# Leveraging Contrastive Learning with BERT for ESG Issue Identification

**Weiwei Wang, Wenyang Wei, Qingyuan Song** and **Yansong Wang**
Technology Innovation Center, Chery HuiYin Motor Finance Service Co.,Ltd.
{wangweiwei, weiwenyang, songqingyuan, wangyansong}@cheryfs.cn

## Abstract

In this study, we, the CheryFS team, present a model solutions dedicated to the task of "Multi-Lingual ESG Issue Identification" in the Chinese track. The objective is to predict the ESG (Environmental, Social, and Governance) label associated with each news article. Our approach integrates supervised and unsupervised data into a comprehensive contrastive learning framework of a MacBERT model with further pretrained. This innovative methodology has resulted in Micro-F1 score of 0.412 on the validation dataset. Furthermore, we perform a meticulous analysis of the model's optimization strategy, providing valuable insights for future research.

## 1 Introduction

Natural Language Processing (NLP) harnesses the capability to extract extensive semantic information from copious volumes of unstructured data, demonstrating immense potential for application in the financial services industry. By analyzing diverse types of unstructured data, including data reports, news articles, text chat records, and research reports, NLP can effectively contribute to scenario recognition and risk analysis in various financial contexts. Commonly, individuals express their opinions on financial products, services, investments, and stock markets through news or social media channels. Thus, the strategic mining of such financial sentiments can inform decision-making, offer valuable advice, and shape user or business understanding.

The "Multi-Lingual ESG Issue Identification"(Chen et al., 2023) subtask aims at uncovering themes related to Environmental, Social, and Corporate Governance (ESG) in Chinese, English, and French news articles. The challenge is defined as follows: Given an article derived from an ESG-focused news website, the model is expected to predict its potentially relevant themes. English and French datasets include a single theme per article, while the Chinese dataset may contain multiple themes. Due to the limitation of time, our team engaged with the Chinese track of this task.

In our research, we incorporated labeled data, unlabeled data obtained through web crawlers, and pseudo-labeled data for data augmentation. Our initial model was constructed around the MacBERT(Cui et al., 2020) architecture. We endeavored to enhance its performance by 1) investigating a variety of data augmentation strategies, 2) implementing further pretrained with all accessible data, 3) fusing our pre-trained model from the second stage with contrastive learning(Khosla et al., 2021) to boost sensitivity to disparate topics, and 4) consolidating the results of several analogous models with different parameters through ensemble methods.

## 2 Related Work

### 2.1 Data Augmentations

Data augmentation(Feng et al., 2021) has recently attracted heightened interest within the Natural Language Processing (NLP) field due to developments in low-resource domains.

Rule-based strategies are straightforward to implement but typically result in incremental performance improvements. Wei and Zou (2019) proposes EDA, a set of token-level random perturbation operations including random insertion, deletion, and swap. Techniques that leverage trained models may entail higher implementation costs but introduce greater data variability, leading to substantial performance enhancements.

Model-based techniques tailored for downstream tasks can significantly impact performance. The popular method, back translation(Sennrich et al., 2016), translates a sequence into another language and then back into the original language. Kobayashi (2018)(contextual augmentation) feeds surrounding words to large model like BERT, RoBERTA(Liu et al., 2019) or XLNET(Yang et al., 2020) to inference the most suitable word.

In our research, we employ a combination of rule-based and model-based techniques to generate pseudo-labeled data from labeled data.

### 2.2 Sentence Representation and Self-supervised

The prevalent paradigm for most NLP research since 2018 entails a two-stage training process. Initially, a neural language model (LM), typically comprising millions of parameters, is trained on extensive unlabeled corpora through various pre-training tasks. Subsequently, the word representations acquired in the pre-trained model are repurposed during fine-tune for a downstream task. Several self-supervised pre-training tasks have been proposed to pre-train language models, such as Masked Language Modeling (MLM) (Devlin et al., 2019), and MAsked Sequence to Sequence pre-training (MASS) (Song et al., 2019). Sun et al. (2020) has proved that further pre-train BERT with masked language model tasks on the domain-specific data can improving the performance of the model.

