WC Team at SemEval-2025 Task 6: PromiseEval: Multinational, Multilingual, Multi-Industry Promise Verification leveraging monolingual and multilingual BERT models

Takumi Nishi
University of Illinois Urbana-Champaign
takumin2@illinois.edu

Nicole Miu Takagi
Waseda University
miu.n.takagi@toki.waseda.jp

Abstract

This paper presents our system developed for SemEval-2025 Task 6: PromiseEval: Multinational, Multilingual, Multi-Industry Promise Verification. The task aims at identifying "promises" made and "evidence" provided in company ESG statements for various languages. Our team participated in Subtasks 1 and 2 for the languages English, French, and Japanese. In this work, we propose using BERT and finetuning it to better address the task. We achieve competitive results, especially for English and Japanese.

1 Introduction

Corporate Environmental, Social, and Governance (ESG) statements often contain forward-looking promises – commitments to sustainability, social responsibility, or ethical governance – that significantly influence public trust and corporate reputation. In recent years, there has been increasing emphasis placed on companies' ESG commitments (Curtis et al., 2021; Li et al., 2021). Ensuring the integrity of ESG promises is not only vital for transparency, but also for upholding stakeholder trust and holding organizations accountable for their commitments.

However, the complexity and lack of standardization in these statements pose a challenge to innovate new natural language processing (NLP) approaches to assess their strength and verifiability.

In this context, promise verification has emerged as an important task: systematically checking if a stated promise is made and supported by evidence in company reports in the form of SemEval Task 6 Chen et al. (2025). The task we address here, is titled PromiseEval: Multinational, Multilingual, Multi-Industry Promise Verification and is divided into the following sub-tasks:

(A) ESG promise identification

(B) Evidence identification in support of promise

Analyzing multilingual corporate commitments means NLP models must handle diverse linguistic expressions of promises – from English and French to Japanese, Chinese, and beyond – often with limited annotated data in each language. Recent efforts have begun to address this: for instance, Seki et al. (2024) introduced ML-Promise, the first multilingual dataset for corporate promise verification, covering five languages (English, French, Chinese, Japanese, Korean) to facilitate cross-lingual ESG promise analysis.

Given that the data implies a classification task, one significant approach is to fine-tune pre-existing language models for each language. This is the methodology taken in this paper.

Advances in Natural Language Processing offer a promising path toward automating promise verification. In particular, transformer-based language models such as BERT have proven highly effective at modeling context and meaning in text, especially against traditional machine learning text classification methods such as TF-IDF (Garrido Ramas et al., 2021). BERT's flexible architecture enables finetuning for custom classification tasks by adding only a simple output layer, yet achieves robust performance.

Verifying ESG promises is inherently a multifaceted and multilingual challenge. As will be discussed earlier, results performed better for single-language BERT models, where learning in multilingual models did not transfer to other languages.

Promise evaluation in natural language processing is an important area of research which focuses on evaluating if promises are made in certain statements and if they are supported by evidence. This paper systematically explores methodologies for training language models to identify and evaluate these things. Recent developments have led to several promising approaches for evaluating multilingual text, which will be explored in Background.

Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 1670–1676

2 Background

The training data sets were provided by the PromiseEval2025 organizers, covering five languages, including English, French, Japanese, Chinese, and Korean (Seki et al., 2024; Chen et al., 2025). Each data set, covering one of the five languages, was created by extracting texts from ESG reports released by various corporate organizations headquartered in countries where the language is the native tongue. For example, the Japanese data set used Japanese ESG reports published by companies primarily operating or originating from Japan, while the French data set used those from France. Following extraction, the texts were segmented into paragraphs or sentences. They were later annotated with labels related to the four subtasks set by the organizers.

Our team focused on subtasks 1 ("Identifying Promises") and 2 ("Linking Evidence to the Promise") for data sets pertaining to the languages English, French, and Japanese. For the goal of each subtask, "Identifying Promises" required creating models that could determine whether each segmented text contained promise statement(s), while "Linking Evidence to the Promise" focused further on being able to identify whether each text (assuming they contained promise statements) have any evidence statements to support their claim. Additionally, for Asian languages (including Japanese), the subtasks further required models to be able to extract the exact sentences. However, given the complexity and challenge of accurately extracting exact promise and evidence texts, we had limited our model to verify whether a dataset contains promise and evidence text. As for the labels, we used "promise_status" and "evidence_status" to create models for subtasks 1 and 2, respectively. Note that in the Asian language data sets (including Japanese), annotators further included "promise_string" and "evidence_string," which referenced the specific sentences within the data where promise and evidence statements were made.

