DICE @ ML-ESG-3: ESG Impact Level and Duration Inference Using LLMs for Augmentation and Contrastive Learning

Konstantinos Bougiatiotis, Andreas Sideras, Elias Zavitsanos, Georgios Paliouras
Institute of Informatics and Telecommunications, NCSR "Demokritos"
Patriarhou Gregoriou and Neapoleos St., Aghia Paraskevi
{bogas.ko, andreasideras, izavits, paliourg}@iit.demokritos.gr

Abstract

We present the submission of team DICE for ML-ESG-3, the 3rd Shared Task on Multilingual ESG impact duration inference in the context of the joint FinNLP-KDF workshop series. The task provides news articles and seeks to determine the impact and duration of an event in the news article may have on a company. We experiment with various baselines and discuss the results of our best-performing submissions based on contrastive pre-training and a stacked model based on the bag-of-words assumption and sentence embeddings. We also explore the label correlations among events from the same news article and the correlations between impact level and impact length. Our analysis shows that even simple classifiers trained in this task can achieve comparable performance with more complex models under certain conditions.

Keywords: ESG, NLP, machine learning, impact, sustainability, duration, fintech

1. Introduction

Environment, Social, and Governance (ESG) in the financial industry includes environmental, social, and governance issues within a company that may impact its performance. Their effect may be mild, moderate, or severe, and their duration may vary. Each of the three aspects of ESG involves various indicators that contribute to the ESG profile of a company. The environmental element focuses primarily on climate considerations, waste management, and resource preservation. The social direction concerns human rights, employee health and safety, training, and consumer rights protection. The governance dimension is related to board issues, business ethics, and issues related to the company's strategic decisions.

ESG has recently become particularly important, forcing organizations to incorporate ESG criteria into their processes and operations. Assembling a company's ESG profile is critical because of the need to evaluate companies' activities and investments, as well as the adoption of regulations and the transparency of communication about their sustainability. Therefore, it is apparent from the business perspective that ESG issues may impact the company and its investors when there is doubt about its decision-making strategies and sustainability. Given the above, companies must periodically release ESG reports, as they represent an essential guide for potential new investors.

In this context, automating the analysis of ESG reports, indicators, or related news has gained much attention in the academic literature. Recently, an ESG shared task was proposed (Kang and El Maarouf, 2022) in the context of the FinNLP workshop series, including two subtasks that focused on ESG taxonomy enrichment and sustainable sentence prediction. The following year, the task was extended to a multilingual ESG issue identification (Chen et al., 2023) that aimed at integrating the ESG paradigm into financial natural language processing (NLP) systems. The objective of the task was to classify news articles into 35 key ESG issues and identify the affected company and the corresponding industry.

This third task on multilingual ESG inference (ML-ESG-3) aims to determine the impact and duration an event in the news article may have on a target company. This challenging task comprises two subtasks: impact level identification and duration identification, including news articles in five languages. In this work, we present the submission of the team *DICE* for ML-ESG-3, along with the baseline models we experimented with. Our primary focus was on the English language. In this setting, our best system ranked in the 6th position out of 32 submissions in the subtask of impact level identification, while our best-performing system in the subtask of impact duration ranked in the 14th position.

The rest of the paper is organized as follows. Section 2 provides an overview of the related work in the ESG domain. Section 3 presents the datasets given by the organizers and the task design. In sections 4 and 5, we discuss our methods and empirical results, while section 6 concludes the paper and highlights future directions.

Proceedings of the Joint Workshop of the 7th FinNLP, the 5th KDF, and the 4th ECONLP, pages 234–243

2. Related Work

The ESG paradigm has gained increasing attention, especially since 2020. The idea of analyzing ESG data and factors has matured over time, and nowadays, the academic community supports the automated analysis of such data using machine learning (ML) and deep learning (DL) methods that target various aspects and use cases.

A body of work focuses on predicting ESG scores and the related variables and factors that affect these scores. The work in (Gupta et al., 2021) is based on statistical analysis and traditional ML to measure the importance of ESG parameters in financial performance and how they affect investment decisions. Similarly, in (D'Amato et al., 2021) and (D'Amato et al., 2022), the authors aim to identify the variables that affect the ESG score by leveraging random forests, and they conclude that balance sheet items, i.e., numerical indices, constitute significant predictors of the ESG score.

In addition, some work focuses on the impact of ESG data on investments and stock returns. The work in (Utkarsh Sharma and Gupta, 2024) investigates whether ESG data can lead to profitable investments. According to this, the higher the ESG scores, the better the financial performance, especially when ESG data are combined with other financial variables. In another study (Yu et al., 2022), the authors tried to discover the relationship between ESG scores and stock returns using credit rating agency data. Finally, the work in (Margot et al., 2021) uses ML to identify patterns between ESG profiles and the financial performance of companies by mapping ESG data to excess returns.