In our research, we utilize all available data for the further pre-trained of the MacBERT model, which results in a more robust representation.

### 2.3 Contrastive Learning

Contrastive learning has proven its efficacy in learning robust representations, particularly within the natural language domain. In recent years, multiple studies have investigated the construction of sentence embeddings using contrastive learning. The fundamental concept of contrastive learning involves generating positive and negative sentence pairs, with the aim of drawing positive pair representations closer while distancing the negative ones.

Several strategies have been proposed to realize this objective. Fang et al. (2020) employs contrastive self-supervised learning at the sentence level with back-translation data augmentation. Gao et al. (2022) uses both unsupervised denoising objective and supervised natural language inference signals to learn sentence embeddings.

In our research, we introduce a contrastive loss function that encourages data with similar semantics to cluster together, while carefully avoiding the repulsion of false negatives.

| **Dataset** | **C** | **L** | $\overline{L}$ | $\hat{L}$ | $\overline{W_c}$ |
|---|---|---|---|---|---|
| Train | 900 | 45 | 2.95 | 59.06 | 1400 |
| Val | 100 | 37 | 2.61 | 7.05 | 1378 |
| Test | 238 | 42 | 2.81 | 15.95 | 1338 |
| Unlabeled | 1000 | - | - | - | 1410 |
| Pseudo | 2000 | 45 | 2.95 | 59.06 | 1396 |

Table 1: Details of the datasets. C: the amounts of the dataset; L: the numbers of labels; $\overline{L}$: average labels per instance; $\hat{L}$: average instances per label; $\overline{W_c}$: the average char per instance in content;

## 3 Dataset and Methods

The ESG dataset comprises columns such as title, content, and corresponding topic labels. The Chinese track training set includes 900 instances, the validation set includes 100 instances, and the test set encompasses 238 instances.

In addition to the labeled data, we have amassed 1000 instances of unlabeled data utilizing website crawlers. The distribution of this unlabeled data aligns with that of the labeled data.

Besides, we implement data augmentation methods such as EDA, back translation, and contextual augmentation yielding 2000 instances of pseudo-labeled data.

The distribution of the dataset is illustrated in Table 1.

### 3.1 MacBERT with Further Pre-trained

Given the impressive results BERT has achieved across various domains, we utilize the MacBERT model as the backbone of our model. However, while the MacBERT model is pre-trained on a general domain corpus, all training data derives from a specific domain's small corpus. Directly fine-tuning our BERT model could lead to overfitting. To mitigate this, we further pre-trained BERT-Chinese with masked language model tasks on all the labeled and unlabeled data.

Following this additional pre-trained, we input a sentence comprising m different tokens into BERT, extracting token embeddings from the last hidden layer as $[CLS, T1, T2, \cdots, T_m]$, where CLS is a special token denoting the start of the sentence for classification. The sentence representation is then obtained by applying mean-pooling to the token embeddings with a fixed length:

$$u = mean - pooling([CLS, T1, T2, \cdots, T_m]) \tag{1}$$

We place a binary classifier at top of the representation derived from the BERT model.

### 3.2 Constrastive Learning

We introduce a contrastive learning objective aimed at attracting similar instances and repelling disparate ones within the embedding space to achieve superior classification scores. For additional details, please refer to section 5.2.

In practice, we begin by encoding the instances with the further pre-trained model described earlier. Then, for a given instance $x_i$, all other instances in the batch sharing the same label $y_j$ with it constitute the positive sample set $S_j$. The set of positive samples under each label is denoted by $S = S_1, S_2, \cdots, S_q$, where q represents the topic number of instance $x_i$. We can then define the contrastive learning loss for each instance across the batch as

$$L_{cl} = \frac{-1}{q} \sum_{S_j \in S} \sum_{s \in S_j} \log \frac{func\left(E_i, E_s\right)}{\sum_{k \in I/\{i\}} func\left(E_i, E_k\right)} \tag{2}$$

$$func\left(u, v\right) = \exp\left(sim\left(u, v\right)/\tau\right) \tag{3}$$

where $E_i$ denotes the sentence representation, $sim(\cdot)$ indicates the cosine similarity function, $\tau$ is the contrastive learning temperature.

