Leveraging Semi-Supervised Learning on a Financial-Specialized Pre-trained Language Model for Multilingual ESG Impact Duration and Type Classification

Jungdae Kim, Eunkwang Jeon, Sanghyun Jeon
KakaoBank Corp.
{jd.kim, weezy.j, ali.jeon}@kakaobank.com

Abstract

This paper presents the results of our participation in the Multilingual ESG Impact Duration Inference (ML-ESG-3) shared task organized by FinNLP-KDF@LREC-COLING-2024. The objective of this challenge is to leverage natural language processing (NLP) techniques to identify the impact duration or impact type of events that may affect a company based on news articles written in various languages. Our approach employs semi-supervised learning methods on a finance-specialized pre-trained language model. Our methodology demonstrates strong performance, achieving 1st place in the Korean - Impact Type subtask and 2nd place in the Korean - Impact Duration subtask. These results showcase the efficacy of our approach in detecting ESG-related issues from news articles. Our research shows the potential to improve existing ESG ratings by quickly reflecting the latest events of companies.

Keywords: ESG, ESG Rating, NLP, SSL

1. Introduction

The importance of Environmental, Social, and Governance (ESG) factors in the investment decision-making process has been increasingly emphasized. ESG factors have emerged as key considerations for corporate sustainability and long-term success, leading to the proposal of various frameworks and approaches to evaluate and quantify companies ESG-related activities. However, existing ESG evaluation methods primarily rely on fixed materials such as annual reports, limiting their ability to promptly reflect the dynamic changes in the market. In this context, an approach has been proposed to infer the impact of the latest events and news articles on companies ESG ratings (Tseng et al., 2023; Kannan and Seki, 2023). Tseng et al. (2023) introduced a new dataset that can identify the ESG impact type and impact duration of corporate events using ESG-related news articles. This dataset has become an important foundation for the Multi-Lingual ESG Impact Duration Inference (ML-ESG-3) shared task proposed at FinNLP-KDF@LREC-COLING-2024. The goal of the ML-ESG-3 shared task is to identify the impact duration or impact type of events that may affect companies using natural language processing (NLP) techniques on news articles written in various languages.

To achieve this goal, we utilized a finance-specialized pre-trained language model and applied semi-supervised learning (SSL) methods using unlabeled data collected through web crawling. This approach achieved 1st and 2nd place in the Korean impact type and impact duration identification tasks, respectively. As part of the research exploring the modernization and dynamic update possibilities of ESG evaluation, this paper presents an NLP-based methodology that can improve ESG evaluation by promptly reflecting the latest corporate events. This is expected to enable investors to make investment decisions considering ESG factors based on more accurate and timely information.

2. Dataset

The Korean task consists of two sub-tasks: Impact Type Identification and Impact Duration Inference. The datasets for these sub-tasks were annotated following the methodology proposed by Tseng et al. (2023).

Impact Type identification is a single-choice question that aims to determine the type of impact a news article might have on a company. The possible labels are "opportunity", "risk", and "cannot distinguish". The "opportunity" label indicates that the news article discusses a potential positive impact or benefit to the company, while the "risk" label suggests that the article highlights a potential negative impact or threat. The "cannot distinguish" label is assigned when the impact type is unclear.

Impact Duration inference is a single-choice question that seeks to determine the duration of the impact a news article might have on a company. Based on the distinction between short-term and long-term, three labels are presented: “less than 2 years”, “2 to 5 years”, and “more than 5 years”. These labels provide a temporal context for the impact, allowing for a better understanding of the Immediate and long-term implications of the news content on the company.

The news articles In the dataset vary In length, with an average of 733 characters per article. The shortest article has 173 characters, while the longest article has 1,768 characters. This variation in article length presents a challenge for the models, as they need to effectively understand texts of different sizes.

Proceedings of the Joint Workshop of the 7th FinNLP, the 5th KDF, and the 4th ECONLP, pages 269–273

To provide a clear understanding of the dataset composition, Tables 1 and 2 show the distribution of labels for the Impact Type and Impact Duration sub-tasks, respectively, within the training set. For both sub-tasks, the dataset provides a train set containing 800 examples and a test set with 200 examples.

Labels	Count
opportunity	462
risk	229
cannot distinguish	109
Total	**800**

Table 1: Label counts in Korean – Impact Type train set.

Labels	Count
less than 2 years	446
2 to 5 years	142
more than 5 years	212
Total	**800**

Table 2: Label counts in Korean – Impact Duration train set.