A common characteristic of the above efforts is that they rely on structured data analysis. However, ESG data are available at several levels and modalities. This variety raises interesting questions from an ESG perspective regarding the implications of differences in ESG data from different providers. For this reason, much work focuses on becoming independent of data providers by using other data sources, such as Corporate Social Responsibility (CSR) reports, company communications, and the news. For example, the work in (Wang et al., 2020) uses the news to classify the relevance and sentiment of the articles to the economy by using DL and traditional ML methods. In (Nugent et al., 2021), the authors analyzed news articles and classified them into twenty ESG categories using domain adaptation and data augmentation techniques to improve classification performance. Using transformer-based language models, the work in (Guo et al., 2020) used news data to examine the impact of ESG issues in financial news and to analyze the predictive power of ESG news on stock volatility.

In the previous multilingual ESG shared task (ML-ESG-2) (Chen et al., 2023) for news classification into ESG issues, most submitted methods focused on large language models. The authors in (Pontes et al., 2023) used RoBERTa and SBERT and found that the best results in both monolingual and multilingual data are achieved with RoBERTa, while the work in (Glenn et al., 2023) relies on fine-tuning multilingual BERT with augmented data produced by GPT-3.5. Similarly, the authors in (Lee et al., 2023) use generative models, zero-shot techniques, and translation to augment the training data and experiment with BERT-based models, such as RoBERTa and FinBERT. The work in (Mashkin and Chersoni, 2023) experiments with transformer representations that were used in traditional ML methods, such as Logistic Regression (LR), Random Forests (RF), and Support Vector Machines (SVM) for classification. Finally, the authors in (Billert and Conrad, 2023) and (Wang et al., 2023) also rely on BERT models. The former exploits a strategy for efficient transfer learning, introduced in (Houlsby et al., 2019), to fine-tune a multilingual BERT, while the latter leverages MacBERT in a contrastive learning framework utilizing pseudo-labeled data.

In this ML-ESG-3 shared task, we experiment with several baselines and focus on our submitted systems based on contrastively pre-trained and stacked models.

3. Datasets and Task Design

The organizers released the datasets in two phases. First, the annotated training data, including five languages, were released, and then, the blind test sets for the corresponding five languages. A training sample from the English dataset with the corresponding fields and values is shown below.


```
{
    "URL": "https://www.esgtoday.com/arabesque-ai-appoints-carolina-minio
        -paluello-as-new-ceo/",

    "news_title":"Arabesque AI Appoints Carolina Minio Paluello as New
        CEO",

    "news_content": "ESG-focused financial technology company Arabesque
        AI announced today the appointment of Dr. Carolina Minio
        Paluello as the company\u2019s new Chief Executive Officer.",

    "impact_level": "low",

    "impact_length": "2 to 5 years"
},
```


As depicted, apart from the news content, we also have the corresponding news title and URL from which the text was extracted. In this work, we focused on the English data, and we submitted systems for the English and French datasets where each text sample is annotated with the following labels:

- **Impact Length**, was selected among "Less than 2 years" ($x < 2$), "2 to 5 years" ($2 < x > 5$), and "More than 5 years" ($x > 5$).

- **Impact Level**, qualifies the opportunity or risk as being “low”, “medium” or “high”.

The English dataset consists of 545/136 train/test samples, while the French dataset is split into 661/146 respectively. The number of samples in each class for the English data is, as shown in Fig. 1, in paired format. The class distribution is not balanced. For the *Impact Length*, 48.62% of the data are annotated as “More than 5 years”, 36.33% of the data are annotated as “2 to 5 years”, and 15.05% of the data concern “Less than 2 years”. On the other hand, the impact level annotations are distributed as follows: 44.59% of the samples belong to the “medium” category, a percentage of 35.96% belongs to “high”, and the remaining 19.45% belongs to the “low” category. An important observation is that the “high” impact level category seems strongly correlated with a duration of “More than 5 years”.

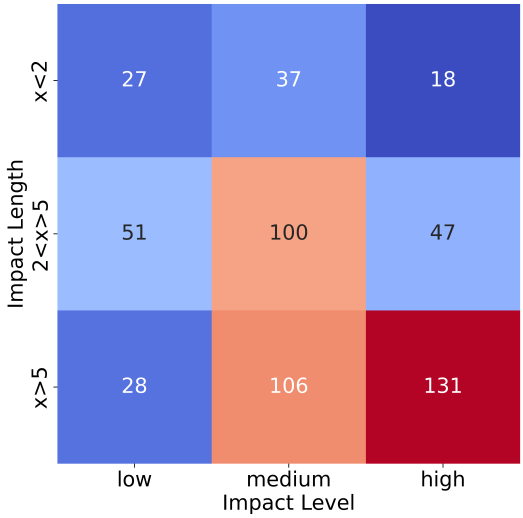

Figure 1: Number of samples in each class for both tasks in the English dataset.

