-test suggests the results are not statistically significant. This makes sense fundamentally because we do not have any priors suggesting that ESG-related news is correlated with future over- or under-performance.

Note that these statistics for all ESG-related news are heavily influenced by the negative sentiment articles since these account for about 70% of the observations in the dataset. Not surprisingly, the statistics for this cohort are very similar to the full dataset.

The average daily return of the Extreme Negative cohort is −0.08% (equating to average annual returns less than −20%), significantly underperforming the market and the broader Negative sentiment cohort. Unlike the other groups, the

return of the Extreme Negative cohort is statistically significant at the 5% level. These results suggest negative sentiment polarity is an important indicator of future underperformance and sustainable investors can choose their preferred threshold. Practically, we find that the trade-off is between more false positives (less negative threshold) and higher potential impact (more negative threshold).

The results on the right-hand side, focusing on the highly relevant ESG articles with strong firm impacts, are even more interesting than the broader study. Negative sentiment articles now account for about 75% of the total observations, slightly higher than the broader dataset on the left-hand side. Within this more focused dataset, the average 1-day return of the Negative cohort is -0.06%, about $2\times$ lower than the stocks in the same cohort of the broader universe. In addition, this result is statistically significant at the 5% level, suggesting that "major" ESG news and firm relevance are correlated with future stock price underperformance. Finally, focusing on the more negatively charged sentiment articles results in even more underperformance (-0.09% per day).

Overall, our findings suggest that stocks with negative ESG news significantly underperform the broad market during the next trading day and these results can be further enhanced by focusing on more charged readings of sentiment. As a result, sustainable investors can leverage these insights to identify potential vulnerabilities from real-time streaming news to optimize their portfolio construction and trading activities.

9 Sentiment Signal Backtests

9.1 Backtesting results for all ESG news with firm impact

We now perform a temporal study to test the efficacy of the sentiment signals over the 1-year period. Each day, we construct five portfolios—one containing all the stocks with ESG news and four containing the stocks with positive, neutral, negative, and extreme negative sentiment. We then monitor these portfolios over the following day, repeat the classification process, and track the performance of the five portfolios over our full research period.

This backtest is different from the event study in the previous section—in this study, we analyze the efficacy of the sentiment signals over time whereas the prior results aggregate all the stocks and articles into a panel dataset. Both studies are important because they test the robustness of the conclusions in different ways.

Exhibit 20 shows the number of stocks within each sentiment cohort over time for all ESG News with firm impact. We note that the Positive and Neutral cohorts do not contain a large number of stocks on a daily basis and the resulting statistics are unlikely to be robust. The other three cohorts contain a reasonable number of stocks on a daily basis.

Exhibit 21 shows the growth of $1 for each sentiment cohort over our study period. The performance patterns for the all ESG, Positive, and Neutral cohorts are not intuitive but the Negative and Extreme Negative cohorts perform as expected. The Negative cohort underperforms the broad market by about 15% over this period while the Extreme Negative stocks underperform by about 25%, consistent with priors.

The results in this exhibit are consistent with the findings in Exhibit 19 with the possible exception of the Neutral cohort. In Exhibit 21, the Neutral cohort outperforms the broad market by about 4% over the 1-year period but in Exhibit 19, the average daily excess return was negative. However, note that the median 1-day return for this cohort is positive, implying that more than 50% of the

Exhibit 20 : Number of stocks with sentiment signals for all ESG stocks with firm sentiment.

