Predicting Company ESG Ratings from News Articles Using Multivariate Timeseries Analysis

Tanja Aue
University of Innsbruck
Innsbruck, Austria
tanja.aue@student.uibk.ac.at

Adam Jatowt
University of Innsbruck
Innsbruck, Austria
adam.jatowt@uibk.ac.at

Michael Färber
Technical University of Dresden
Dresden, Germany
michael.faerber@tu-dresden.de

Abstract

In recent years, corporate environmental, social, and governance (ESG) engagement has received significant public attention. As mandatory ESG reporting is increasingly adopted and investors place greater emphasis on sustainability in their decisions, the demand for transparent and reliable ESG ratings is growing. However, existing automatic approaches to ESG rating prediction remain limited. Many rely on traditional machine learning methods like random forests or social network analysis, rather than leveraging incoming news article streams and large multivariate time series data, which, for the first time, enables capturing the dynamic relationships between topics, sentiments, and events. In this paper, we propose a novel approach to predicting ESG ratings from news articles by uniquely combining multivariate time series construction with advanced deep learning techniques. We create an extensive dataset of 3.7 million news articles spanning three years and covering 3,000 U.S. companies, providing a robust foundation for training and evaluating our approach. Our approach achieves high accuracy and outperforms existing approaches, underscoring its potential as a scalable, data-driven solution for ESG rating prediction.

CCS Concepts

• **Information systems → Content analysis and feature selection**.

Keywords

ESG ratings, financial applications, news articles, fintech

ACM Reference Format:

Tanja Aue, Adam Jatowt, and Michael Färber. 2025. Predicting Company ESG Ratings from News Articles Using Multivariate Timeseries Analysis. 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, 7 pages. https://doi.org/10.1145/3701716.3717509

1 Introduction

For many companies, enhancing ESG performance and minimizing ESG-related risks are now strategic priorities [24]. Negative ESG news, in particular, poses significant reputational and financial risks. Despite the growing importance of ESG, research on automatic ESG assessment remains underdeveloped [4, 8, 12, 20]. For example, Sokolov et al. [20] focus on classifying Twitter content for ESG relevance using basic indices, while Borms et al. [4] predict ESG scores from news data based on simple aggregation functions and a small, manually labeled dataset. Other approaches avoid advanced machine learning, relying on simpler models like tuned feedforward neural networks [8, 12]. These limitations hinder scalability and the integration of complex data sources, particularly time-sensitive, multivariate data like news articles, which serve as a natural source for signals about ESG status and ESG-related events and incidents.

To address the current gaps in automated ESG rating prediction, we present a framework that efficiently constructs multivariate time series from a comprehensive news dataset and applies advanced deep learning techniques for ESG rating predictions. Our approach integrates ESG ratings for 3,000 U.S. companies with a dataset of over 3.7 million news articles, spanning three years and sourced from over a thousand outlets worldwide. This scale and diversity in data enable a robust analysis of company behavior over time, capturing trends across various geographic regions and sectors. Our pipeline processes news data in several steps: First, we classify articles for ESG relevance; next, we analyze sentiment and extract content features to produce a broad informational foundation for ESG assessment. Finally, this multivariate time series representation feeds into our deep learning models, specifically CNN and transformer-based architectures, to predict ESG ratings. In our evaluation on our large-scale dataset, we can show that our approach outperforms several competitive baselines. Overall, our approach provides the basis for (1) reducing costs by automating ESG predictions and augmenting human judgment with NLP, which increases accessibility also for small and medium-sized enterprises; (2) enhancing transparency by allowing stakeholders to review and understand the rating process; and (3) enabling scalability by efficiently processing continuously updated, time-sequenced data, making rapid and data-driven ESG assessment feasible.

To summarize, our main contributions are as follows:

(1) We compile and analyze an extensive ESG-focused news dataset spanning 3,000 US companies over three years and drawing from more than a thousand sources worldwide. This dataset surpasses previous work limited to single-source or region-specific news data [4, 15, 19].

(2) We develop a multivariate time series approach that captures key ESG signals, including sentiment, semantic, and topical dimensions, and apply CNN and transformer-based models to predict ESG ratings with high accuracy.

WWW Companion '25, Sydney, NSW, Australia

ACM ISBN 979-8-4007-1331-6/2025/04
https://doi.org/10.1145/3701716.3717509

(3) We rigorously evaluate our model against state-of-the-art methods and provide an in-depth analysis of its capabilities and performance.