4. Methods and System Selection

The current task entails several intricacies. As previously emphasized, there is a discrete correlation not only between the classes of impact length and level but also between the text snippets originating from the same article. Such instances occur in both the training and test data. Also, we operate within a low-resource environment with limited data. Thus, we experiment with methods that encapsulate the above observations. All our experiments were performed five times, using different splitting seeds on the full English training set, splitting the data in $70\%/10\%/20\%$ train/val/test stratified (concerning class label) splits in each run. The evaluation is performed in terms of macro-averaged F1, also reporting the standard deviations.

4.1. Features and Task Engineering

First, we experimented independently for the length and level identification tasks with ML methods, such as Logistic Regression and input representations like TF-IDF, to establish baseline performance and gain insights regarding the feature importance and problem difficulty. This analysis indicated that the model highly correlates specific people and company names with its prediction. By exploring the dataset, we validated that there are companies (e.g., Microsoft) that are almost always classified into the same classes for both prediction tasks. Also, given that multiple texts belong to the same article, we noticed that their labels match rather frequently. Consequently, we experimented with several pre- and post-processing techniques, as well as different ways to split the data for model selection.

Using a simple TF-IDF vectorization process, we noticed that specific words highly correlate with specific classes. Table 1 provides such examples and shows the number of occurrences of each word, alongside its distribution over the classes. The first set of words, namely “2035”, “2050”, and “trillion”, correspond to simple cases where it is straightforward to deduce the label of the texts containing them, solely using these context words. For instance, it is easy to understand that when talking about things that have a horizon up to 2050, the time context is probably “More than 5 years” ($x > 5$), or when talking about matters in the context of trillions of dollars, the impact level is probably “high”.

Table 1: Example of specific word occurrences and their distribution among both task labels.

Word (Occur.)	**Distribution**
2035 (5)	{x>5:4}, {high:3}
2050 (7)	{x>5:7}, {high:7}
trillion (9)	{x>5:8}, {high:9}
water (38)	{x>5:32}, {high or medium:34}
appoint (30)	{2<x>5:29}, {low:29}
hydrogen (11)	{x>5:7}, {high:10}
microsoft (6)	{x>5:5}, {high:6}
verizon (6)	{x>5:4}, {high:6}
hsbc (6)	{x<2 or 2<x>5:6}

The second group of words, which contains words like “water” and “appoint”, captures ESG-related issues. As expected, the “water-themed” news is mainly of “high” or “medium” impact and always corresponds to $x > 5$ years in terms of impact length, showcasing the long-term gravitas of water management. On the other hand, the word

"appoint" refers to changes in personnel, mainly on the board of directors, and corresponds to "low" impact levels in terms of ESG.

The final group of words focuses on specific companies, for which all related news usually corresponds to "high" and long-term (i.e., $x > 5$ years) impact. One intuitive explanation for these "company-related" news exhibiting the same class could be the size of the companies, as any news related to companies of large capitalization may have severe implications in terms of ESG risks and opportunities. However, another explanation could be that many text samples that refer to specific companies originate from the same URL, hosting a specific news item, and the impact level/length class label is common among the samples in the same news article.

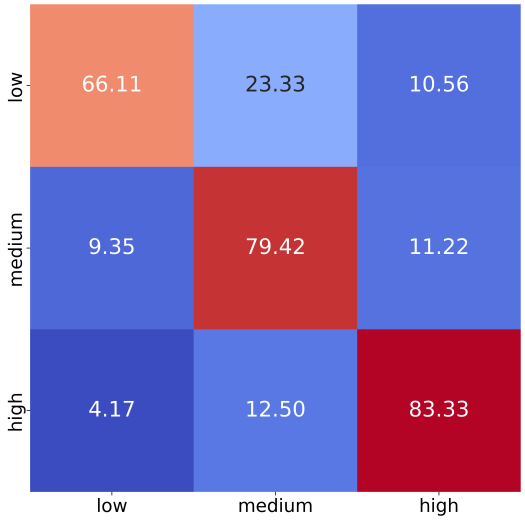

Figure 2: Correlation of class labels among same-URL instances regarding impact level.

Motivated by the above, we measured the correlation between class labels among same-URL samples, as shown in Fig. 2. Each cell $C[i, j]$ indicates the probability of encountering a sample with class label j, related to a specific URL, given that we have already seen a sample with class label i from the same URL (intra-URL class correlation). For example, for a news article containing two distinct text samples and given that one of them has a "high" impact level label, the prior probabilities of the labels for the second text are indicated by the last row of the table, i.e., $\{\texttt{low} : 4.17, \texttt{medium} : 12.50, \texttt{high} : 83.33\}$. This means that with a very high probability, the second text sample will have a "high" impact in most cases, regardless of its content. Finally, the diagonal of the correlation matrix has the highest values, validating our intuition that, in most cases, the news items found in a specific article exhibit the same impact level label.