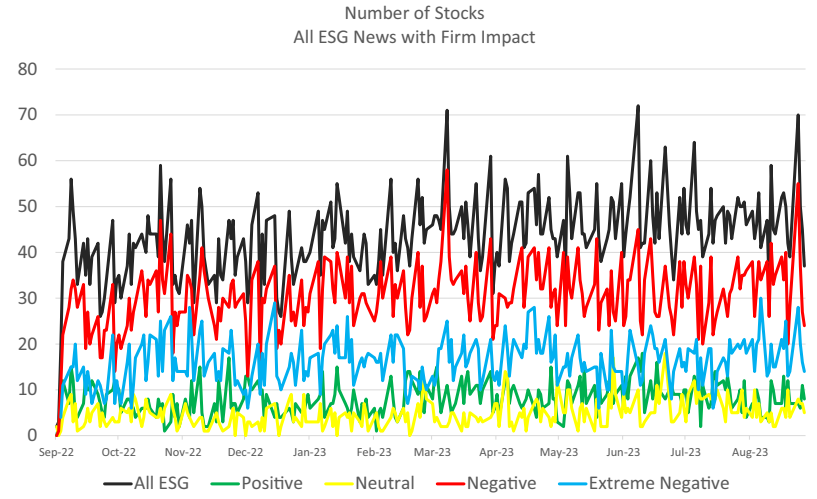
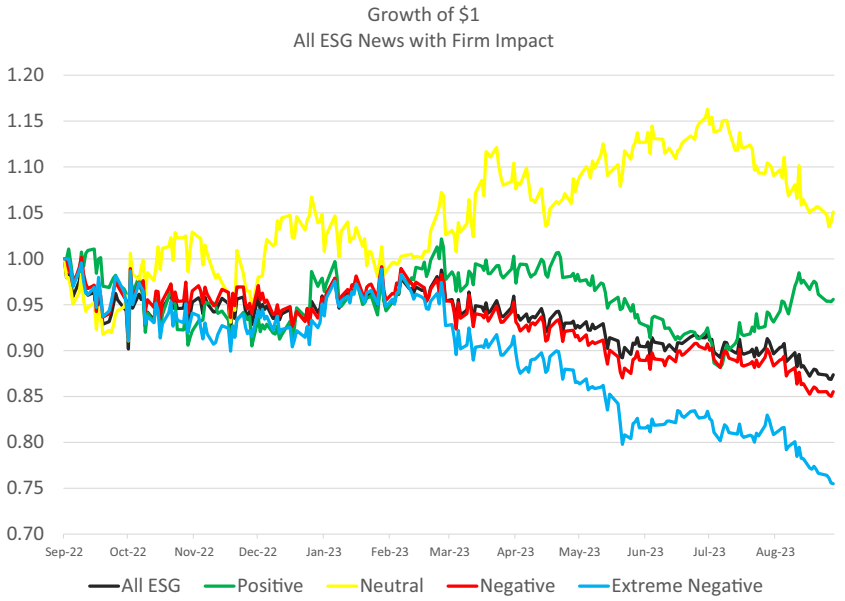


Exhibit 21 : Cumulative excess return versus S&P 500 for all ESG News with firm impact.



excess returns for this cohort are positive. This realization about the skewness of the stocks in the Neutral cohort makes the results in Exhibit 21 more consistent.

9.2 Backtesting results for core ESG News with firm impact

We now look at the smaller dataset containing only the Core ESG Stocks with Firm Sentiment. Exhibit 22 shows the number of stocks for each sentiment cohort in this dataset and Exhibit 23 shows the corresponding performance over time. Consistent with the results in Exhibit 21, stocks with negative ESG news tend to underperform, with the Extreme Negative cohort performing the worse.

We summarize the backtesting results in Exhibit 24. Note that many of the cohorts contain

Exhibit 22: Number of stocks with sentiment signals for core ESG stocks with firm sentiment.