2 Related Work

Machine Learning Approaches for ESG Rating Prediction. For ESG rating prediction, a few works have explored machine learning models. Krappel et al. [12] utilize an ensemble approach combining models like XGBoost, CatBoost, and a Feedforward Neural Network, while D'Amato et al. [8] rely on a random forest algorithm. Del Vitto et al. [6] apply Linear, Lasso, and Ridge regression, as well as a fully-connected neural network, but focus on structured financial and company data rather than news articles. Ang et al. [2] predict ESG ratings by examining dynamic inter-company relationships and assess how different types of financial and network information impact ESG prediction.

Semi-Supervised and Advanced NLP Techniques Borms et al. [4] demonstrate that ESG indices based on Flemish-Dutch news articles can predict negative ESG score adjustments several months in advance. Using a semi-supervised text mining method, they develop ESG indices for 291 European companies from a corpus of 365,319 Dutch-language articles. Manually defining seed words for the ESG dimensions and negative sentiment, they expand these keywords using GloVe embeddings. Their approach involves setting up a matrix with eleven frequency-based and six sentiment-adjusted indicators to quantify ESG information from news. In contrast, Sokolov et al. [20] leverage the BERT model to classify ESG-relevant social media content, showing that accuracy in identifying ESG-related content improves with a BERT classifier that includes an additional fully connected layer. They prefilter each ESG category with keywords and human review, labeling 1,468 ESG-relevant tweets as positive or negative out of 6,000 tweets from Twitter.

Gap to Related Works. Building on the idea of predicting ESG ratings from news data [4], our approach extends this by employing a multivariate time series model that captures various facets of company dynamics, including shifts in article volume, sentiment trends, and ESG-related semantic information. This model processes news streams alongside large-scale multivariate time series data, enabling, for the first time, a deeper capture of evolving relationships between topics, sentiments, and events. Using these time series, we train a deep learning model to predict ESG ratings. Unlike the manual keyword-based approach in [4] and [20], our method is fully automated and utilizes a substantially larger dataset, encompassing a broader spectrum of companies and news articles.

3 Dataset

In this section, we first outline how we collect ESG ratings for U.S. companies. We then show the process of creating our news article dataset, which includes articles about these companies linked to their ESG ratings.

3.1 ESG Ratings Dataset Creation

We leverage Asset4 ESG ratings from Refinitiv[1], a global provider of financial market data, as a reference point for our predictions. We collected yearly ratings for 3,343 U.S. companies rated by Refinitiv from 2018 to 2020, allowing us to focus on English-language news.

Refinitiv's ESG scores are organized around the three ESG key dimensions, each comprising several themes:

- *Environmental*: resource use, emissions, innovation
- *Social*: workforce, human rights, community, product responsibility
- *Governance*: management, shareholders, CSR strategy

Refinitiv provides three distinct ratings: (1) basic ESG scores, (2) ESG controversy scores, and (3) combined (controversy-adjusted) ESG scores. The basic ESG score is based mainly on information published in company reports, while the controversy score derives from global media coverage, penalizing companies with more negative and controversial news. The combined score, adjusted downward by controversy levels, offers a more comprehensive view of a company's ESG performance [16]. Each rating is on a scale from 0 to 100, with higher scores indicating better ESG performance. We use the combined score as our ground truth for prediction.

In addition to ESG ratings, we retrieved each company's full and abbreviated names and market capitalization from Refinitiv. Market capitalization, calculated as the stock price multiplied by outstanding shares, is widely used to classify company size in finance. We use this measure later to analyze how company size affects prediction accuracy.[2]

3.2 News Dataset Creation

To predict ESG ratings from news articles, a comprehensive and high-quality dataset of ESG-related news is required. For this purpose, we utilize the GDELT project database [10], a widely-used resource in interdisciplinary research [3, 23]. GDELT offers an extensive database of news covering print, web, broadcast, and television sources from most countries around the world in over 100 languages, with updates every 15 minutes. This global scope and real-time nature make GDELT an ideal source for capturing ESG-related news coverage across diverse contexts.

Using the GDELT Doc API Client, we extracted up to the top 250 articles per month over three years (2018, 2019, and 2020) for each company. Monthly aggregation proved optimal in terms of data relevance; shorter time units (e.g., daily or weekly) increased the inclusion of unrelated articles, while monthly intervals yielded a more balanced representation of relevant ESG content. Thus, each company's initial dataset includes a maximum of 3,000 articles per year, ensuring both depth and specificity in ESG news coverage.

To capture ESG themes accurately, we designed our keyword search to include the 10 primary themes within the three ESG dimensions (see Sec. 3.1) along with both short and full company names (e.g., "Visa" and "Visa Corporation"). Following [4], we required that the company's name appear at least twice in each article to ensure focused content and exclude incidental mentions.