To empirically validate our intuition, we devised a small-scale experiment, starting with a baseline classification model with a Bag-of-Words (BoW), TF-IDF weighted feature representation for each text, and a Logistic Regression (LR) model on top. We create two variants of this model. The first one uses a Named Entity Recognition (NER) component (we use spaCy (Honnibal and Montani, 2017)) and masks each named-entity identified in the text with a corresponding label string for the entity (e.g. "*Jeff Bezos*" is mapped to "*PERSON*", "*Microsoft*" to "*ORG*" etc.) to anonymize the text and mitigate any information that is bound to specific entities. The second deploys a simple post-processing strategy (dubbed *PostProcess*) using the prior-probability table of Fig. 1. Specifically, at inference time, if the sample for which we predict the labels originates from a URL that was already seen in training, we weigh the predicted class probabilities of the LR model with the corresponding prior probability for this specific URL based on the class labels of the other same-URL texts seen in training. This is a simple way to "steer" the predictions of the classifier toward the "expected" distribution of same-URL texts.

Table 2: Results using a baseline model and its variants on impact level prediction under different stratification splits.

Model	**Impact Level**	
	Class	*Class + URL*
BoW-LR	52.51 $\pm$ 3.05	47.68 $\pm$ 3.55
+*NER*	52.71 $\pm$ 3.91	**48.43** $\pm$ **2.30**
+*PostProcess*	**56.75** $\pm$ 4.06	47.68 $\pm$ 3.55

The performance of these models is reported in Table 2, in terms of macro-F1 averaged across five different runs. We also test their performance under two different stratification methods. The first one, corresponding to the second column in the Table, denotes the vanilla stratification setup based on the class labels. The second one, corresponding to the third column in the Table, is a stratified group split where the samples also follow a group split based on their URLs. In this setup, samples belonging to the same URL are always found in the same split, either train, validation, or test, so there is no intra-URL "leak" among the splits.

Focusing on the vanilla setup first, we observe that adding the NER pre-processing step does not improve the generalization capabilities of the model much. On the other hand, the post-process strategy improved the performance of this simple model significantly, which empirically validates the usefulness of knowing other same-URL labels. For the group-based split, we observe that the performance of the models drops for all variants, indicat-

Table 3: Results of different baseline models on both tasks, on either the full test dataset or focusing only on those test samples belonging to a URL already seen in training. The reported score is macro-F1, averaged over five runs alongside the standard deviation.

Model	**Impact Length**		**Impact Level**	
	Full	*SameURL*	*Full*	*SameURL*
BoW + LR	50.36 $\pm$ 4.17	48.32 $\pm$ 2.30	52.51 $\pm$ 3.05	45.82 $\pm$ 8.47
Emb + kNN(k=5)	47.01 $\pm$ 4.70	43.73 $\pm$ 3.56	50.14 $\pm$ 1.90	50.12 $\pm$ 5.20
SameURL-Labels	-	46.34 $\pm$ 9.23	-	56.59 $\pm$ 4.96
SameURL-BoW + LR	-	45.61 $\pm$ 8.62	-	55.67 $\pm$ 5.24
SameURL-Emb + kNN(k=5)	-	47.53 $\pm$ 4.28	-	56.51 $\pm$ 2.57
Stacked Model (*DICE 1*)	**51.52 $\pm$ 3.87**	**49.54 $\pm$ 5.89**	**59.68 $\pm$ 3.26**	**60.78 $\pm$ 4.66**

ing a much harder setup for the BoW-based model. This can be of interest to the organizers of similar future challenges if they want to restrict the models from taking advantage of the whole news article and making predictions based solely on the given text. Moreover, the post-process variant performs the same as the original baseline. This is expected since there are no cases where the test samples' URLs are in any of the training samples. Finally, the NER variant is the best-performing one (while also decreasing the standard deviation in performance), indicating that over-fitting on specific words that correspond to entities is not good for generalization. Thus, adding a NER pre-processing step could be helpful if the test set was created following this regime. For our submissions, we did not add the NER pre-processing, as the splits given by the organizers did not conform to this setup.

4.2. Baseline Approach

Following the observations mentioned above, we aimed to create a system that could do the following:

1. Capture specific words that are highly correlated with labels. To this end, we use a *BoW+LR* model as before (with no pre-/post-processing techniques).

2. Generalize to cases where the (highly) label-correlated vocabulary from (1) is not useful. To this end, we use a sentence embedding model (Reimers and Gurevych, 2019), specifically *all-mpnet-base-v2*, first to embed the news content of each item and then use a k-NearestNeighbor (kNN) classifier on-top. We denote this model as *Emb + kNN*.

