QM-AI at SemEval-2025 Task 6: an Ensemble of BERT Models for Promise Identification in ESG Context

Zihang Sun and **Filip Sobczak**
QualityMinds GmbH, Munich, Germany
{zihang.sun, filip.sobczak}@qualityminds.de

Abstract

This paper presents our approach and findings in the SemEval-2025 Task 6: Multinational, Multilingual, Multi-industry Promise Verification (PromiseEval), which focuses on verifying promises in the industrial Environmental, Social, and Governance (ESG) reports. Specifically, we participate in the first subtask of the PromiseEval shared task, promise identification. We tackle this subtask by building an ensemble of four BERT models trained in different experimental configurations, and deploying logistic regression as meta-model. Each configuration has a different combination of two variables: whether augmented data is used, and whether English translation is used. We find out that the BERT model trained without augmented data or English translation not only has the best evaluation results on the test data in most languages, but also has higher robustness than the meta-model. We submitted results from the meta-model to the leaderboard, and rank the first place in Japanese and Korean, the second place in French and Chinese, and the seventh place in English.

1 Introduction

As the public's emphasis on the environment protection and the importance of Environmental, Social, and Governance (ESG) aspects of industries grow, a strong implementation of ESG framework creates value for companies in various ways (Barangă and Țanea, 2022). In order to gain such benefits without paying the cost, it is revealed that companies with environmental violations try to appear more environmentally friendly by producing more frequent, abundant reports with less readability to deflect reader's attention from their violations (Gorovaia and Makrominas, 2025). To verify the promises made in the industrial ESG reports, Seki et al. (2024) proposes a four-step approach: 1. identifying promise 2. linking supporting evidence to the promise, 3. assessing clarity of the promise-evidence pair, and 4. inferring timing for verifying the promise. These four steps correspond to the four subtasks in the SemEval-2025 Task 6: Multinational, Multilingual, Multi-industry Promise Verification (PromiseEval) (Chen et al., 2025). In this paper, we describe our submission to the first subtask of the SemEval-2025 Task 6: promise identification.

The subtask 1 is a multi-lingual binary classification task. The used dataset is the proposed multilingual dataset, ML-Promise, that includes ESG reports from various industries in English, French, Japanese, Korean, and Chinese (Chen et al., 2025). Each instance in the ML-Promise dataset contains the origin of a PDF and a page number. Additionally, a text snippet is included in each instance for the English, French, and Japanese languages. These text snippet can be used directly as input for classification. For the Korean and Chinese languages, the participants need to either extract the text from the whole page for classification or use the given page of the PDF as direct input. The output of the task 1 is a boolean label indicating whether the input contains any promise.

We approach this task by building an ensemble of BERT models (Devlin et al., 2019). Each BERT model is exposed to a different experiment configuration. Our hypothesis is that with each BERT model trained differently, the robustness of our ensemble will improve. After evaluating all base BERT models and the meta-model on the test data, we find out that the BERT model trained with original data without data augmentation or English translation has the highest performance and robustness than the other BERT model and the meta-model.

2 Background

BERT models have been proven to have promising performance in other ESG related tasks. In the

Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 228–233

Multi-lingual ESG Issue Identification shared task, where multi-lingual ESG news articles are to be classified into 35 key ESG issues, BERT-like language models with data augmentations by LLMs have leading performance in all languages (Chen et al., 2023a). In the Multi-lingual ESG Impact Type Identification shared task, a classification task with three classes, fine-tuned RoBERTa model (Liu et al., 2019) is proven to have the best results in English and French, while Fin-BERT model (Araci, 2019) with English translation achieves the highest performance in Chinese (Winatmoko and Septiandri, 2023; Vardhan et al., 2023; Chen et al., 2023b). In the Multi-Lingual ESG Impact Duration Inference shared task, a classification task with three classes, DeBERTa-v3 (He et al., 2021) is the best-performing model for English, while for Korean and Japanese, the XLM-RoBERTa model (Conneau et al., 2020) is among the best models (Chen et al., 2024). Based on the findings in the previous tasks, we introduce English translation as a variable into our experimental configurations. Furthermore, we pick the DeBERTa-v3 model to be the model receiving English translation, and the XLM-RoBERTa model to receive the original multi-lingual input. More details regarding our classification system and experimental setup is presented in Section 3 and 4.