3. Methods

We first designate a model that has been fine-tuned using supervised learning with KF-DeBERTa (Jeon et al., 2023), a Korean language model specialized for the financial domain, as our baseline model. Subsequently, to improve performance compared to the baseline model, we collect additional ESG-related news articles from the web and conduct semi-supervised learning using the collected data.

3.1 Finance-specialized Pre-trained Language Model

KF-DeBERTa (Jeon et al., 2023) is trained on a large-scale Korean financial corpus and follows the architecture and methods of DeBERTa (He et al., 2020). KF-DeBERTa is suitable for ESG-related tasks because it showed state-of-the-art performance in most evaluations of general and financial domains. In particular, the DeBERTa architecture has a significant advantage in understanding long sequences like in this dataset because it uses relative position embeddings, compared to BERT (Devlin et al., 2018) architecture models that use absolute position embeddings. To take advantage of this, we used the number of max position embeddings used for relative position embedding allocation as a hyperparameter during fine-tuning. Table 3 shows the performance of the validation set of Korean - Impact Type according to the number of max position embeddings. We chose 1,792 as the max position embeddings to be used for all future experiments.

Max Position Embeddings	Micro-F1	Macro-F1
512	0.8197	0.7417
768	0.8279	0.7553
1024	**0.8361**	0.7466
1280	0.8279	0.7613
1536	**0.8361**	0.7555
1792	**0.8361**	0.7814
2048	0.8179	**0.7881**

Table 3: Effects of max position embeddings on performance in Korean – Impact Type validation set.

3.2 Semi-supervised Learning

Semi-supervised learning has been shown to be effective in improving model performance when labeled data is scarce (Tarvainen and Valpola, 2017; Bertheolot et al., 2019; Xie et al., 2020; Shon et al., 2020). In the case of this task, we believed that semi-supervised learning utilizing unlabeled data would be effective since the number of labeled data is only 800. We collected 2,916 unlabeled data by crawling ESG-related news articles from the web and applied the ideas of UDA (Xie et al., 2020) and FixMatch (Shon et al., 2020), which are consistency training-based semi-supervised learning methods. Consistency training methods regularize model predictions to be invariant to noise injected into input examples or hidden states. UDA utilizes high-quality augmentation methods that have traditionally been effective in supervised learning as noise to be injected into unlabeled data. In each iteration, UDA calculates the supervised loss for a mini-batch of labeled data and the consistency loss for a mini-batch of unlabeled data using the model prediction of the unlabeled example as a soft pseudo-label for the augmented unlabeled example. It then calculates the final loss by summing the two losses. Generally, a larger batch size is used for consistency loss than for supervised loss.

FixMatch employs both weak and strong augmentation techniques for processing unlabeled data. Weak augmentation is applied to unlabeled examples to create hard pseudo-labels, and strong augmentation is applied to unlabeled examples to create model predictions and calculate consistency loss.

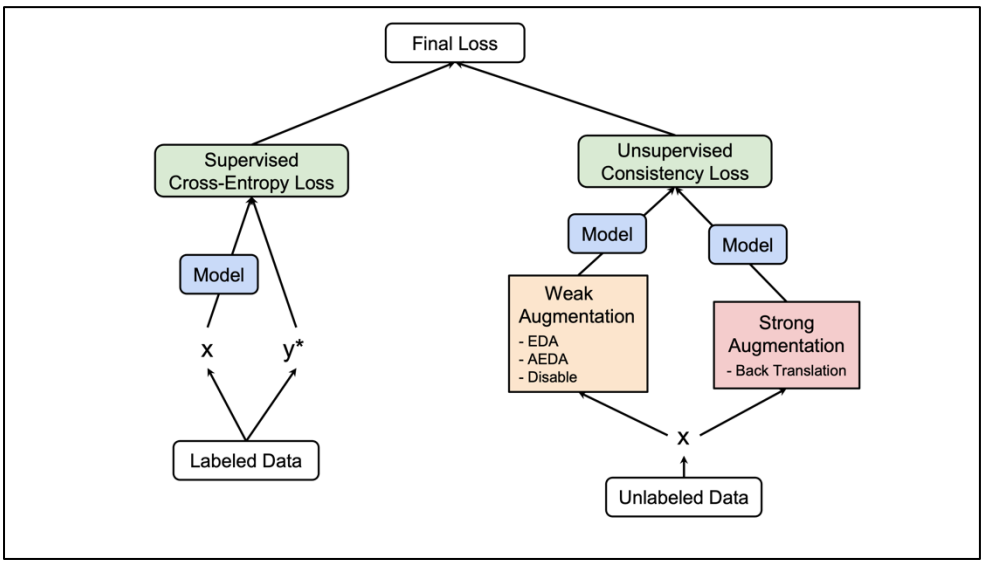

Figure 1: The entire process of the semi-supervised learning we used.