3. Encapsulate information from same-URL training samples when possible to do so. To do this, we create three simple models that activate only in cases where a sample originates from a URL already seen in training.

   (a) *SameURL-Labels*: Calculates the probability of each label based on the frequency of the labels of all same-URL training samples.

   (b) *SameURL-BoW + LR*: Retrieves the BoW representations of all same-URL training samples and aggregates them by summation, using an LR classifier on the resulting feature vectors.

   (c) *SameURL-BoW + LR*: Retrieves the sentence embedding representations of all same-URL training samples and aggregates them by summation, using a kNN classifier on the resulting embeddings.

Having these five base models in place, our first submission is a stacked model that considers the probabilities for each class according to these models as input (i.e., a feature vector of length $3(labels) \times 5(models) = 15$) and uses an LR model for the final classification. The final LR classifier is trained using the predictions of the base models on the validation split. No hyper-parameter tuning is performed here.

The results of these models for both tasks on the English dataset are shown in Table 3. We report the performance both on the vanilla setup of the full (5-fold created) test sets (denoted with *Full*) and focusing only on the test samples that we've already seen in training (denoted with *SameURL*). The *SameURL-* models can only generate predictions for the *SameURL* subset of the test samples, so their performance is omitted (denoted with $-$). Essentially, that means that for the cases where a test sample originates from a URL not seen during training, the stacked model only utilizes the predicted probabilities of *Bow+LR* and *Emb+kNN*.

Regarding the performance of the models, predicting impact length seems much more difficult across all settings than impact level. If we focus on the difference under the *Full* setting between the two tasks, we see that the ensemble of *Bow+LR* and *Emb+kNN* is much more effective in the impact level task, denoting that these models make complementary predictions, while the slight increase in the performance of the ensemble indicates that

they probably make the same mistakes when predicting impact length.

Regarding the *SameURL* setting and models, in the impact length task, the information from the *SameURL* models is not as helpful as in the level task. Interestingly, when we focus only on the *SameURL* test samples, the *SameURL-X* models, which use aggregates of information between the intra-URL data, perform better than the *Bow+LR* and *Emb+kNN* that use the actual test sample. This provides evidence that we should exploit the information from the *SameURL* samples.

4.3. Deep Learning Approaches

Having created the stacked baseline model, we now focus on improving performance, mainly on the impact length task with DL approaches. We experimented with models that utilize contextualized embeddings and incorporate prior knowledge from their pre-training process, whether domain-specific or general. Table 4 presents the performance of all such models.

We began with the generic BERT model (Devlin et al., 2018) in a frozen state, using it as an embedding model for the news content by averaging over the token embeddings of the last layers. Subsequently, we appended two additional layers and trained the model independently on impact level and length tasks. The results were much worse than the previously established baseline. Thus, we moved on to experimenting only with fine-tuned models. The performance of the fine-tuned BERT model, with the same classification heads as above, is shown in the second line in Table 4.

Since ESG-related narrative is too specific and domain-oriented and the amount of available data is limited, there is strong evidence that generic pre-trained models may not capture the linguistic semantics of this particular task. Thus, we experimented with RoBERTa (Liu et al., 2019) and FinBERT (Araci, 2019), which are trained on larger and domain-specific data, respectively. However, they both failed to surpass generic BERT's performance. We therefore focused on learning representations for our data that uncover the actual ESG semantics. SetFit (Tunstall et al., 2022) is an efficient framework for few-shot tuning in low-resource scenarios, where a pre-training representation learning step is evolved. SetFit finetunes a sentence encoder while optimizing a triplet loss. Each triple tuple consists of three samples: two that share the same label (positive pair) and one sample of a different label. Then, it builds a classifier on top. SetFit achieved an improved performance at the expense of being too slow to train. However, it inspired us to implement a Contrastive Learning pre-training step.

Contrastive representation learning (Le-Khac et al., 2020) tries to distinguish between similar and dissimilar samples by comparing them. This unsupervised technique can be used as a pre-training step where the model tries to learn meaningful features to address a downstream task. What we contrast upon is called the "pretext task" and has to be aligned with the downstream task. In other words, when the model addresses this pretext task, it should learn highly informative features for the downstream task.

The pretext task we define is to distinguish between sentences that refer to the same ESG issue. Such sentences would be rephrases of a single news text. Thus, the task involves taking a news text, providing a rephrased version of it, and several other unrelated news texts, with the objective of learning a metric space that brings the original and rephrased sentences closer while distancing the irrelevant ones. We assume that this pre-training step will uncover the underlying semantics of ESG news and that the ensuing classifier will capitalize on this information.