The PromiseEval shared task has one separate leaderboard for each language. The leaderboard only accepts submission files containing results for all four subtasks. Although our focus is only to solve the subtask 1, we utilize the GPT-4o-mini to gain labels for the other three subtasks in order to submit our results to the leaderboard. We present our approach with GPT-4o-mini to generate predictions in Section 4. The leaderboard scores are aggregated over labels from all subtasks. This means that the leaderboard scores do not reflect the ranking and performance of participant's system in one individual subtask. To showcase our system's performance in subtask 1, we evaluate our system on the test data and show the results in Section 5.

3 Methods

3.1 Data Augmentation

We divide the given training dataset into a training split and a development split, maintaining an 8:2 ratio. To tackle class imbalance in the training split and boost the number of training samples, we expand the training split by augmenting data from the original PDFs. The development split is left unaltered to ensure that our system is evaluated using the actual data distribution during training time.

We expand the training split by drawing unused pages or sentences from the PDFs within the training split. Specifically, we extract sentences for the English, French, and Japanese languages and sample pages for Chinese and Korean. This approach ensures the augmented data mirrors the original data entries in length. The amount of augmented data for each language is determined by the difference between the sample sizes of the positive and negative classes in the respective language. Following this, OpenAI's GPT-4o model is utilized to assign labels to the sampled data. As shown in Table 1, the class imbalance issue still exists after data augmentation, but is lessened, especially for the total count of classes over all languages.

3.2 The Classification System

Our classification system is an ensemble system consisting of four BERT base models and a logistic regression model as meta-model.

We train two XLM-roberta-large models using the original multi-lingual input text. One model uses the multi-lingual augmented training split and the other uses the training split without augmented data. Both models are evaluated on the same development split. The goal is to let one model learn in quantity, i.e. with augmented data, while the other model should not be influenced by the potential noises introduced by augmented data.

To tackle multi-lingual nature of data and its varying class imbalance in different languages as shown in Table 1, we translate all non-English texts into English using google translate API[1]. Similar to multi-lingual setting, we train two Deberta-v3-large models (He et al., 2021) respectively on the translated augmented training split and the translated training split without augmentation.

We use a logistic regression model as the meta-model to produce the final prediction. Except from the predictions from the four base BERT models, we also provide the meta-model with the base models' probabilities for the predictions as well as two pieces of meta data: the language of the original input text and a boolean value about whether the input text is augmented. The prediction from the meta-model is the final output and submitted to the

[1]Version: 4.0.2. URL: https://pypi.org/project/googletrans/

	English	French	Japanese	Korean	Chinese	Sum
Before	71 / 251	68 / 247	29 / 295	92 / 290	211 / 110	471 / 1193
After	261 / 287	248 / 305	278 / 358	197 / 443	282 / 232	1266 / 1625

Table 1: A comparison table of the count of positive and negative classes for each language in training split. In each diagonal box, the left number stands for the number of instances in positive class, while the right number is for the negative class. A class is positive when the promise status is True. The first row shows the comparison *before* data augmentation. The second row shows that *after* data augmentation

evaluation website.

4 Experimental setup

In the process of data augmentation, we exclude pages or sentences containing fewer than 16 words to prevent the dataset from being populated with simple word phrases or page headings. The training datasets for English, French, and Japanese languages include the text for classification directly. In contrast, the Chinese and Korean datasets only list the page number and PDF source. Hence, we utilize the PyPDF2 library[2] to retrieve text from specified pages of the Chinese and Korean PDFs, serving as the input for the BERT models.

There are two experimental variables: with augmented data or without, and either using multilingual or monolingual text input (English translation), yielding a total of four configurations. As described in Section 3.2, one BERT model is trained for each configuration. When a model is trained using monolingual data, it will likewise be evaluated on the English translations of the input text in both the development and test datasets. All BERT models share the same hyper-parameters: they are fine-tuned with a batch size of 16, and a learning rate of 6e-6 over 20 epochs. The model with the best macro F1 score is selected as the checkpoint.

Each of the four fine-tuned BERT models is evaluated on the development data split, and the resulting labels and probabilities are recorded. A logistic regression model, built using the Scikit-learn library[3], is then trained on the full development data split, including BERT models' outputs and additional metadata as detailed in Section 3.2. The positive promise status is given the label id 0 and the negative promise status is 1. The trained logistic regression model is our meta-model. Finally, we deploy our fine-tuned BERT models on the test dataset, and use the meta-model on the BERT models' outputs and other meta data to generate final predictions for the test data.