[1] https://www.refinitiv.com/en/sustainable-finance/esg-scores

[2] For this analysis, we categorize companies as large caps (>$10 billion), mid caps ($2-$10 billion), and small caps (<$2 billion) based on market capitalization, though exact thresholds may vary [1].

Table 1: Overview of the News Article Dataset.

	2018	2019	2020	In Total
# of companies	2,665	2,940	2,879	8,484
# of articles	1,165,840	1,399,277	1,174,754	3,739,871
# of articles per company				
average	435	473	406	440
max	1,984	2,154	2,005	2,154

Once we retrieved the list of URLs from GDELT, we crawled each article and preprocessed the dataset to enhance quality and consistency. We removed articles shorter than 200 characters and those with HTML errors, then extracted text content using the Beautiful Soup library[3]. We excluded companies with fewer than five collected articles to maintain a robust sample for each entity.

In a final cleaning step, we removed extraneous noise, such as links, email addresses, dates, Unicode characters, and newline representations, which are common byproducts of scraping. Due to the length limitations of the BERT models we employ, each article was reduced to its title and the first five sentences, ensuring manageable input size while retaining core content.

Table 1 provides an overview of the cleaned dataset. Following the outlined cleaning process, our final dataset comprises 3,739,871 articles corresponding to 8,484 ESG ratings. The majority of articles (72%) originate from U.S. sources, while coverage from the U.K., India, and Canada accounts for 7%, 3.4%, and 2.5% of the total, respectively. The dataset is diverse in domain representation, with no single source dominating; the largest sources, *prnewswire.com* and *news.yahoo.com*, represent only 7% and 4% of the articles.

4 Methodology

An overview of our approach is shown in Figure 1. The pipeline begins with a classification step, where articles are categorized as either ESG-relevant or irrelevant, ensuring a targeted focus on content pertinent to ESG factors. In the second step, we analyze the sentiment for both relevant and irrelevant articles on a per-company basis, capturing nuances in how companies are portrayed in relation to ESG topics. In the third step, articles are clustered by semantic similarity, forming coherent groupings that reflect distinct ESG themes (see Sec. 3.1).

This process results in nine distinct time series for each company each year: one time series representing Article-to-ESG relevance, two sentiment time series (one for relevant and one for irrelevant articles), and six semantic time series, one per cluster. These time series together form a multidimensional input that captures both the quantity and thematic variety of ESG-related content for each company.

To predict ESG ratings based on this enriched input, we employ various neural network configurations, including Transformer architectures and convolutional layers, which are tailored to model the temporal and thematic patterns in the time series.

Each step of our approach is detailed below.

[3]https://pypi.org/project/beautifulsoup4/

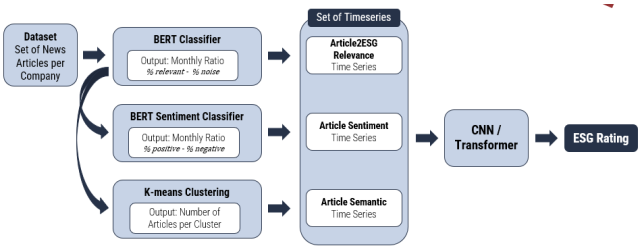

Figure 1: Overview ESG Prediction Model

4.1 Text-based ESG Relevance Classification

Determining whether an article contains relevant information about a company's ESG performance and activities is a crucial step in our pipeline. For this task, we fine-tune BERT specifically for ESG relevance classification. Since manually labeling thousands of articles would be time-consuming, we adopt a weak-supervision approach, structured as a two-step process:

In the **first step**, we pre-label a subset of articles by assessing their semantic similarity to standard definitions of the three ESG dimensions. We use ESG definitions from the Corporate Finance Institute (CFI), a respected resource in finance, to guide the labeling.[4] The pre-labeled subset, consisting of articles on 100 randomly selected companies, serves as the training, validation, and testing base for the classifier. We create pre-trained text embeddings for these ESG definitions and the news articles using the sentence transformer model 'all-distilroberta-v1' (SBERT) from Huggingface [17, 22]. Pairwise cosine similarity between an article's content and each of the 10 ESG category descriptions is computed to determine relevance, with articles scoring above a 0.1 threshold labeled as relevant.[5]

In the **second step**, we fine-tune a pre-trained DistilBERT model [18] on this pre-labeled dataset to predict labels for the entire corpus. Articles are tokenized with the DistilBertTokenizer from Huggingface [22], and the tokenized inputs are then fed into DistilBERT, with padding applied for shorter articles. The last hidden state corresponding to the classifier token 'CLS' serves as the input to the subsequent layers. This vector is processed by a dense layer of size 265, followed by a dropout layer with a rate of 0.5, before reaching a final dense layer with softmax activation to classify articles as relevant or irrelevant. The model architecture draws from [20].[6]

The DistilBERT classifier is fine-tuned in two stages. In the first stage, only the customized classification layers are optimized, while in the second stage, the entire network is fine-tuned using a lower learning rate to ensure stability [19, 20]. Both training stages are run for 4 epochs; however, in most configurations, the second stage stops after 2 epochs, as early stopping criteria are met. Using the trained classifier, we then predict the relevance label for each article.

