CriticalMinds: Enhancing ML Models for ESG Impact Analysis Categorisation Using Linguistic Resources and Aspect-Based Sentiment Analysis

Iana Atanassova[*,†], **Marine Potier**[*], **Maya Mathie**[*], **Marc Bertin**[‡], **Panggih Kusuma Ningrum**[*]
[*]Université de Franche-Comté, CRIT
F-25000 Besançon, France
{iana.atanassova, panggih_kusuma.ningrum}@univ-fcomte.fr,
{marine.potier02, maya.mathie}@edu.univ-fcomte.fr

[†]Institut Universitaire de France (IUF)

[‡]ELICO, Université Claude-Bernard Lyon 1
43 Bd. du 11 novembre 1918, 69622 Villeurbanne cedex, France
marc.bertin@univ-lyon1.fr

Abstract

This paper presents our method and findings for the ML-ESG-3 shared task for categorising Environmental, Social, and Governance (ESG) impact level and duration. We introduce a comprehensive machine learning framework incorporating linguistic and semantic features to predict ESG impact levels and durations in English and French. Our methodology uses features that are derived from FastText embeddings, TF-IDF vectors, manually crafted linguistic resources, the ESG taxonomy, and aspect-based sentiment analysis (ABSA). We detail our approach, feature engineering process, model selection via grid search, and results. The best performance for this task was achieved by the Random Forest and XGBoost classifiers, with micro-F1 scores of 47.06 % and 65.44 % for English Impact level and Impact length, and 39.04 % and 54.79 % for French Impact level and Impact length respectively.

Keywords: ABSA, ESG, Impact level, Impact length, ESG taxonomy, linguistic resources

1. Introduction

After the establishment of Environmental, Social, and Governance (ESG) criteria in 2004 (United Nations, 2004), the incorporation of ESG principles within corporations has become a topic of extensive discussion (Berg et al., 2022). The advent of FinNLP challenges explore the opportunity to employ Natural Language Processing methodologies in this domain (Aue et al., 2022; Del Vitto et al., 2023; Schimanski et al., 2024).

The ML-ESG 2024 shared task focuses on multilingual ESG impact type and duration inference, particularly in languages including English and French. The tasks for English and French involve annotations for "Impact Level" (low, medium, high) and "Impact Length" (less than 2 years, 2 to 5 years, more than 5 years) based on the MSCI ESG rating guidelines (Chen et al., 2024).

Our objective in participating in this task, as CriticalMinds team, is to propose a competitive Machine Learning (ML, low resource) approach and evaluate the contribution of several types of features: manually crafted linguistic resources exploiting the ESG taxonomy, and features derived from aspect-based sentiment analysis (ABSA).

2. Method

In this section, we first introduce the datasets employed in the analysis. We then detail the feature types implemented in our experiments with ML models, along with specifications regarding the feature sets' dimensions. Finally, we describe the procedure for model selection and present the corresponding results.

2.1. Data Description

The datasets used in this experiment cover two languages, English and French. For both languages, the training and test sets were provided in json format, with the following variables for each news article: URL, news_title, news_content, impact_level, impact_length. The latter two variables contain the annotated categories in the training set.

We identified a total of 48 duplicate entries within the French training dataset. These duplicates were excluded from subsequent analyses due to inconsistencies between the 'impact_level' and 'impact_length' labels, which rendered the determination of the correct labels ambiguous. Following this data cleaning processes, Table 1 presents the distributions of annotations for 'Impact Length' and 'Impact Level' for the training datasets.

Proceedings of the Joint Workshop of the 7th FinNLP, the 5th KDF, and the 4th ECONLP, pages 248–253
20 May, 2024.

Table 1: Distribution of annotations in the training sets in English and in French

Category		En	Fr
Impact length	Less than 2 years	82	110
	Between 2 and 5 y.	198	218
	More than 5 years	265	285
Impact level	low	106	117
	medium	243	305
	high	196	191
Total		545	613

2.2. Features extraction and selection

In our experiment, we tested combinations of different types of features that we describe below. We designed five types of features:

1. FastText embeddings (Bojanowski et al., 2017; Grave et al., 2018) word vectors;
2. TF-IDF vectors;
3. Features derived from the ESG taxonomy;
4. Linguistic resources to capture expressions of uncertainty and temporal data;
5. Aspects extracted by ABSA.