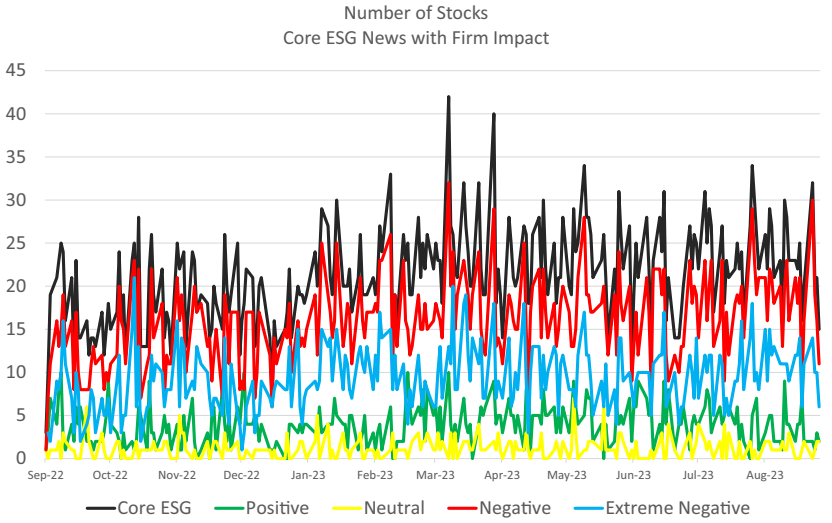
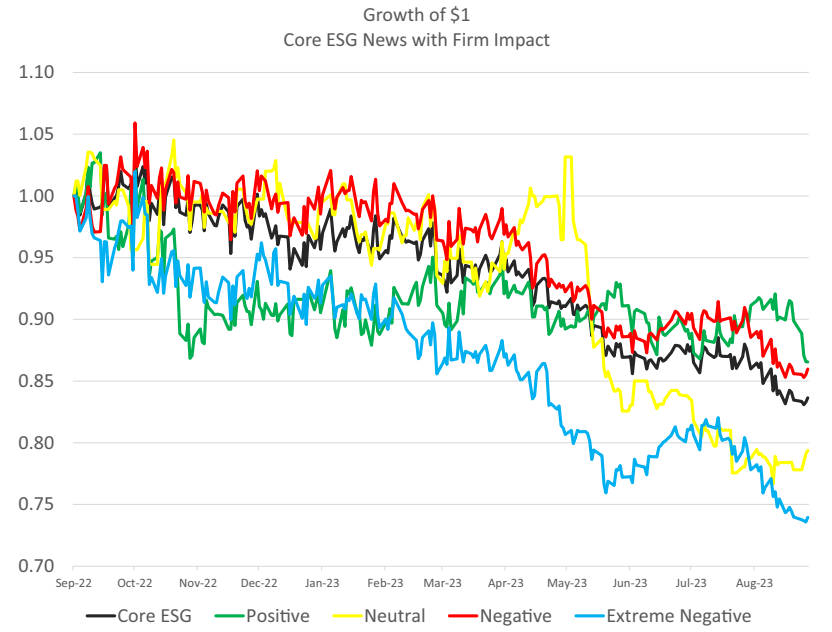


Exhibit 23: Cumulative excess return versus S&P 500 for core ESG news with firm impact.



a small number of stocks so we should view the results with care. The Negative sentiment cohorts of both tables contain a reasonable number of stocks and the backtests suggest under-weighing those stocks can be profitable, with information ratios around 0.60. In addition, we note that while the number of stocks in the Extreme Negative cohorts are small, investors can choose the appropriate filtering threshold for this cohort. For example, systematic investors may choose to include more stocks in this cohort to minimize idiosyncratic risks, while fundamental investors may opt to have a more narrow set of stocks to delve into on a daily basis.

Exhibit 24 : Summary statistics for sentiment signal backtests.