[4]Definitions available at https://corporatefinanceinstitute.com/resources/knowledge/other/esg-environmental-social-governance/

[5]A low threshold was chosen because the ESG definitions are broad, making high thresholds overly restrictive. Testing various thresholds, we found that 0.1 produced a balanced distribution of 54% relevant to 46% noise, while a 0.2 threshold led to only 17% relevant articles, which was too limiting given our initial filtering by ESG-related keywords.

[6]We compared this architecture to alternatives, including additional dense layers or bidirectional LSTM layers, but found no improvement in classification performance.

To generate the Article2ESG Relevance time series, which reflects the monthly relevance of articles to ESG topics, we group the articles by month and by company. The monthly relevance value, $rel - noise_m$, is calculated according to the following formula:

$$rel - noise_m = (Relevant_m/N_m) - (Noise_m/N_m) \quad (1)$$

Here, the monthly relevance-to-noise ratio, $rel - noise_m$, is computed by subtracting the noise ratio from the relevance ratio for each month. $Relevant_m$ is the count of relevant articles for a given company in month m, while $Noise_m$ represents the count of noise articles. N_m is the total number of articles for that month. We use subtraction rather than a direct ratio to avoid division issues when the noise ratio is zero. This approach results in a positive $rel - noise_m$ value if the proportion of relevant articles exceeds that of noise articles for a given month, and a negative value otherwise.

The resulting values range between -1 (indicating only noise articles) and 1 (indicating only relevant articles). Thus, for each company, the *Article2ESG Relevance time series* comprises 12 $rel - noise_m$ values, one for each month of the analyzed year.

4.2 Sentiment Analysis

In the second step, we analyze the sentiment of the labeled ESG-relevant articles. For this task, we employ the SieBERT model [11], which provides reliable sentiment predictions. The model's output labels are used to generate a monthly article-sentiment time-series for each company, calculated separately for ESG-relevant and irrelevant articles.

To compute the sentiment time series, we group articles by company and month and calculate a monthly sentiment ratio:

$$pos - neg_m = (Positive_m/N_m) - (Negative_m/N_m) \quad (2)$$

Similar to the $rel - noise_m$ ratio (Eq. (1)), the monthly sentiment ratio, $pos - neg_m$, is determined by subtracting the negative sentiment ratio from the positive sentiment ratio. Here, $Positive_m$ represents the count of positively labeled articles for a given company in month m, while $Negative_m$ is the count of negatively labeled articles. N_m is the total number of articles for that month. This approach yields a positive $pos - neg_m$ value if positive sentiment articles outnumber negative ones and a negative value otherwise.

The final *Article Sentiment time series* is constructed by concatenating these monthly values, resulting in a time series of 12 sentiment values per company for the analyzed year.

4.3 Semantic Analysis

The third step, semantic analysis, groups articles by content, enabling us to track the evolution of key ESG-related topics discussed in the news over time.

We begin by fine-tuning the DistilBERT model for the task of ESG classification to create task-specific embeddings for the articles. These embeddings serve as input for the k-means clustering algorithm.[7] To determine the optimal number of clusters k, we apply the elbow method, which identified 6 as the most suitable number of clusters.

[7]We also experimented with the spherical k-means algorithm, but the results were similar to those obtained with standard k-means.

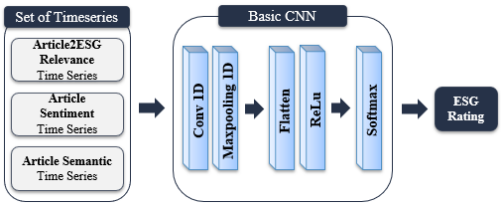

Figure 2: Basic CNN Model

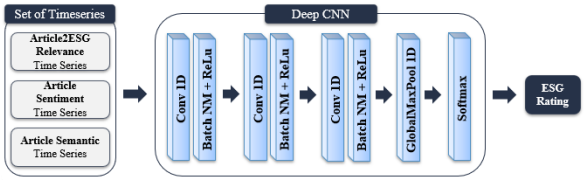

Figure 3: Deep CNN Model

The results of this semantic analysis are integrated into the final ESG rating prediction model as time series data, with each time series corresponding to one of the six clusters. This process yields 6 semantic time series per company, created through the following steps. First, articles are grouped by cluster to identify the total set of articles within each cluster. Then, within each cluster, articles are grouped by company to quantify the overall topic representation for each company.