We chose the idea of using both weak augmentation and strong augmentation from FixMatch for augmentation diversity and the idea of using soft pseudo-labels from UDA to mitigate the model's overconfidence in unlabeled data. We used EDA (Wei and Zou, 2019) and AEDA (Karimi et al., 2021) for weak augmentation and also considered not using weak augmentation. When weak augmentation is not used, it is the same as UDA. EDA augmentation applies Synonym Replacement (SR), Random Insertion (RI), Random Swap (RS), and Random Deletion (RD) to some of the words in a sentence. We only used SR and RS for augmentation in EDA, as they were empirically suitable for Korean data. AEDA augmentation randomly selects some of all positions between words in a sentence and inserts one of the six punctuation marks {".", ";", "?", ":", "!", ","} randomly selected at each position. We also used back translation for strong augmentation, where we first translated the Korean unlabeled data into English using machine translation and then back into Korean.

To summarize our method, we calculate the supervised loss using labeled data, create soft pseudo-labels by applying weak augmentation to unlabeled data, and calculate consistency loss by applying strong augmentation to create model predictions. We then calculate the final loss by summing the two losses. Figure 1 shows the entire process of the semi-supervised learning we used. The loss used for training can be formulated as follows:

$$L = L_s + L_c \quad (1)$$

$$L_s = \frac{1}{B}\sum_{i=1}^{B} CE\left(y^*, p_\theta\left(y|x_i^l\right)\right) \quad (2)$$

$$L_c = \frac{1}{\mu B}\sum_{i=1}^{\mu B} CE\left(p_{\tilde{\theta}}\left(y|\alpha(x_i^u)\right), p_\theta\left(y|\mathcal{A}(x_i^u)\right)\right) \quad (3)$$

where L is the total loss, L_s is the supervised loss, L_c is the consistency loss, $p_\theta\left(y|x_i^l\right)$ is the model's predicted probability distribution for the target given the i-th labeled example x_i^l, $\tilde{\theta}$ is a fixed copy of the current parameters θ indicating that the gradient is not propagated through $\tilde{\theta}$, x_i^u is the i-th unlabeled example, B is the batch size of labeled data, μ is the ratio of unlabeled data to labeled data, μ is the multiplier used to determine the batch size of unlabeled data μB by multiplying it with the batch size of labeled data B. CE is the cross-entropy loss function, y^* is the one-hot encoded label for labeled example, α is the weak augmentation function, $\mathcal{A}$ is the strong augment-ation function.

Table 4 shows the performance on the Korean-Impact Type validation set for each configuration. The batch size of the unlabeled data was most effective when it was 4 to 5 times the batch size of the labeled data. In the weak augmentation setting, AEDA led to decreased performance.

μ	weak aug.	strong aug.	Micro-F1	Macro-F1
4	-	BT	0.8361	**0.7901**
4	EDA	BT	**0.8443**	0.7525
4	AEDA	BT	0.8279	0.7506
5	-	BT	**0.8443**	0.7603

Table 4: Performance on the Korean - Impact Type validation set by augmentation methods. "BT" stands for Back Translation, and "aug." is short for

augmentation. μ is the multiplier used to determine the batch size of unlabeled data μB by multiplying it with the batch size of labeled data B.

4. Experiments

4.1 Training Setup

We used 120 samples out of the 800 samples in the train set as a validation set. For training, we used the AdamW optimizer (Loshchilov and Hutter, 2017) with a linear learning rate schedule having a warmup of 100 steps and an initial learning rate of 2.5×10^{-5}. Batch size was set to 4, weight decay to 0.01, and gradient clipping to 1.0. We conducted training for 5 to 12 epochs and also utilized the exponential moving average (EMA) of weights with decay rates of 0.99 and 0.999.

4.2 Results

We evaluated the Korean - Impact Type dataset and the Korean - Impact Duration dataset using the Micro-F1 and Macro-F1 performance metrics.

Our SSL method worked well on the Korean - Impact Type dataset. The model trained with SSL showed improved Micro-F1 and Macro-F1 performance on the validation set compared to the supervised learning baseline model. On the other hand, the model with EMA applied did not show performance improvement compared to the baseline. We submitted the baseline model and two SSL models based on validation set performance. In the final results, one of the SSL models achieved 1st place with Test Micro-F1 of 0.8400 and Test Macro-F1 of 0.7985. Table 5 shows the experimental results on the Korean - Impact Type dataset.