To calculate the first two types of features, FastText embeddings and TF-IDF, we used the text from the news_title and news_content fields. These were concatenated, then tokenized and lemmatized using `nltk WordNetLemmatizer`. Stop words were also removed. To reduce the dimension of TF-IDF vectors, we used only the 25 terms having the highest discriminatory power. This value was adjusted experimentally.

For the rest of the features, the original values of news_title and news_content fields were used. We describe these features in more detail in the following subsections.

2.2.1. Features derived from ESG taxonomy

As the task of classifying EGS impact durations and levels is essentially related to the semantics of the ESG taxonomy[1], we used the terms denoting ESG issues, sectors and subsectors in the following way. We defined as features the number of occurrences of the issues, sectors and subsectors in the ESG taxonomy. Moreover, for each issue, sector and subsector, we consider lists of synonym expressions that can be present in the news articles and that were curated manually and represented as regular expressions. The figure 1 shows an example of regular expressions in English related to the ’energy’ subsectors.

```
'energy': [
r'\b[Oo]il\s+(?:and|&)\s+[Gg]as\s+[Dd]rilling|[Oo]il\s+[Dd]rilling|[Gg]as\s+[Dd]rilling\b',
r'\b[Oo]il\s+(?:and|&)\s+[Gg]as\s+[Ee]quipments?\s+(?:and|&)\s+[Ss]ervices?|[Oo]il\s+(?:and
r'\b[Ii]ntegrated\s+[Oo]il\s+(?:and|&)\s+[Gg]as|[Ii]ntegrated\s+[Oo]il|[Ii]ntegrated\s+[Gg]
r'\b[Oo]il\s+(?:and|&)\s+[Gg]as\s+[Ee]xplorations?\s+(?:and|&)\s+[Pp]roductions?|[Oo]il\s+(
r'\b[Oo]il\s+(?:and|&)\s+[Gg]as\s+[Rr]efining\s+(?:and|&)\s+[Mm]arkets?(?:ing)?|[Oo]il\s+(?
r'\b[Oo]il\s+(?:and|&)\s+[Gg]as\s+[Ss]torage\s+(?:and|&)\s+[Tt]ransports?(?:ation)?|[Oo]il\
r'\b[Cc]oal\s+(?:and|&)\s+[Cc]onsumable\s+[Ff]uels?|[Cc]oal\s+[Ff]uels?|[Cc]onsumable\s+[Ff
```

Figure 1: Excerpt from the lists of regular expressions related to the ’energy’ subsectors

2.2.2. Linguistic resources

The prediction of Impact level is related to the notion of uncertainty. For this reason, we used as features the number of occurrences of lists of uncertainty and hedging cues in news_title and news_content. In particular, we used the lists defined in (Atanassova et al., 2018).

For the prediction of Impact length, we created lists of temporal expressions that denote various time spans such as "over the next 2 years", "by 2026", etc. They were implemented as regular expressions and their numbers of occurrences were used as features.

Experimentally, we found that these linguistic resources features improve the micro-F1 scores of our models of about 1 % to 2 %.

2.2.3. Aspects extraction

In our study, we leveraged Aspect-Based Sentiment Analysis (ABSA) to dissect and extract significant aspects from textual content, marking it as an advanced segment of sentiment analysis that precisely pinpoints text components and evaluates the sentiments tied to them (Hua et al., 2023). By integrating a combination of linguistic, statistical, and machine learning techniques, and utilizing resources like annotated datasets, lexicons, and ontologies, ABSA achieves a high level of analytical precision (Fan et al., 2020).

ABSA provides a way to examine the textual aspects, which is particularly useful when working with complex datasets such as ESG news articles. These articles often contain discussions on multiple aspects of ESG criteria within the same paragraph or article. By employing a transfer learning approach with a fine-tuned ABSA model[2], we could effectively parse and understand the nuanced sentiments associated with specific ESG aspects. This selected model, optimized within the SetFit ABSA framework and utilizing Sentence Transformer embeddings (Tunstall et al., 2022), is

[1] https://www.msci.com/our-solutions/esg-investing/esg-industry-materiality-map

[2] *joshuasundance/setfit-absa-all-MiniLM-L6-v2-laptops-aspect* from Hugging Face

particularly suited for natural language understanding tasks, enabling precise analysis at the sentence level in ESG news dataset.