The monthly values for each time series corresponding to cluster l are calculated by counting the number of articles per month that belong to that cluster. The resulting 12 monthly values per company and cluster are concatenated, producing 6 semantic time series for each company. These time series capture the monthly prominence of each topic cluster, showing how dominant specific ESG topics are for each company over time.

4.4 ESG Rating Prediction

In this section, we propose four models for generating ESG ratings based on the prepared time series data: two CNN-based approaches and two Transformer-based approaches.

CNN-based Approaches. The first model we experiment with is a basic convolutional neural network, shown in Fig. 2. The nine constructed input time series for each of the 12 months—specifically, the *Article2ESG Relevance*, *Article Sentiment* (2 time series), and *Article Semantic time series* (one for each of the 6 clusters)—form the model input as a (9 x 12) matrix. Before feeding the data into the model, we standardize the time series using standard scaling (i.e., subtracting the mean and dividing by the standard deviation). The standardized time series are then processed by a 1D convolutional layer, followed by a 1D max pooling layer. The output of these layers is flattened, and a dense layer with ReLU activation is added to introduce non-linearity into the model [5]. In the final classification layer, the softmax function is applied to generate the probability of each class (i.e., ESG rating values). For each sample, the class with the highest probability is selected as the predicted ESG rating.

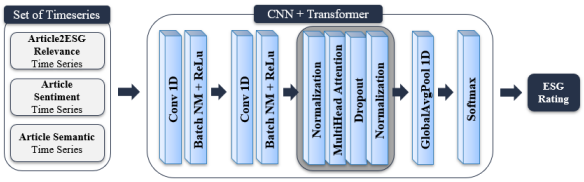

Figure 4: CNN with a Single Transformer Layer

Table 2: Classification Results

Max Length	*512*	*Training Epochs*	*4*	*2*
Batch Size	*32*	*Learning Rate*	*1e-4*	*2e-5*
	precision	**recall**	**f1-score**	**support**
noise	0.80	0.79	0.80	23,938
ESG relevant	0.85	0.86	0.85	33,150
accuracy			0.83	57,088
macro avg	0.83	0.83	0.83	57,088
weighted avg	0.83	0.83	0.83	57,088

Building on this basic CNN, we introduce a deeper model with three convolutional blocks, as illustrated in Figure 3. Here, each of the three convolutional blocks consists of a convolutional layer, a batch normalization layer, and a ReLU activation. Batch normalization standardizes the input for each layer, which has been shown to accelerate model training [5]. Following these convolutional blocks, a global max pooling layer is applied.

Transformer-based Approaches. Alongside the CNN models, we explore two Transformer-based models: (a) a basic model integrating CNN with a single Transformer layer, and (b) a deeper model that combines CNN with a multi-layer Transformer encoder. The basic model, incorporating two convolutional blocks and a single Transformer layer, is shown in Figure 4.

The Transformer encoder block [21] includes a multi-head attention layer, a dropout layer, and two normalization layers (highlighted in gray in Figure 4). Normalization is applied to the input from the second convolutional block as well as to the output of the attention block. The core multi-head attention layer is complemented with dropout to improve generalization. As with the deep CNN, global max pooling followed by the softmax function is applied to produce class predictions.

The deeper model expands on this architecture by replacing the single Transformer layer with three consecutive Transformer encoder blocks, as shown in Figure 4. The CNN component consists of two convolutional layers with filter sizes of 64 and 128, and kernel sizes of 3 and 1, respectively. Each Transformer layer contains 8 heads of size 200, with a dropout rate of 0.2.

Each of the four model architectures operates as a time series classification model, where each potential rating is treated as a class. For instance, Asset4 ratings, which range from 0 to 100, can be modeled as 100 discrete classes. However, ESG rating prediction can also be approached as a regression problem. To test both methods, we replace the softmax layer with a linear dense layer in the regression setup, producing a continuous output between 0 and 100 as the predicted ESG rating.

5 Experiments

5.1 Experimental Settings

In addition to comparing all our proposed models, we evaluate our approach on our dataset against several baselines, including (a) random selection, (b) majority class (mean Asset4 Rating), and (c) the method proposed by [20]. The approach by [20] constructs ESG scores using the output of their ESG classification model and a BERT classifier. Specifically, the scores are calculated by averaging the predicted probabilities for each set of input documents per company and day. Since the Asset4 ratings are on an annual basis, we aggregate the daily scores to a yearly score for comparison by averaging daily predictions across each year.