In the contrastive learning setting, we need to define a similarity distribution to sample a positive or a negative sample pair (according to the pretext task). A common approach is to use augmentation techniques to get a positive pair for each sample and treat all the rest as negative pairs. We want an augmentation technique that keeps the ESG-related information intact. We used OpenAI's *gpt-3.5-turbo* model and generated three augmentations per sample with the following prompt: "*Rephrase the following in 3 ways. Use synonyms and keep the length close to the original*". An example of the original text and the corresponding generated augmentations can be seen below:

Original text: ESG-focused financial technology company Arabesque AI announced today the appointment of Dr. Carolina Minio Paluello as the company's new Chief Executive Officer.

Augmentation 1: Arabesque AI, a fintech firm with an emphasis on ESG, today declared the induction of Dr. Carolina Minio Paluello as their new CEO.

Augmentation 2: Today, Arabesque AI, a finance technology corporation focused on ESG, introduced Dr. Carolina Minio Paluello as its latest Chief Executive Officer.

Augmentation 3: Dr. Carolina Minio Paluello was announced today as the new CEO of ESG-dedicated fintech company Arabesque AI.

Having multiple ways to express the same ESG news, we consider a pair consisting of the original text and its augmented version as positive and two randomly selected original texts as negative pairs. In the generated pairs, we always include one original sample. It is possible for the augmentations to include texts with vocabulary that may not necessarily align with the narrative of our original data, along with ambiguities or even meaningless passages. This is why we demanded three of them and also added the sampling technique to address such cases and add variability to the vocabulary.

The contrastive loss used to learn this metric space is the following NTXent loss (Sohn, 2016).

Table 4: Final results on both tasks for the English language. The reported score is macro-F1, averaged over five runs alongside the standard deviation

Models	Impact Length	Impact Level
Stacked Model (*DICE 1*)	**51.52 ± 3.87**	**59.68 ± 3.26**
BERT	48.67 ± 5.20	55.17 ± 6.00
RoBERTa	45.50 ± 4.07	52.65 ± 4.37
FinBERT	45.89 ± 6.11	53.45 ± 3.31
SetFit	50.00 ± 4.00	57.22 ± 6.20
CL Variant 1 (*DICE 2*)	**51.77 ± 6.06**	53.66 ± 3.36
CL Variant 2 (*DICE 3*)	50.57 ± 7.48	52.23 ± 2.38
CL Variant 3	47.25 ± 6.00	50.17 ± 3.43

$$L_q = -\log \frac{\exp\left(q \cdot k_+/\tau\right)}{\sum_{i=0}^{K} \exp\left(q \cdot k_i/\tau\right)} \quad (1)$$

where q is an original sample in our batch of size K, τ is the temperature parameter set to 0.07. k_i is any original sample within the current batch, and k_+ is a positive augmentation of the current q original sample, uniformly sampled from the three available augmentations. We computed the loss per sample and optimized its mean aggregation over the batches. This loss maximizes the dot product (similarity) of the positive samples on the enumerator while pushing away the negative pairs (denominator). We pre-trained the generic BERT with two extra linear layers of 768 dimensions and a ReLU applied between them on the available data for 70 epochs, with a learning rate of $5e-6$, a batch size of 32 and an early stopping criterion of 7 epochs. Keeping only the BERT backbone, we froze its parameters and appended three linear layers, applying ReLU to the first two and dropout to the first one for the downstream classification task. We trained for 30 epochs, with a patient of 6, a batch size of 32, and a learning rate of $1e-3$.

We report three variants of this setting that involve the same pre-trained model. The first one targets impact length and level independently. The second adds the post-processing of the predictions as described earlier, and the third one uses two classification heads and tries to solve both tasks simultaneously. Table 4 summarizes all results plus the Stacked Model for comparison reasons.

The unexpected dominance of BERT over RoBERTa and FinBERT has already been noted. However, we should stress that we did not conduct thorough hyperparameter tuning for these models. SetFit was very promising but too slow and did not allow further experimentation. Additionally, it exhibited considerable variation among the five runs, especially in the impact level task, where it achieved the best macro F1 score. Regarding the Contrastive Learning setting, it is interesting that the post-processing step (Variant 2) resulted in a performance drop, unlike the baseline models, where we observed the opposite effect. That is also the case with the third variant, where we tried to leverage the tasks' correlation depicted in Fig. 1.

5. Official Results

Table 5: Final results on the official test sets, macro-F1 reported.

Models	Length (Rank)	Level (Rank)
DICE 1 - Eng.	37.07 (30)	53.11 (10)
DICE 2 - Eng.	**42.53** (14)	**55.27** (6)
DICE 3 - Eng.	37.84 (29)	55.08 (7)
DICE 1 - Fr.	34.45 (19)	44.80 (11)

Table 5 presents the results for our submissions in the blind test set. There is a noticeable deviation between our anticipated performance and the official evaluation, particularly concerning the impact length task. However, for the impact level task, we are much more aligned with our expectations and rank relatively high on the leaderboard. *DICE 1*, although it was our best-performing model in our evaluation setting, performed poorly. We also noticed a significant decrease in performance when applying our post-processing step to the contrastive pre-trained model to the impact length task. However, the impact level appears to remain unaffected.