Upon reviewing the ESG news dataset, we noted a predominance of neutral sentiments (82.4 %), reflecting the objective presentation style typical of news articles. However, this neutrality does not diminish the utility of ABSA; on the contrary, it allows us to mine the texts for the specific aspects they discuss, shedding light on crucial ESG themes relevant to corporate conduct. This aspect-oriented analysis method, as supported by Hua et al. (2023), provides a deeper dive into key detail information in texts, reaching beyond the surface level of sentiment polarization.

These extracted aspects were then incorporated as features in our ML model, grouping them by their impact_level and impact_length. We calculated the frequency of these aspect occurrences in the news_title and news_content, where the numbers of occurrences were calculated with respect to several cut-off values of the lists for French and for English. The choice of the cut-off values was optimized through grid search.

Figure 2 shows the aspects detected from the English training set grouped by category.

Table 2 shows the cut-off values that were used for English and French, leading to 17 and 11 derived features, respectively.

Table 2: Aspect lists cut-off values N

En	[10, 25, 50, 75, 100, 150, 200, 250, 300, 350, 400, 500, 600, 700, 800, 900, 1100]
Fr	[25, 50, 100, 150, 200, 300, 400, 500, 750, 1000, 1500]

2.3. Feature set dimensions

We employed Principal Component Analysis (PCA) (Jolliffe, 2002) to reduce the dimensions of some of the sets of features, namely the number of dimensions for the FastText embeddings and for the features derived from the ESG taxonomy. This was necessary for two reasons. Firstly, high-dimensional data can complicate model training and possibly lead to overfitting. Secondly, the features that are based on the linguistic resources and the aspects have a fixed dimension, and therefore we need to find the correct balance between the number of dimensions for these features and the ones derived from the embeddings and the ESG taxonomy.

During the grid search phase of our model optimization, we tested various combinations for the numbers of these dimensions, ranging from 5 to 80 dimensions, to find the best configuration for the prediction of each category. Table 3 presents the dimensions of the different types of features that were used with the best model configurations.

2.4. Model Selection

In order to identify the optimal Machine Learning (ML) models, hyperparameters, and to adjust the number of dimensions that were used for the FastText embeddings and TF-IDF features, we performed grid search on the training set. 20 % of the dataset was used for performance evaluation and the rest was used for training with 4-fold cross validation. We used grid-search by maximizing the micro F1 score to test models, including Support Vector Machines (SVM), Random Forest, Gradient Boosting, Logistic Regression, K-Nearest Neighbors (KNN), Extreme Gradient Boosting (XGBoost), LightGBM, and CatBoost. Key hyperparameters tested included kernel types and regularization parameters for SVM, number of estimators and depth for tree-based models, to distance metrics and weights for KNN. For the implementation of the models we used the python `sklearn`, `xgboost`, `catboost` and `lightgbm` libraries.

Table 4 presents the two best models with their hyperparameters, dimensions of features after PCA and results on the training set.

3. Results

Table 5 shows the results obtained by the CriticalMinds team on the test set. To obtain these results, we executed both the Random Forest (RF) and Extended Gradient Boosting (XGB) models five times each, and then selected the most consistently observed predictions across these iterations.

To show the contribution of the different types of features, table 6 presents the results of both models and compares the scores obtained using: the features derived from embeddings (Emb), for TF-IDF and linguistic resources (LR), with adding the features derived from the ESG taxonomy (F-ESG), and those from ABSA. These results show that the features derived from the ESG taxonomy and ABSA improve the performance in most cases. In particular, adding ABSA derived features improves the micro-F1 scores in 4 cases with 2.85 % on average, while it reduces the performance in three cases but with only 1.87 % on average.

4. Discussion

The use of Aspect-Based Sentiment Analysis (ABSA) as strategy in feature engineering is an original approach that aims to improve the semantic representation of textual data. The results in table 6 show the variable impact of ABSA across

Table 3: Number of dimensions for the different models and types of features

Category	Embeddings	TF-IDF	ESG taxonomy	Linguistic	ABSA-derived	Total
Random Forest						
En Impact Level	19	25	12	4	17	77
En Impact Length	75	25	10	3	17	130
Fr Impact Level	12	25	36	4	11	88
Fr Impact Length	70	25	28	3	11	137
XGBoost						
En Impact Level	20	25	15	4	17	81
En Impact Length	75	25	20	3	17	140
Fr Impact Level	18	25	40	4	11	98
Fr Impact Length	75	25	36	3	11	150