Default model parameters are applied across all four proposed models to ensure a certain level of comparability. We utilize the rectified Adam optimizer proposed by [14], which has shown improved performance over earlier optimizers. A standard learning rate of 1e-3 is used, and early stopping criteria are applied to mitigate overfitting. Additionally, the learning rate is reduced if five consecutive steps show minimal validation loss changes below a 0.01 threshold. While most models are trained for a maximum of 25 epochs, early stopping is typically triggered earlier. An exception is the Transformer regression model with three Transformer blocks, which is trained for up to 50 epochs as it benefits from extended training.

For the classification models, we use sparse categorical accuracy and sparse categorical cross-entropy loss to evaluate performance. The regression models are assessed using mean absolute percentage error and mean squared error, which are standard metrics for regression evaluation. Although the metrics and losses from the classification and regression models are not directly comparable, the deviation from the Asset4 ratings provides a common basis. Since the difference between the Asset4 and predicted ratings can be positive or negative, we use the average of the absolute differences for comparison.

5.2 Evaluation of ESG Relevance Classification

We begin by evaluating the classification accuracy of ESG-related articles used to construct the Article2ESG relevance time series. Table 2 (upper half) presents the training statistics of the DistilBERT model with the selected configuration. Various model parameters were tested, with the best results achieved using a maximum input length of 512, a batch size of 32, a learning rate of $1e^{-4}$ for the first training stage, and $2e^{-5}$ for the second stage. The model is trained for 4 epochs in each stage; however, in the second stage, overfitting begins after the second epoch. The model was trained on approximately 10,500 articles, with a validation set of 2,100 articles.

Using the fine-tuned weights, the DistilBERT classification model is evaluated on a test set of 57,088 articles. This test set represents the remaining articles from a subset of 100 randomly sampled companies, after excluding training and validation data. The classification report in Table 2 shows an overall accuracy score of 83%. For the ESG-relevant class, which is of primary interest, the model achieves a precision of 85% and a recall of 86%. For the noise or irrelevant class, the model attains a precision of 80% and a recall of 79%. Although not perfect, the proposed approach is a solid basis for generating the relevance time series to be used as input in the final model at a sufficient level.

Table 3: Model Results for the year 2020

	Overall Results				*Absolute Difference to Asset4 Rating*		
Classification Models	**acc**	**scc**			**mean**	**std**	**max**
Basic CNN	2.2%	4.29			15.27	12.23	64.60
Deep CNN	**2.3%**	**4.38**			15.72	12.43	58.10
CNN + Transformer	**2.3%**	4.49			15.81	12.57	62.40
CNN + Deep Transformer	2.2%	5.29			16.44	12.80	67.80
Regression Models			**mape**	**mse**			
Basic CNN			51.31	285.07	13.43	10.20	52.62
Deep CNN			50.12	**263.18**	**13.04**	9.66	48.63
CNN + Transformer			44.18	368.73	15.05	11.88	54.41
CNN + Deep Transformer			50.66	294.75	13.36	10.59	73.38
Baselines	**acc**		**mape**	**mse**			
Mean Asset4 Rating	2.2%		**43.64**	365.63	14.79	12.02	55.94
Random Selection	0.9%		102.79	1167.99	28.10	19.48	83.80
Sokolov et al. [20]	0.5%		176.27	2108.27	41.46	19.74	89.89

Table 4: Model Results for the year 2018

	Overall Results				*Absolute Difference to Asset4 Rating*		
Classification Models	**acc**	**scc**			**mean**	**std**	**max**
Basic CNN	**3.1%**	**4.23**			14.35	12.66	62.56
Deep CNN	2.8%	4.38			14.55	12.79	65.60
CNN + Transformer	2.8%	4.49			14.79	12.28	65.30
CNN + Deep Transformer	2.6%	5.59			15.91	13.18	64.80
Regression Models			**mape**	**mse**			
Basic CNN			49.75	254.29	12.44	9.94	59.77
Deep CNN			47.35	**227.42**	**11.85**	9.27	53.07
CNN + Transformer			**42.47**	322.88	13.67	11.62	64.88
CNN + Deep Transformer			49.66	243.50	12.44	9.41	52.77
Baselines	**acc**		**mape**	**mse**			
Mean Asset4 Rating	2.0%		62.82	319.09	14.62	10.25	50.89
Random Selection	1.2%		127.53	1426.66	31.07	21.48	85.11
Sokolov et al. [20]	1.2%		160.89	1530.52	33.62	20.01	89.64