Concerning ESG specifically, there has been growing interest in methods to assess the credibility of ESG promises. Early NLP research in this domain involves analyzing ESG reports for insights, such as Shi Bowen (2023) analyzing similarities in sustainability reports through the application of latent dirichlet allocation, and support vector machine models to analyze ESG scores against a list of company mergers and acquisitions, and Petridis et al. (2022) utilizing a retrieval-augmented generation approach to idenfity Sustainable Development Goals in environmental impact assessments.

Aside from the aforementioned methods, some of the algorithms that can be found in the literature for identification and classification tasks are Naive Bayes (Bayes, 1968), stochastic gradient descent (Zhang, 2004), k-nearest neighbors (Liao and Vemuri, 2002), decisions trees (Charbuty and Abdulazeez, 2021), Convolutional Neural Networks (Zhang and Wallace, 2015) and Support Vector Machines (Tong and Koller, 2001; Joachims et al., 1999). In 2018, however, a revolutionary language model was released by Google, titled Bidirectional encoder representations from transformers (BERT) (Devlin et al., 2019). Utilizing an encoder-only transformer architecture, it dramatically improved over previous state-of-the-art models, and remains to perform exceptionally well.

An example in utilizing BERT in classifying text to quantify ESG ratings, Schimanski et al. (2024) trained RoBERTa and DistilRoBERTa, while Pasch and Ehnes (2022) utilized transformer-based models to train an ESG sentiment model based on ESG ratings and text documents.

Relatedly, an annotated environmental claim detection dataset was presented by Stammbach et al. (2023), focusing on identifying statements about environmental actions or impacts in corporate communications. This work, along with efforts by Seki et al. (2024) further highlight the importance of identifying ESG claims. Our work builds on these efforts, thus contributing to the field of promise and claim detection and verification.

3 System Overview

Given our team's focus on creating robust models for Promise-verification tasks by fine-tuning existing state-of-the-art LLM models, our strategy for detecting promise and evidence statements in both subtasks revolved around leveraging the base BERT model and its derivatives for French, Japanese and multilingual.

The decision to adopt BERT models for our task is primarily associated with its flexible architecture. It enables researchers to train task-specific models with only another output layer added to the base model, producing highly satisfactory results (Devlin et al., 2019). The successful application of BERT has also been noted in text classification tasks for ESG-related research. Addi-

tionally, BERT's transfer learning feature further allows models to be created from relatively small datasets with limited computational resources. This was another reason to adopt BERT models, as the datasets for each language contained only 400 rows of data, which is even smaller than the traditional number of what is considered a 'small dataset.'

Our team's approach to creating accurate models that can verify Promise and its associated evidence was to use monolingual models explicitly dedicated to each language, alongside a multilingual one encompassing all three languages. For English, we adopted the base BERT model for fine-tuning, while French and Japanese used language-specific BERT models.

For the French model, we chose to fine-tune CamemBERT, a BERT model dedicated to French texts. CamemBERT's architecture is based on RoBERTa, an improved iteration of the original BERT model, and its training was done using French datasets extracted from the OSCAR corpus (Martin et al., 2020). In terms of performance, it has produced superior results in downstream tasks, such as part-of-speech tagging and natural language inference, against mBERT. In addition, research by Kelodjoue et al. (2022), further found that CamemBERT and in text-classification tasks on verbatim transcripts alongside online posts compared to FlauBERT (another BERT model fine-tuned on French dataset).

Regarding the Japanese models, we used tohoku BERT[1], a well-known BERT model that was fine-tuned for Japanese data and is one of the widely adopted models in Japanese NLP research. In particular, research by Shibayama and Shinnou (2021), found that fine-tuning Tohoku BERT led to creating a Japanese-based sentence-BERT that demonstrated higher performance than other models they created.

Finally, we also utilized mBERT[2] to create multilingual model for both promise and evidence verification tasks. Previous researches that compare performances of monolingual and multilingual models in various classification tasks often provide differing results on which strategies are overall better than the other. Nonetheless, in many of the same studies, researchers have also highlighted how multilingual models offer competitive results with similar performance metrics to monolingual models (Velankar et al., 2022; Zhao and Aletras, 2024; Lothritz et al.). Most critically, however, multilingual models have been found to produce superior results in complex classification and retrieval tasks (Conneau et al., 2020; Ranaldi et al., 2025). For example, in the paper by Hu et al. (2020), they found models like XLM-R perform well in natural langauge inference tasks. Another research by Dementieva et al. (2022) which is directly related to evidence verification, showcased models trained on cross-lingual fake news evidence datasets can yield advantageous results in evidence classifications.