Table 4: Best models and results on the training set

Category	Hyperparameters	Micro-F1
Random Forest		
En Impact Level	'criterion': 'gini', 'n_estimators': 400, 'max_depth': None	86.24 %
En Impact Length	'criterion': 'log_loss', 'n_estimators': 400, 'max_depth': None'	79.82 %
Fr Impact Level	'criterion': 'log_loss', 'n_estimators': 500, 'max_depth': None'	71.54 %
Fr Impact Length	'criterion': 'log_loss', 'n_estimators': 200, 'max_depth': None	66.67 %
XGBoost		
En Impact Level	'learning_rate': 0.1, 'n_estimators': 200, 'max_depth': 9	84.40 %
En Impact Length	'learning_rate': 0.1, 'n_estimators': 400, 'max_depth': 9	77.06 %
Fr Impact Level	'learning_rate': 0.1, 'n_estimators': 300, 'max_depth': 7	65.04 %
Fr Impact Length	'learning_rate': 0.1, 'n_estimators': 400, 'max_depth': 5	68.29 %

Table 5: Micro-F1 and Macro-F1 Scores for Impact Length and Impact Level on the test set

		English		French	
	Model	Impact Length	Impact Level	Impact Length	Impact Level
micro F1	CriticalMinds_1 (RF)	64.71 %	47.06 %	54.79 %	36.30 %
	CriticalMinds_2 (XGB)	59.56 %	42.65 %	46.58 %	39.04 %
	CriticalMinds_3 (RF + XGB)	65.44 %	45.59 %	54.11 %	36.30 %
macro F1	CriticalMinds_1 (RF)	42.81 %	43.16 %	30.33 %	22.48 %
	CriticalMinds_2 (XGB)	41.53 %	39.59 %	32.19 %	37.96 %
	CriticalMinds_3 (RF + XGB)	43.86 %	40.64 %	32.88 %	26.21 %

Table 6: Micro-F1 scores on the training set with different subsets of features. Emb = Embeddings, LR = Linguistic resources, F-ESG = ESG taxonomy features. The last column presents the final results (as in table 5) using Emb+TF-IDF+LR+F-ESG and also Aspect-based Sentiment Analysis features.

Category	Features		
	Emb+TF-IDF+LR	Emb+TF-IDF+LR+F-ESG	All
Random Forest			
En Impact Level	44.85 %	45.59 %	47.06 %
En Impact Length	61.76 %	62.50 %	64.71 %
Fr Impact Level	36.30 %	37.67 %	36.30 %
Fr Impact Length	54.11 %	54.79 %	54.79 %
XGBoost			
En Impact Level	42.65 %	45.59 %	42.65 %
En Impact Length	61.76 %	57.35 %	59.56 %
Fr Impact Level	38.36 %	33.56 %	39.04 %
Fr Impact Length	45.89 %	47.95 %	46.58 %

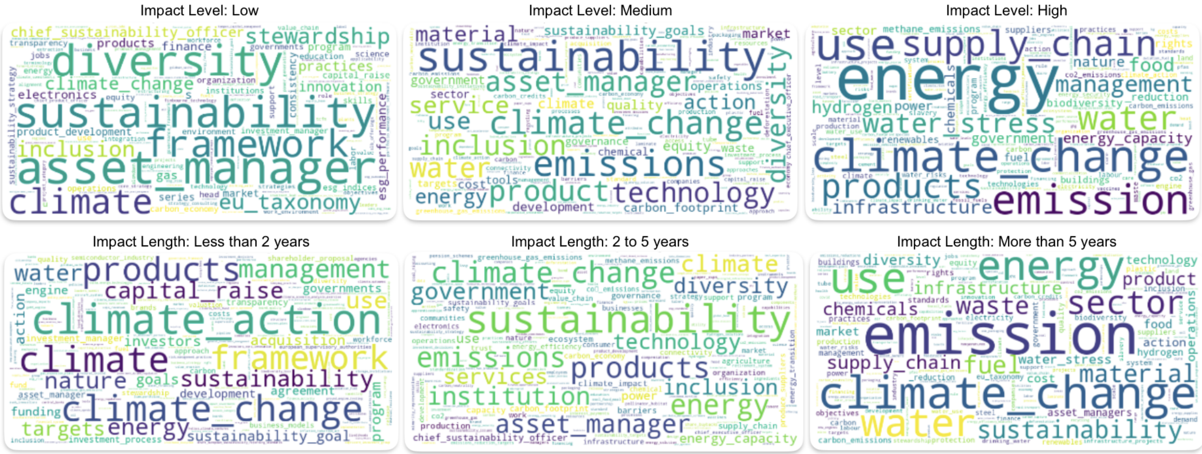

Figure 2: Wordclouds representing aspects detected from the English training set grouped by impact_level and impact_length