Besides, we combine the constrastive loss with cross-entropy and train them jointly. The overall training objective is calculated as follows:

$$L = \alpha \cdot L_{cl} + (1 - \alpha) \cdot L_{ce} \tag{4}$$

where $\alpha$ is a parameters which determined the importance of the contrastive loss.

### 3.3 Ensemble

We also construct an ensemble model using various sizes of MacBERT. Specifically, we train two instances of MacBERT-Large and two instances of MacBERT-Base, each with a different seed. We amalgamate all the models' predictions by averaging their probabilities, thereby enhancing the overall accuracy of the prediction.

## 4 Experiments

### 4.1 Training Setup

We adopt MacBERT-Large and MacBERT-Base models as our backbone model. For self-supervised pre-training, we employ all the labeled and unlabeled data with a batch size of 32 across 25 epochs, implementing early-stopping validated with a patience of 100 steps. The pre-training learning rate for all models is set to $1e - 5$.

| Models | Micro-F1 | Macro-F1 |
|---|---|---|
| Base | 0.389 | 0.173 |
| Large | 0.407 | 0.178 |
| Ensemble | 0.412 | 0.181 |

Table 2: Performance of all the models on the validation set.

When fine-tuning with constrative learning, we utilize all the labeled and pseudo-labeled data with a batch size of 16 for 20 epochs. The learning rate for the BERT-Chinese-Large model is set to $5e - 5$, and for the MacBERT-Base model, it's set to $4e - 5$. All models are trained across 15 epochs.

### 4.2 Results

Table 2 shows the appearance on the validation set. The table shows that the MacBERT-Large model with further pretrained performs the best on the validation set for single model with an Micro-F1 score of 0.407. The last submitted ensemble models achieve an Micro-F1 score of 0.412 on the validation set, while achieve 0.3914 on the test set. Unfortunately, due to time constraints, we were unable to record additional results on the test set.

## 5 Analysis

### 5.1 Effect of Data Augmentaion Methods

We experimented with three different data augmentation methods: (1) Easy Data Augmentation (EDA); (2) Back-Translation (BT); and (3) Contextual Augmentation (CA). These experiments were built upon the further pre-trained BERT-Chinese-Base model, with the augmented data utilized for contrastive learning.

The results, displayed in Table 3, show that among the single data augmentation methods, CA yielded the highest improvement in model performance, achieving a Micro-F1 score of 0.384. Among the combined augmentation methods, CA and BT had the most significant impact on model performance, securing an increase of 0.389. As a result, we ultimately selected a combination of CA and BT for data augmentation.

We delved into the differences between these three methods and discovered a potential reason for the ineffective EDA data augmentation scheme. It appeared that the key tokens edited by the method

| Models | Micro-F1 | Macro-F1 |
|---|---|---|
| EDA | 0.378 | 0.164 |
| BT | 0.383 | 0.168 |
| CA | 0.384 | 0.169 |
| EDA+BT | 0.381 | 0.167 |
| EDA+CA | 0.382 | 0.167 |
| BT+CA | 0.389 | 0.173 |
| EDA+BT+CA | 0.386 | 0.171 |

Table 3: Performance of all the data augmentaion methods on the validation set.

was not relevant to the topic label corresponding to the original sentence, or some key words were omitted, leading to incorrect annotation. Here are some examples:

Original content: "···但随著全球零排放航空旅行的兴趣增加，···", the related topic label is "E01 - 气候变化|碳排放量(Carbon Emissions)".

BT content: "···但当全球零排放航空旅行的兴趣增加，···"

MG content: "···但随著全球零排放旅游的兴趣增加，···"

EDA content: "···但随著全球航空旅行的增加兴趣，···"

We can observe that the lack of token "零排放" has resulted in a disconnection between sentence semantics and their corresponding topic.