Since the workshop organizers released the test set ground truths, we also performed an error analysis. Following our intuitions regarding the information shared between same-URL samples, we analyzed the performance of the models separately on two subsets. The first subset contains all the test samples with URLs that exist in our training set (denoted as *SameURL*). The second contains those that originate from unseen URLs (denoted as *!SameURL*). Tables 6 and 7 display the corresponding test results.

Overall, there is a massive increase in the scores concerning not previously seen URL articles except

Table 6: Performance of submitted models on the test set for impact length, when grouping samples on whether we've encountered a same-URL sample in training (*SameURL*) or not *!SameURL*.

Model	Impact Length		
	Full	*SameURL*	*!SameURL*
DICE 1	37.07	29.15 ↓	48.09 ↑
DICE 2	42.53	35.92 ↓	**52.86** ↑
DICE 3	37.84	28.80 ↓	**52.86** ↑

Table 7: Performance of submitted models on the test set for impact level, when grouping samples on whether we've encountered a same-URL sample in training (*SameURL*) or not *!SameURL*.

Model	Impact Level		
	Full	*SameURL*	*!SameURL*
DICE 1	53.11	**59.52** ↑	44.99 ↓
DICE 2	55.27	48.52 ↓	**63.22** ↑
DICE 3	55.08	44.46 ↓	**63.22** ↑

for the *DICE 1* model on the level task. All the models seem to have overfitted entities found in the training data, with the contrastive models being the ones that generalize better in both cases. Moreover, it is essential to note the effectiveness of the *DICE 1* in utilizing information on the *SameURL* group for the impact level task, as shown in Table 7. This is the only case that performs better on the *SameURL* group than the entire test set. This is in line with the findings of our analysis, as also shown in Table 3, where the models that utilize information from other *SameURL* articles perform very well when predicting the impact level of the sample at hand. This effect is not observed, though, for impact length in both cases as expected (i.e., both in Tables 3, 6), which is due to the much lower intra-URL label correlation.

Concerning the contrastive learning models, we observe a drop in performance for the *SameURL* setting. This drop is probably related to the way we conducted the contrastive pretraining. Due to the pretext task we defined, the embeddings of *SameURL* samples are forced apart because they constitute negative pairs in this context. This, when combined with high intra-URL label correlation (e.g., impact level), has a negative effect on the final downstream task. It would be interesting to incorporate the above observations in the contrastive learning setting, which we leave as a future work.

6. Conclusion

The complex nature of the ML-ESG-3 shared task provides an excellent opportunity to experiment with various methods in the domain of ESG under challenging conditions. In this work, we focused on identifying the impact level and length duration of ESG issues found in news articles, based on the English dataset that the organizers distributed. In this setting, we demonstrated how the correlation between texts originating from the same articles impacts the overall performance of different models. Our explanatory analysis revealed that the class labels, at least in the English data, were closely linked to specific tokens, such as the names of companies, nouns, and verbs related to specific ESG issues. To mitigate this bias in the data, we experimented with various baseline systems, pre-/post-processing techniques, and contrastive pre-training. In both subtasks of the ML-ESG-3, our best-performing system was the one based on contrastive pre-training.

Regarding future directions and following our findings regarding news events originating from the same news article, as well as correlations between impact length and impact level, we plan to focus on methodologies that consider multiple sources of information. For example, information stemming from the latest SEC filing regarding any ESG disclosure or other news sources at the time of the news event under examination, alongside other historical information regarding ESG-related activities of the company.

7. Acknowledgements

The authors would like to acknowledge the financial support of Qualco SA for this project. The opinions of the authors expressed herein do not necessarily state or reflect those of Qualco SA.

8. Bibliographical References

Dogu Araci. 2019. Finbert: Financial sentiment analysis with pre-trained language models. *CoRR*, abs/1908.10063.

Fabian Billert and Stefan Conrad. 2023. Team HHU at the FinNLP-2023 ML-ESG task: A multi-model approach to ESG-key-issue classification. In *Proceedings of the Fifth Workshop on Financial Technology and Natural Language Processing and the Second Multimodal AI For Financial Forecasting*, pages 146–150, Macao. -.

Chung-Chi Chen, Yu-Min Tseng, Juyeon Kang, Anaïs Lhuissier, Min-Yuh Day, Teng-Tsai Tu, and Hsin-Hsi Chen. 2023. Multi-lingual ESG issue identification. In *Proceedings of the Fifth Workshop on Financial Technology and Natural Language Processing and the Second Multimodal*

AI For Financial Forecasting, pages 111–115, Macao.