different models and languages. Specifically, the inclusion of features from ABSA appears to enhance the predictions in English, underscoring the value of capturing sentiment at a granular level. However, the results also indicate a complex interplay between aspects and other features, suggesting that the contribution of ABSA depends on the model and the linguistic characteristics of the dataset.

The results of our study should be interpreted in the light of several limitations. Firstly, the dependence on linguistic resources makes this approach difficult to deploy for multilingual processing. We specifically curated the lists of regular expressions for English and for French. This task is often time-consuming. We will publish all resources in order to ensure the reproducibility of this experiment.

We choose to use FastText embeddings because of the relatively small size of the models and the ease of use on low-resource machines. FastText embeddings capture subword information and allow representing out-of-vocabulary words. This makes them particularly relevant for processing news articles that may contain numerous new terms and named entities. However, other types of embeddings should be tested as they might improve the results.

The quality of the training data is critical for the success of ML models. During our investigation, we encountered several cases of duplicated annotations, particularly within the French dataset, which were inconsistent and required meticulous cleaning before proceeding with data processing.

Furthermore, in our experimentation, we explored whether augmenting the training set with translated datasets can improve the performance of the models. Specifically, we augmented the training datasets by translating the English dataset into French and vice versa, using ChatGPT-4. Contrary to our expectations, we observed a systematic decline in the performance of all models when the training sets were augmented in this manner. This suggests that the expression of ESG-related information is highly language-specific. This finding underscores the importance of developing language-specific models and training sets for such tasks.

5. Acknowledgements

This work was supported by French ANR grant number ANR-21-CE38-0003-01 (Project InSciM).

6. Bibliographical References

Iana Atanassova, François-C. Rey, and Marc Bertin. 2018. Studying Uncertainty in Science: a distributional analysis through the IMRaD structure. In *7th International Workshop on Mining Scientific Publications (WOSP 2018) at 11th edition of the Language Resources and Evaluation Conference (LREC 2018)*, Miyazaki, Japan.

Tanja Aue, Adam Jatowt, and Michael Färber. 2022. Predicting companies' esg ratings from news articles using multivariate timeseries analysis.

Florian Berg, Julian F Kölbel, and Roberto Rigobon. 2022. Aggregate Confusion: The Divergence of ESG Ratings*. *Review of Finance*, 26(6):1315–1344.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Chung-Chi Chen, Yu-Min Tseng, Juyeon Kang, Anaïs Lhuissier, Hanwool Lee, Yohei Seki, Min-Yuh Day, Teng-Tsai Tu, and Hsin-Hsi Chen. 2024. Multi-Lingual ESG Impact Duration Inference. In *Proceedings of Joint Workshop of the 7th Financial Technology and Natural Language Processing and the 5th Knowledge Discovery from Unstructured Data in Financial Services*.

Alessandro Del Vitto, Daniele Marazzina, and Davide Stocco. 2023. ESG ratings explainability through machine learning techniques. *Annals of Operations Research*, pages 1–30.

Shouxiang Fan, Junping Yao, Yangyang Sun, and Ying Zhan. 2020. A summary of aspect-based sentiment analysis. *Journal of Physics: Conference Series*, 1624(2):022051.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning Word Vectors for 157 Languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Yan Cathy Hua, Paul Denny, Katerina Taskova, and Jörg Wicker. 2023. A systematic review of aspect-based sentiment analysis (absa): Domains, methods, and trends.

Ian T. Jolliffe. 2002. *Principal Component Analysis*, 2nd edition. Springer Series in Statistics. Springer New York, NY.

Tobias Schimanski, Andrin Reding, Nico Reding, Julia Bingler, Mathias Kraus, and Markus Leippold. 2024. Bridging the gap in ESG measurement: Using NLP to quantify environmental, social, and governance communication. *Finance Research Letters*, 61:104979.

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

United Nations. 2004. Who cares wins: Connecting financial markets to a changing world. Technical report, United Nations Global Compact.