	Sentiment for All ESG News with Firm Impact					Sentiment for Core ESG News with Firm Impact				
	All ESG	Positive	Neutral	Negative	Extreme Negative	Core ESG	Positive	Neutral	Negative	Extreme Negative
Average 1-Day Excess Return (%)	−0.04	−0.01	0.04	−0.05	−0.10	−0.06	−0.05	−0.12	−0.05	−0.11
Median 1-Day Excess Return (%)	0.01	0.04	0.17	0.00	−0.09	−0.03	−0.03	−0.14	−.07	−0.06
Daily Volatility (%)	1.39	1.47	1.81	1.42	1.50	1.44	1.68	1.79	1.48	1.57
Average Annual Return (%)	−12.63	−4.41	5.11	−14.48	−24.50	−16.35	−13.45	−20.61	−14.03	−26.95
Information Ratio	0.58	0.19	0.18	0.65	1.04	0.72	0.52	0.89	0.60	1.06
Daily Average Number of Stocks	43	8	5	30	8	21	4	2	16	4

10 Limitations and Future Directions

Given the sheer volume of news articles, our study only analyzed data over a 1-year period. Extending this research to a longer period over different market cycles would strengthen the findings. Additionally, we only focused on S&P 500 stocks—we expect that the results may be even more significant for stocks outside of this large cap universe given that the smaller cap cohort may be less efficient with stronger reactions to new information.

Finally, with LLMs improving at a rapid pace, newer models may overcome some of the fine-tuning benefits we found in our experiments. In other work, we have found that the more recent models are highly flexible with the potential to excel across many domains.

11 Conclusion

Our research provides a roadmap for sustainable investors to leverage NLP and advanced language models to extract actionable insights from real-time news articles. NLP encompasses various techniques for analyzing text, ranging from dictionary-based approaches, which classify words based on pre-defined lexicons, to context-aware deep learning models. Among these, BERT-based models, such as FinBERT, specifically fine-tuned for financial language, excel at capturing nuanced meanings in structured financial texts. Expanding beyond BERT, transformer-based models, particularly LLMs like GPT, Llama, and Mistral, comprehend contextual sentiment and nuanced language with greater depth. We find that investors can apply these transformer-based models to identify ESG-relevant articles and assess sentiment on those articles to create impactful investment signals.

We compare various language models and find that while pre-trained models can serve as a reasonable starting point for sustainable investors, fine-tuning significantly improves performance in ESG classification and sentiment analysis. Interestingly, while expert annotations from investment analysts can improve model performance through fine-tuning, the effects of few-shot or in-context learning are often limited and can even prove counterproductive. Moreover, language models are effective at evaluating the impact of news on specific companies, providing valuable insights into the causal relationship between news events and company performance. These insights

enable asset managers to strategically optimize the use of expert annotations, enhancing their competitive edge.

The event studies and the simple backtests validate our hypothesis that negative ESG news tend to predict stock underperformance. Our studies show promising results with average annual excess returns of over 10% versus the broad market. Additionally, including a model's confidence of negatively charged sentiment articles can further enhance performance.

Appendix

Exhibit A.1 : Expert annotated dataset; sentiment distribution for top 20 industries.

Industry	Sentiment of All ESG News with Firm Impact (%)				Sentiment of Core ESG News with Firm Impact (%)			
	Negative	Neutral	Positive	Irrelevant	Negative	Neutral	Positive	Irrelevant
Automobiles	0.65	1.07	1.24	5.58	0.30	0.36	0.84	7.03
Media	0.15	0.25	1.62	5.60	0.06	0.04	1.01	6.52
Banks	0.17	0.90	1.39	4.76	0.06	0.19	0.67	6.31
Entertainment	0.38	0.32	0.84	5.41	0.13	0.11	0.50	6.21
Interactive Media & Services	0.57	0.51	1.30	3.45	0.30	0.21	0.78	4.54
Hotels Restaurants & Leisure	0.50	0.48	0.63	3.33	0.21	0.11	0.42	4.19
Technology Hardware, Storage & Peripherals	0.36	0.40	0.44	3.47	0.29	0.17	0.25	3.96
Software	0.55	0.51	0.72	2.72	0.25	0.10	0.48	3.70
Consumer Staples Distribution & Retail	0.44	0.17	0.59	3.26	0.27	0.06	0.29	3.85
Passenger Airlines	0.42	0.46	1.16	1.73	0.15	0.23	0.67	2.72
Broadline Retail	0.15	0.27	0.42	2.78	0.00	0.10	0.23	3.30
Capital Markets	0.15	0.13	0.32	2.92	0.08	0.08	0.21	3.16
Semiconductors & Semiconductor Equipment	0.17	0.15	0.19	2.21	0.08	0.02	0.06	2.57
Oil, Gas & Consumable Fuels	0.10	0.25	0.30	1.39	0.10	0.10	0.25	1.60
Ground Transportation	0.00	0.10	0.95	0.72	0.00	0.04	0.82	0.91
Aerospace & Defense	0.15	0.10	0.25	1.16	0.10	0.04	0.19	1.33
Food Products	0.11	0.11	0.25	1.16	0.06	0.04	0.21	1.33
Biotechnology	0.17	0.19	0.27	0.97	0.08	0.02	0.13	1.37
Health Care Providers & Services	0.23	0.13	0.19	1.05	0.10	0.04	0.15	1.31
Specialty Retail	0.23	0.10	0.10	1.18	0.13	0.02	0.02	1.43