Table 5: Model Results for the year 2019

	Overall Results				*Absolute Difference to Asset4 Rating*		
Classification Models	**acc**	**scc**			**mean**	**std**	**max**
Basic CNN	2.1%	**4.21**			13.60	10.92	55.50
Deep CNN	2.4%	4.33			14.39	11.86	59.50
CNN + Transformer	**2.5%**	4.44			14.90	12.08	62.40
CNN + Deep Transformer	**2.5%**	5.23			15.69	12.65	67.00
Regression Models			**mape**	**mse**			
Basic CNN			46.47	242.51	12.28	9.56	53.38
Deep CNN			45.71	**227.83**	**11.97**	9.19	46.33
CNN + Transformer			**40.25**	348.70	14.47	11.77	60.15
CNN + Deep Transformer			45.90	245.45	12.36	9.59	56.95
Baselines	**acc**		**mape**	**mse**			
Mean Asset4 Rating	**2.5%**		55.88	313.37	14.39	10.40	49.70
Random Selection	0.9%		120.66	1436.93	31.11	21.30	86.60
Sokolov et al. [20]	0.7%		158.60	1713.39	36.34	20.49	90.93

5.3 Evaluation of ESG Rating Prediction

As detailed in Section 4.4, four different models are applied to predict ESG ratings from the constructed time series, each configured for classification and regression tasks. Table 3 presents the results for all model variations as well as baseline models for the year 2020. The reported results represent average values over 10 runs, each conducted with different, randomly initialized train-test splits. Since evaluation metrics for classification and regression models are not directly comparable, we also consider the average absolute difference between the Asset4 ratings and the predicted ratings to assess model performance. For the baseline models, all metrics except for the sparse categorical cross-entropy loss (scc) are calculated, as scc requires class probabilities, which are unavailable.

We can observe that the classification model accuracy is relatively low due to the challenging nature of predicting the correct rating out of 100 classes, and results vary significantly depending on the train-test split. When comparing all models based on the mean absolute difference, the best-performing model is the proposed deep CNN regression model with three convolutional blocks. This model achieves the lowest mean absolute difference, a smaller standard deviation, and a lower maximum deviation, outperforming others. Among the regression models, it also has the lowest mean squared error, although its mean absolute percentage error is slightly higher.

Notably, three out of four regression models outperform all other models significantly in terms of mean difference, with values of 13.04 and 13.34 compared to values between 15.05 and 16.44 for other approaches. All regression models perform better than classification models, indicating that ESG rating prediction is best approached as a regression task. This is reasonable because a prediction slightly above or below the actual rating is treated as incorrect in classification but is more acceptable in regression. We also observe poor performance for the approach by Sokolov et al. [20].

The prediction results for 2018 and 2019, shown in Table 4 and Table 5 respectively, reveal similar trends. The classification models again perform worse than regression models in both years, with the deep CNN regression model consistently achieving the best performance across all periods. In general, the mean absolute difference to the Asset4 ESG ratings indicates slightly better performance in 2018 and 2019, with values between 11.85 and 15.91, compared to 13.04 to 16.44 in 2020. This decrease in accuracy in 2020 may be related to the economic impact of Covid-19.

In addition to the top-performing deep CNN, the basic CNN and deep Transformer regression models show strong performance in 2020. In 2018 and 2019, the basic CNN classification model also outperforms all baselines. The deep CNN classification model and basic Transformer regression model perform better than all three baselines in 2018. The mean Asset4 ratings serve as the best baseline across all three years, outperforming both random selection and the method from [20].

6 Conclusion

Automatically predicting ESG ratings from news, without human intervention, has the potential to enable various stakeholders, including private investors and government agencies, to monitor companies' ESG compliance in a time- and resource-efficient manner. Using news articles from over 30,000 websites in approximately 200 countries, we have shown that ESG ratings can be predicted reliably based on information from news sources. Our input processing pipeline comprises three main steps—article classification for ESG relevance, sentiment estimation, and cluster-based time series generation—to derive inputs for the prediction model. Results indicate that these derived time series significantly enhance ESG prediction performance, underscoring their importance as key features. Additionally, our models demonstrate stronger performance on small-cap companies compared to larger ones.

A limitation of our study is the reliance on a single type of rating, despite its widespread use in both practice and research (e.g., [6, 7]). Given recent discussions around the consistency of ESG ratings [13], future work will explore alternative rating sources. Additionally, we will consider bias detection techniques in news articles [9].

References

[1] Andrew, L. [2021], 'Small-cap vs. Mid-cap vs Large-cap: Why the differences matter for your investments', "https://finance.yahoo.com/news/small-cap-vs-mid-cap-160235008.html". Retrieved on March 21, 2022.