### 5.2 Effect of Contrastive Learning

We explored three implementations of contrastive learning to determine the most effective method in MacBERT-Base model. For our analysis, let's consider a batch composed of K samples, denoted as $Batch = (X_1, Y_1), (X_2, Y_2), \cdots, (X_K, Y_K)$. For a given sample i, where $X_i$ represents a text sequence and its topic label set is denoted as $Y_i$, the model's encoding provides us the sentence representation $E_i$ and the topic probability $Q_i$ of $X_i$. Here, $Q_i = Q_{i1}, Q_{i2}, \cdots, Q_{iL}$, with L representing the total number of topic labels.

We represent $Y_i$ as the one hot encoding of the label, defined as $Y_i = y_1, y_2, \cdots, y_L$. For a given i-th topic label $y_i \in 0, 1$, $y_i = 0$ signifies the absence of this type of label in the text, while $y_i = 1$ implies its presence.

We tested three implementations of contrastive learning:

(1) strictly contrastive learning(SCL) This approach mandates that a sample can serve as a positive contrastive sample of the anchor point only when their label sets exactly match. SCL is rigorous and does not consider samples that partially overlap with the anchor label set.

| Models | Micro-F1 | Macro-F1 |
|---|---|---|
| SCL | 0.377 | 0.168 |
| JSCL | 0.385 | 0.171 |
| SLCL | 0.389 | 0.173 |

Table 4: Performance of different contrastive learning methods on the validation set.

(2) Jaccard Similarity Contrastive Loss (JSCL)(Li et al., 2022): This method works on samples to varying degrees based on the similarity of their labels. For a given sample, JSCL draws samples with the exact same label as closely as possible, while only slightly pulling in samples that share some labels.

(3) Stepwise Label Contrastive Loss (SLCL): While the previous two methods primarily consider multiple emotions simultaneously, SLCL considers different labels separately, computes the contrast loss independently, and then combines each emotion's losses.

As the result shown in Table 4, SLCL achieve the best score and we choose this method as our contrastive learning method.

## 6 Conclusion

In this paper, we discussed the methodologies employed for the multi-lingual ESG issue identification (ML-ESG) shared task at FinNLP 2023. Our team's proposed MacBERT model, equipped with further pre-trained and contrastive learning strategies, achieved the highest ranking in the Chinese track. Our experimental results underscored the efficiency of further self-supervised pre-training and contrastive learning approaches. Comprehensive experiments confirmed our method's efficacy and helped discern the aspects contributing to our performance enhancements.

## 7 Limitations

Despite our promising results, our study was limited by time and resource constraints. Consequently, we could not undertake semi-supervised experiments and few-shot learning experiments. These methodologies present intriguing prospects for future exploration.

## References

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 (FinNLP) and the Second Multimodal AI For Financial Forecasting (Muffin)*.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. 2020. Revisiting pre-trained models for Chinese natural language processing. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 657–668, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Hongchao Fang, Sicheng Wang, Meng Zhou, Jiayuan Ding, and Pengtao Xie. 2020. Cert: Contrastive self-supervised learning for language understanding.

Steven Y. Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard Hovy. 2021. A survey of data augmentation approaches for nlp.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2022. Simcse: Simple contrastive learning of sentence embeddings.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2021. Supervised contrastive learning.

Sosuke Kobayashi. 2018. Contextual augmentation: Data augmentation by words with paradigmatic relations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 452–457, New Orleans, Louisiana. Association for Computational Linguistics.

Junjie Li, Yixin Zhang, Zilei Wang, and Keyu Tu. 2022. Probabilistic contrastive learning for domain adaptation.

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.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2019. Mass: Masked sequence to sequence pre-training for language generation.

Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2020. How to fine-tune bert for text classification?

Jason Wei and Kai Zou. 2019. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6382–6388, Hong Kong, China. Association for Computational Linguistics.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2020. Xlnet: Generalized autoregressive pretraining for language understanding.