Valeria D'Amato, Rita D'Ecclesia, and Susanna Levantesi. 2021. Fundamental ratios as predictors of esg scores: a machine learning approach. *Decisions in Economics and Finance*, 44(2):1087–1110.

Valeria D'Amato, Rita D'Ecclesia, and Susanna Levantesi. 2022. Esg score prediction through random forest algorithm. *Computational Management Science*, 19(2):347–373.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Parker Glenn, Alolika Gon, Nikhil Kohli, Sihan Zha, Parag Pravin Dakle, and Preethi Raghavan. 2023. Jetsons at the FinNLP-2023: Using synthetic data and transfer learning for multilingual ESG issue classification. In *Proceedings of the Fifth Workshop on Financial Technology and Natural Language Processing and the Second Multimodal AI For Financial Forecasting*, pages 133–139, Macao. -.

Tian Guo, Nicolas Jamet, Valentin Betrix, Louis-Alexandre Piquet, and Emmanuel Hauptmann. 2020. Esg2risk: A deep learning framework from esg news to stock volatility prediction. *arXiv preprint arXiv:2005.02527*.

Akshat Gupta, Utkarsh Sharma, and Sandeep Kumar Gupta. 2021. The role of esg in sustainable development: An analysis through the lens of machine learning. In *2021 IEEE International Humanitarian Technology Conference (IHTC)*, pages 1–5.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp.

Juyeon Kang and Ismail El Maarouf. 2022. FinSim4-ESG shared task: Learning semantic similarities for the financial domain. extended edition to ESG insights. In *Proceedings of the Fourth Workshop on Financial Technology and Natural Language Processing (FinNLP)*, pages 211–217, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Phuc H. Le-Khac, Graham Healy, and Alan F. Smeaton. 2020. Contrastive representation learning: A framework and review. *IEEE Access*, 8:193907–193934.

Hanwool Lee, Jonghyun Choi, Sohyeon Kwon, and Sungbum Jung. 2023. EaSyGuide : ESG issue identification framework leveraging abilities of generative large language models. In *Proceedings of the Fifth Workshop on Financial Technology and Natural Language Processing and the Second Multimodal AI For Financial Forecasting*, pages 127–132, Macao. -.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Vincent Margot, Christophe Geissler, Carmine De Franco, Bruno Monnier, France Advestis, and France Ossiam. 2021. Esg investments: filtering versus machine learning approaches. *Applied Economics and Finance*, 8(2):1–16.

Ivan Mashkin and Emmanuele Chersoni. 2023. HKESG at the ML-ESG task: Exploring transformer representations for multilingual ESG issue identification. In *Proceedings of the Fifth Workshop on Financial Technology and Natural Language Processing and the Second Multimodal AI For Financial Forecasting*, pages 140–145, Macao. -.

Tim Nugent, Nicole Stelea, and Jochen L Leidner. 2021. Detecting environmental, social and governance (esg) topics using domain-specific language models and data augmentation. In *Flexible Query Answering Systems: 14th International Conference, FQAS 2021, Bratislava, Slovakia, September 19–24, 2021, Proceedings 14*, pages 157–169. Springer.

Elvys Linhares Pontes, Mohamed Benjannet, and Lam Kim Ming. 2023. Leveraging BERT language models for multi-lingual ESG issue identification. In *Proceedings of the Fifth Workshop on Financial Technology and Natural Language Processing and the Second Multimodal AI For Financial Forecasting*, pages 121–126, Macao.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Kihyuk Sohn. 2016. Improved deep metric learning with multi-class n-pair loss objective. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Lewis Tunstall, Nils Reimers, Unso Eun Seo Jo, Luke Bates, Daniel Korat, Moshe Wasserblat, and Oren Pereg. 2022. Efficient few-shot learning without prompts.

Akshat Gupta Utkarsh Sharma and Sandeep Kumar Gupta. 2024. The pertinence of incorporating esg ratings to make investment decisions: a quantitative analysis using machine learning. *Journal of Sustainable Finance & Investment*, 14(1):184–198.

Jingli Wang, Ashok Bhowmick, Mucahit Cevik, and Ayse Basar. 2020. Deep learning approaches to classify the relevance and sentiment of news articles to the economy. In *Proceedings of the 30th Annual International Conference on Computer Science and Software Engineering*, pages 207–216.

Weiwei Wang, Wenyang Wei, Qingyuan Song, and Yansong Wang. 2023. Leveraging contrastive learning with BERT for ESG issue identification. In *Proceedings of the Fifth Workshop on Financial Technology and Natural Language Processing and the Second Multimodal AI For Financial Forecasting*, pages 116–120, Macao. -.

Guangliang Yu, Yukun Liu, William Cheng, and Chun-Te Lee. 2022. Data analysis of esg stocks in the chinese stock market based on machine learning. In *2022 2nd International Conference on Consumer Electronics and Computer Engineering (ICCECE)*, pages 486–493.