Exhibit A.2: Overview of language models.

Models	Model description	Usage
FinBERT-Sentiment	FinBERT-Sentiment (Yang *et al.*, 2020) is a pre-trained NLP model used to analyze sentiment in financial text. It is built by further training the BERT language model using a large financial corpus and subsequently fine-tuning it for financial sentiment classification.	Pre-trained (baseline)
FinBERT-ESG	FinBERT-ESG (Huang *et al.*, 2023) is a FinBERT model fine-tuned on 2,000 manually annotated sentences from firms' ESG reports and annual reports.	Pre-trained (baseline)
GPT-35-turbo	GPT-3.5 models can understand and generate natural language or code. GPT-3.5-turbo is a cost-effective model in the GPT-3.5 family that has been optimized for chat and traditional completion tasks.	Pre-trained
GPT4o_mini	GPT-4o mini ("o" for "omni") is the most advanced model in the small model category. This multimodal model has higher intelligence than GPT-3.5-turbo but is just as fast.	Pre-trained
GPT4	GPT-4 is more creative and collaborative than previous GPT models. It can generate, edit, and iterate with users on creative and technical writing tasks.	Pre-trained
Mistral-7B-v0.1	The Mistral-7B-v0.1 LLM is a pre-trained generative text model with 7 billion parameters. Its creators have claimed it outperforms the Llama 2, 13B model on all benchmarks.	Pre-trained and fine-tuned
Mistral-7B-Instruct-v0.2	The Mistral-7B-Instruct-v0.2 LLM is an instruct fine-tuned version of Mistral-7B-v0.2.	Pre-trained and fine-tuned
Llama-3-8B-hf	The Meta Llama 3.1 collection of multilingual LLMs contains various pre-trained and instruction-tuned generative models.	Pre-trained and fine-tuned
Llama-3-8B-Instruct	The Llama 3.1 instruction tuned text only models (8B, 70B, and 405B) are optimized for multilingual dialogue use cases and outperform many of the available open source and closed chat models on common industry benchmarks.	Pre-trained and fine-tuned

Exhibit A.3 : Sentiment classification for core ESG news with firm impact (dataset split into 70% training, 15% validation and 15% testing).

	Mistral-7B-Instruct-v0.2		Mistral-7B-v0.1		Llama-3-8B-Instruct		Llama-3-8B-hf		GPT-35-turbo	GPT4o_mini	GPT4	
F1-Score	Pre-trained	Fine-tuned	Pre-trained	Fine-tuned	Pre-trained	Fine-tuned	Pre-trained	Fine-tuned	Pre-trained	Pre-trained	Pre-trained	Support
Negative	0.43	0.42	0.00	0.45	0.00	**0.46**	0.00	0.40	0.10	0.37	0.23	75
Neutral	0.10	0.10	0.00	0.00	0.00	**0.17**	0.00	0.00	0.05	0.06	0.00	20
Positive	0.29	0.28	0.09	0.13	0.11	**0.50**	0.05	0.32	0.07	0.23	0.05	29
Irrelevant	0.82	0.92	0.91	0.92	0.88	**0.92**	0.84	0.92	0.84	0.63	0.88	664
Accuracy	0.72	0.86	0.83	0.85	0.79	**0.86**	0.69	0.85	0.73	0.50	0.79	788