To generate labels for subtasks 2-4, we used the GPT-4o-mini model with retrieval augmented generation as a classifier. In all cases, we set the model's temperature value to 1.0. First, we encoded all provided data points with the OpenAI text-embedding-3-small embedding model. For each subtask, we devised a system prompt that described the problem and the expected model output. For each test sample, we retrieved the three most similar examples from the training data, included them in the sample-specific prompt, and generated the output label.

The Chinese and Japanese datasets included two additional subtasks. For data points containing a promise or evidence, we needed to extract the corresponding string that included the promise or evidence text. To do this, we split the input text of each sample into a list of sentences and then individually classified each sentence to determine whether it contained a promise or evidence. Lastly, we concatenated the relevant sentences for the final output.

5 Results

Table 2 shows the results from all base BERT models and the meta-model on the test data performing promise status classification task. The XLM-roberta-large model trained in the multi-lingual setting with original data (multi_ori) yields the best results in three languages. The meta-model has the best performance in Korean, while the augmented data helps the XLM-roberta-large model to achieve best result in Chinese in the multi-lingual setting. Though the ensemble model does not have the best performance for the most languages, it exhibits constantly above average results compared to the base models, and shows comparable robustness to the best-performing base model in *multi_ori* setup.

Examining the results from using augmented

[2]Version: 3.0.1 URL: https://pypi.org/project/PyPDF2/

[3]Version: 1.3.1, URL: http://scikit-learn.org

	English	French	Japanese	Korean	Chinese
mono_aug	0.7288	0.7010	0.6496	0.7348	0.6659
mono_ori	0.7595	0.7748	0.6436	0.7667	0.6547
multi_aug	0.7546	0.6823	0.6771	0.7742	**0.7010**
multi_ori	**0.7921**	**0.7864**	**0.7368**	0.7727	0.6802
meta-model	0.7767	0.7566	0.6631	**0.7839**	0.6821

Table 2: The table presents the Macro F1 scores of four BERT models and the meta-model on the test dataset performing promise status classification task. The term *mono* refers to the *monolingual* setup, in which all languages are translated into English, whereas *multi* denotes the *multi-lingual* setup, where the original languages of the texts remain unchanged. *Aug* and *ori* represent the *augmented* and *original* configurations, respectively. In the augmented setting, both the augmented and original data are utilized, whereas the original configuration relies solely on the labeled data in the provided dataset. A logistic regression model is employed as the meta-model.

data versus solely utilizing original data without augmentation reveals that augmented data consistently enhances the model's performance for Chinese. This improvement might be attributed to the fact that the original training data is significantly biased towards the positive class, with Chinese being the only language exhibiting a class imbalance favoring the negative class, as described in Table 1. A similar data distribution pattern is observed in the test data. The augmented data helps to dampen the positive class bias, thereby enhancing the model's performance for Chinese. This proves that data augmentation is beneficial in reducing the impact of class biases.

When comparing the outcomes of using English translations versus not using them, it is noticeable that in most cases, experiments with monolingual text perform worse than those with multi-lingual text. This suggests that translating multi-lingual content into a single language might not enhance the model's learning capabilities. However, This difference could also stem from the variation in model selection between multi-lingual and monolingual contexts. In the future, a more detailed investigation could be conducted using the same BERT model across all experimental conditions to examine the helpfulness of English translations.

Utilizing logistic regression as our meta-model allows us to assess how each feature contributes to the final output through its coefficients. Table 3 illustrates that all models' predictions positively influence the logistic regression model's result, with the *multi_ori* experimental setup having the greatest impact. This aligns with our findings in Table 2, where the *multi_ori* model demonstrates superior or competitive performance in all languages. Additionally, Table 3 reveals that logistic regression

	prediction	probability
mono_aug	1.232	-0.507
mono_ori	1.391	-0.586
multi_aug	0.535	-0.079
multi_ori	**1.436**	-0.09

Table 3: The table shows the logistic regression model's coefficients for base BERT model's prediction and corresponding probability. The coefficients are rounded to 3 decimal places.

assigns negative coefficients to the base models' *probabilities* for their predictions, thereby penalizing their confidence in their predictions. This suggests that greater confidence from a base model results in less trust from the meta-model. Moreover, this skepticism towards the confidence of base models is less pronounced in multi-lingual contexts compared to monolingual contexts, further corroborating our other observation that monolingual settings underperform relative to multi-lingual settings, and therefore, the meta-model places less trust in them.