The EMA technique was effective on the Korean - Impact Duration dataset. EMA is a technique that calculates the exponential moving average of model weights to reduce noise and decrease variability, thereby stabilizing the learning process (Izmailov et al., 2018). It helps prevent overfitting and improves generalization performance. The model with EMA applied showed improved Micro-F1 performance on the validation set compared to the supervised learning baseline model, and some models also showed improved Macro-F1 performance. In contrast, the model trained with SSL did not show performance improvement over the baseline. We submitted three EMA models based on validation set performance. In the final results, one of the EMA models achieved 2nd place with Test Micro-F1 of 0.6750 and Test Macro-F1 of 0.6198. Table 6 shows the experimental results on the Korean - Impact Duration dataset.

Model	Valid. Micro-F1	Valid. Macro-F1	Test Micro-F1	Test Macro-F1
baseline	0.8361	0.7814	0.8050	0.7343
EMA	0.8279	0.7522	-	-
SSL #1	0.8361	**0.7901**	0.8150	0.7398
SSL #2	**0.8443**	0.7603	**0.8400**	**0.7985**

Table 5: Experimental results in Korean - Impact Type.

Model	Valid. Micro-F1	Valid. Macro-F1	Test Micro-F1	Test Macro-F1
baseline	0.7869	0.7438	-	-
EMA #1	**0.7951**	0.7579	**0.6750**	**0.6198**
EMA #2	**0.7951**	**0.7608**	0.6650	0.6102
EMA #3	**0.7951**	0.7339	**0.6750**	0.6154
SSL	0.7705	0.7164	-	-

Table 6: Experimental results in Korean - Impact Duration.

5. Conclusion

In this paper, we presented our approach and results for the Multilingual ESG Impact Duration Inference (ML-ESG-3) shared task at FinNLP-KDF@LREC-COLING-2024. Our methodology, which employed semi-supervised learning and exponential moving average of weights on a finance-specialized pre-trained language model, demonstrated strong performance in the Korean - Impact Type and Korean - Impact Duration subtasks. Our model achieved 1st place in the Korean - Impact Type subtask and the 2nd place in the Korean - Impact Duration subtask. These results highlight the potential of our methodology in identifying ESG-related issues from news articles.

6. Bibliographical References

Tseng, Y. M., Chen, C. C., Huang, H. H., & Chen, H. H. (2023, October). DynamicESG: A Dataset for Dynamically Unearthing ESG Ratings from News Articles. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management* (pp. 5412-5416).

Jeon, E., Kim, J., Song, M., & Ryu, J. (2023). KF-DeBERTa: Financial Domain-specific Pre-trained Language Model. In *Proceedings of the 35th Annual Conference on Human and Cognitive Language Technology* (pp. 143-148).

He, P., Liu, X., Gao, J., & Chen, W. (2020). Deberta: Decoding-enhanced bert with disentangled

attention. *arXiv preprint arXiv:2006.03654*.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Xie, Q., Dai, Z., Hovy, E., Luong, T., & Le, Q. (2020). Unsupervised data augmentation for consistency training. *Advances in neural information processing systems*, *33*, 6256-6268.

Sohn, K., Berthelot, D., Carlini, N., Zhang, Z., Zhang, H., Raffel, C. A., ... & Li, C. L. (2020). Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in neural information processing systems*, *33*, 596-608.

Tarvainen, A., & Valpola, H. (2017). Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, *30*.

Wei, J., & Zou, K. (2019). Eda: Easy data augmentation techniques for boosting performance on text classification tasks. *arXiv preprint arXiv:1901.11196*.

Karimi, A., Rossi, L., & Prati, A. (2021). AEDA: an easier data augmentation technique for text classification. *arXiv preprint arXiv:2108.13230*.

Loshchilov, I., & Hutter, F. (2017). Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Izmailov, P., Podoprikhin, D., Garipov, T., Vetrov, D., & Wilson, A. G. (2018). Averaging weights leads to wider optima and better generalization. *arXiv preprint arXiv:1803.05407*.

Kannan, N., & Seki, Y. (2023). Textual Evidence Extraction for ESG Scores. In *Proceedings of the Fifth Workshop on Financial Technology and Natural Language Processing and the Second Multimodal AI For Financial Forecasting* (pp. 45-54).

Berthelot, D., Carlini, N., Goodfellow, I., Papernot, N., Oliver, A., & Raffel, C. A. (2019). Mixmatch: A holistic approach to semi-supervised learning. *Advances in neural information processing systems*, *32*.