Our comprehensive study provides substantial evidence supporting the influence of ESG sentiment on the financial performance of public companies. By using advanced language models to extract real-time actionable insights from news articles, sustainable investors have new pathways to integrate qualitative ESG insights into investment decision-making.

Acknowledgments

We thank Evan Follis for preparing the raw news article data, integrating the SASB topic descriptions to facilitate expert annotation, and providing valuable insights during the initial phases of the research. Additionally, we gratefully acknowledge the significant contributions of Tiffany Lee, Inchul Yang, and Ruiqi Zhou at Vanderbilt University for their exploratory work in information retrieval and ESG classification in the early stages of the work. Finally, we appreciate the support and guidance provided by Professors Dana Zhang, Jesse Blocher, and Jesse Spencer-Smith to students throughout our collaboration with the Vanderbilt Data Science Institute.

References

Berg, F., Kölbel, J. F., and Rigobon, R. (2019). "Aggregate Confusion: The Divergence of ESG Ratings," https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3438533.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., *et al.* (2020). "Language Models are Few-Shot Learners," https://arxiv.org/abs/2005.14165.

Chen, C.-C., Huang, H.-H., and Chen, H.-H. (2020). "Fine-Grained Financial Opinion Mining: A Survey and Research Agenda," https://arxiv.org/abs/2005.01897.

Derrick, K. (2024). "ESG Sentiment Analysis: Comparing Human and Language Model Performance Including GPT," https://arxiv.org/abs/2402.16650.

Dorfleitner, G. and Zhang, R. (2024). "ESG News Sentiment and Stock Price Reactions: A Comprehensive Investigation via BERT," *Schmalenbach Journal of Business Research* **76**, 197–244. https://link.springer.com/article/10.1007/s41471-024-00185-3.

Hu, Z., Meng, Y., Wu, F., and Li, J. (2018). "Conditional BERT Contextual Augmentation," https://arxiv.org/abs/1812.06705.

Huang, A. H., Wang, H., and Yang, Y. (2023). "FinBERT: A Large Language Model for Extracting Information from Financial Text," *Contemporary Accounting Research* **40**, 806–841.

Loughran, T. and McDonald, B. (2011). "When is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks," *The Journal of Finance* **66**, 35–65. https://doi.org/10.1111/j.1540-6261.2010.01625.x.

Maslej, N., Fattorini, L., Perrault, R., Parli, V., Reuel, A., Brynjolfsson, E., Etchemendy, J., Ligett, K., Lyons, T., Manyika, J., Carlos Niebles, J., Shoham, Y., Wald, R., and Clark, J. (2024). "Artificial Intelligence Index Report 2024," https://arxiv.org/pdf/2405.19522.

Sustainability Accounting Standards Board. https://sasb.ifrs.org/.

Tetlock, P. C., Saar-Tsechansky, M., and Macskassy, S. (2008). "More Than Words: Quantifying Language to Measure Firms' Fundamentals," *The Journal of*

Finance **63**, 1437–1467. https://doi.org/10.1111/j.1540-6261.2008.01362.x.

Yang, Y., Mark Christopher Siy, U. Y., and Huang, A. 2020. "FinBERT: A Pretrained Language Model for Financial Communications," https://arxiv.org/abs/2006.08097.

Keywords: AI; expert-annotation; ESG; LLMs; language models; sentiment; sustainable investing

JEL Classification: C38, C45, C54, G11, G41