[2] Ang, G., Guo, Z. and Lim, E.-P. [2023], 'On predicting esg ratings using dynamic company networks', *ACM Trans. Manage. Inf. Syst.* **14**(3).
URL: *https://doi.org/10.1145/3607874*

[3] Azhar, N. A., Pan, G., Seow, P.-S., Koh, A. and Tay, W.-Y. [2019], 'Text analytics approach to examining corporate social responsibility', *Asian Journal of Accounting and Governance* **11**, 85–96.

[4] Borms, S., Boudt, K., Van Holle, F. and Willems, J. [2021], Semi-supervised text mining for monitoring the news about the ESG performance of companies, *in* 'Data Science for Economics and Finance', Springer, Cham, pp. 217–239.

[5] Chollet, F. [2021], *Deep learning with Python*, Simon and Schuster.

[6] Del Vitto, A., Marazzina, D. and Stocco, D. [2023], 'Esg ratings explainability through machine learning techniques', *Annals of Operations Research* pp. 1–30.

[7] Drempetic, S., Klein, C. and Zwergel, B. [2020], 'The influence of firm size on the esg score: Corporate sustainability ratings under review', *Journal of Business Ethics* **167**(2), 333–360.

[8] D'Amato, V., D'Ecclesia, R. and Levantesi, S. [2022], 'Esg score prediction through random forest algorithm', *Computational Management Science* **19**(2), 347–373.

[9] Färber, M., Burkard, V., Jatowt, A. and Lim, S. [2020], A multidimensional dataset based on crowdsourcing for analyzing and detecting news bias, *in* 'Proceedings of the 29th ACM international conference on information & knowledge management', pp. 3007–3014.

[10] GDELT [n.d.], 'Overview of the GDELT project', "https://www.gdeltproject.org/". Retrieved on 09.11.2021.

[11] Hartmann, J., Heitmann, M., Siebert, C. and Schamp, C. [2022], 'More than a feeling: Benchmarks for sentiment analysis accuracy', *International Journal of Research in Marketing* .

[12] Krappel, T., Bogun, A. and Borth, D. [2021], 'Heterogeneous ensemble for ESG ratings prediction', *CoRR* **abs/2109.10085**.
URL: *https://arxiv.org/abs/2109.10085*

[13] Larcker, D. F., Pomorski, L., Tayan, B. and Watts, E. M. [2022], 'Esg ratings: A compass without direction', *Rock Center for Corporate Governance at Stanford University Working Paper Forthcoming* .

[14] Liu, L., Jiang, H., He, P., Chen, W., Liu, X., Gao, J. and Han, J. [2019], On the variance of the adaptive learning rate and beyond, *in* 'International Conference on Learning Representations'.

[15] Nugent, T., Stelea, N. and Leidner, J. L. [2021], Detecting environmental, social and governance (esg) topics using domain-specific language models and data augmentation, *in* 'International Conference on Flexible Query Answering Systems', Springer, pp. 157–169.

[16] Refinitiv [2021], 'Environmental, Social and Governance Scores from Refinitiv', "https://www.refinitiv.com/content/dam/marketing/en_us/documents/methodology/refinitiv-esg-scores-methodology.pdf". Retrieved on Nov 09, 2021.

[17] Reimers, N., Gurevych, I., Reimers, N., Gurevych, I., Thakur, N., Reimers, N., Daxenberger, J., Gurevych, I., Reimers, N., Gurevych, I. et al. [2019], Sentence-bert: Sentence embeddings using siamese bert-networks, *in* 'Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing', Association for Computational Linguistics, pp. 671–688.

[18] Sanh, V., Debut, L., Chaumond, J. and Wolf, T. [2019], 'Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter', *arXiv preprint arXiv:1910.01108* .

[19] Sokolov, A., Caverly, K., Mostovoy, J., Fahoum, T. and Seco, L. [2021], 'Weak supervision and black-litterman for automated esg portfolio construction', *The Journal of Financial Data Science* **3**(3), 129–138.

[20] Sokolov, A., Mostovoy, J., Ding, J. and Seco, L. [2021], 'Building machine learning systems for automated ESG scoring', *The Journal of Impact and ESG Investing* **1**(3), 39–50.

[21] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. and Polosukhin, I. [2017], 'Attention is all you need', *Advances in neural information processing systems* **30**.

[22] Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M. et al. [2019], 'Huggingface's transformers: State-of-the-art natural language processing', *arXiv preprint arXiv:1910.03771* .

[23] Wu, C. and Gerber, M. S. [2017], 'Forecasting civil unrest using social media and protest participation theory', *IEEE Transactions on Computational Social Systems* **5**(1), 82–94.

[24] Zumente, I. and Bistrova, J. [2021], 'Esg importance for long-term shareholder value creation: Literature vs. practice', *Journal of Open Innovation: Technology, Market, and Complexity* **7**(2), 127.