In addition to the base model's predictions and probabilities, we incorporate two metadata variables into the logistic regression: the language of the data instance and whether it is augmented. Both of these variables exert minimal to no influence on meta-model's output. The coefficient for the *language* variable is 0.03. Being a logistic regression model, our meta-model treats each variable independently, and thus, the language information contributes little to the final result. Similarly, as anticipated, the coefficient of the *augmented* variable rounds to 0 when rounded to three decimal places. As the meta-model is trained using the development data split that consists solely of original data,

this variable lacks any decision-making authority.

6 Conclusion

This paper describes our contribution to the first subtask of SemEval-2025 Task 6, promise identification. We deploy an ensemble of four BERT models trained in different experimental configurations and use logistic regression as meta-model. Our results show that the BERT model trained without augmented data or English translation has the best performance in most languages.

For future work, one can try out other meta-models that take relations between base model predictions and meta data into account. On the base model side, we can use multi-modal models to take PDF page directly as input to improve the varieties of base models. Furthermore, we believe a more sophisticated data cleaning pipeline for extracted text from PDFs can also potentially improve the base BERT model's performance.

References

Dogu Araci. 2019. Finbert: Financial sentiment analysis with pre-trained language models. *Preprint*, arXiv:1908.10063.

Laurențiu Paul Barangă and Elena-Ioana Țanea. 2022. Introducing the esg reporting–benefits and challenges. *Journal of Financial Studies*, 7(13):174–181.

Chung-Chi Chen, Yohei Seki, Hakusen Shu, Anaïs Lhuissier, Juyeon Kang, Hanwool Lee, Min-Yuh Day, and Hiroya Takamura. 2025. SemEval-2025 task 6: Multinational, multilingual, multi-industry promise verification. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, Vienna, Austria. Association for Computational Linguistics.

Chung-Chi Chen, Yu-Min Tseng, Juyeon Kang, Anaïs Lhuissier, Min-Yuh Day, Teng-Tsai Tu, and Hsin-Hsi Chen. 2023a. 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. -.

Chung-Chi Chen, Yu-Min Tseng, Juyeon Kang, Anaïs Lhuissier, Yohei Seki, Min-Yuh Day, Teng-Tsai Tu, and Hsin-Hsi Chen. 2023b. Multi-lingual ESG impact type identification. In *Proceedings of the Sixth Workshop on Financial Technology and Natural Language Processing*, pages 46–50, Bali, Indonesia. Association for Computational Linguistics.

Chung-Chi Chen, Yu-Min Tseng, Juyeon Kang, Anais Lhuissier, Yohei Seki, Hanwool Lee, Min-Yuh Day, Teng-Tsai Tu, and Hsin-Hsi Chen. 2024. Multilingual ESG impact duration inference. In *Proceedings of the Joint Workshop of the 7th Financial Technology and Natural Language Processing, the 5th Knowledge Discovery from Unstructured Data in Financial Services, and the 4th Workshop on Economics and Natural Language Processing*, pages 219–227, Torino, Italia. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, 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.

Nina Gorovaia and Michalis Makrominas. 2025. Identifying greenwashing in corporate-social responsibility reports using natural-language processing. *European Financial Management*, 31(1):427–462.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. *Preprint*, arXiv:2111.09543.

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. *Preprint*, arXiv:1907.11692.

Yohei Seki, Hakusen Shu, Anaïs Lhuissier, Hanwool Lee, Juyeon Kang, Min-Yuh Day, and Chung-Chi Chen. 2024. Ml-promise: A multilingual dataset for corporate promise verification. *Preprint*, arXiv:2411.04473.

Harsha Vardhan, Sohom Ghosh, Ponnurangam Kumaraguru, and Sudip Naskar. 2023. A low resource framework for multi-lingual ESG impact type identification. In *Proceedings of the Sixth Workshop on Financial Technology and Natural Language Processing*, pages 57–61, Bali, Indonesia. Association for Computational Linguistics.

Yosef Ardhito Winatmoko and Ali Septiandri. 2023. The risk and opportunity of data augmentation and translation for ESG news impact identification with language models. In *Proceedings of the Sixth Workshop on Financial Technology and Natural Language*

Processing, pages 66–71, Bali, Indonesia. Association for Computational Linguistics.