[1]https://github.com/cl-tohoku/bert-japanese

[2]https://github.com/google-research/bert/blob/master/multilingual.md

4 Experimental Setup

4.1 Pre-processing

The pre-processing process for both subtasks was generally the same regardless of the dataset's language. We first began by label encoding the values of "promise_status" and "evidence_status" using the preprocessing module of the Scikit-learn library. For the Japanese dataset, several values within the "evidence_status" also contained NA values aside from the standard boolean values of "Yes" and "No." As the organizers did not give specific explanations or instructions about handling NA values, we treated them as being the same as "No" values (evidence statement is not present inside the data) and converted them accordingly before the encoding process. For the multilingual model training, after label encoding all 3 language datasets, we concatenated them into a single single dataset.

Since our initially strategy was to create monolingual models specialized for each language, we used the AutoTokenizer class from the HuggingFace library for the tokenization process. AutoTokenizer is a practical tool that automatically chooses the correct tokenizer based on the BERT model we set, including CamemBERT and TohokuBERT. Thus, making it easier to fine-tune the three languages without alternating parts of our code. After the tokenization, the data and labels were split into training and testing data with a ratio of 8:2.

4.2 Model Training and Hyper-parameters

The four models we fine-tuned included bert-base-uncased, camembert-base, and cl-tohoku/bert-base-japanese for English, French, and Japanese data, respectively alongside mBERT for multilingual training. As we noticed during the pre-processing phase, all three datasets had a problem with the balance of labels, with the "promise_status" being the most

serious with an average ratio of "Yes" to "No" being 3:1. While "evidence_status" was relatively more balanced, "Yes" was more prevalent for all datasets. Therefore, to mitigate the potential issue of model bias against the minority label, we added class weights as part of our code.

For hyper-parameters, we set the sequence length for all models to 256 instead of the traditional 128 to account for any variations in the length of each dataset. Others, specifically, batch size, epochs, and learning rate, were determined via multiple trials using metrics of training and validation loss to determine the optimum number. The final results for hyper-parameters for each model used for the competition are shown below.

Model	Bs	Ep	LR
English Subtask 1	8	5	1e-05
French Subtask 1	8	5	2e-05
Japanese Subtask 1	16	4	2e-05
Multilingual Subtask 1	16	4	2e-05
English Subtask 2	16	6	5e-06
French Subtask 2	16	6	2e-05
Japanese Subtask 2	16	5	1e-05
Multilingual Subtask 2	16	4	1e-05

Table 1: Batch Size (Bs), Epoch (Ep) and Learning Rate (LR) for each models submitted for the competition

5 Results

5.1 Monolingual Model Performances

5.1.1 Subtask 1 Results

Language	Accuracy	F1-Score (4.s.f)
English	80.25%	0.8672
French	76.75%	0.8558
Japanese	94%	0.9687

Table 2: Accuracy and F1-Score for Subtask 1 models

In terms of general accuracy for the subtask 1 models which were submitted for evaluation, English was 80.25%, French 76.75%, and Japanese scored 94%. As promise_status had a serious issue of label imbalance, we further calculated each model's F1 score as well, and the results, which were rounded to 4 significant figures, are as follows (in the same language order): 0.8672, 0.8558, and 0.9687. We find that our model performs generally well from the accuracy results, producing competitive results with new, unseen data. Additionally, the F1 score for all models can be interpreted that the class weights helped mitigate potential bias against the minority class of "No" values for our models.

The Japanese model is particularly noteworthy in both accuracy and F1 score. Not only did the fine-tuned Tohoku model outperform the other languages in terms of accuracy, but it also had the highest F1 score. The latter result was especially impressive for our team, given the imbalance in the "promise_status" of the Japanese dataset, with "No" values representing only around 11% of the data while others averaged at least 20%.

5.1.2 Subtask 2 Results

For the models submitted for subtask 2, accuracy-wise, it was 71.75%, 76.75%, 73.5% for English, French, and Japanese, respectively. Regarding the F1 score, again in the same order of language, it was 0.7483, 0.8239, and 0.80. For Japanese, as the evidence_status in the final test dataset used for evaluating our Japanese models included NA values, we used the same pre-processing steps during the training phase. We converted them to "No" before calculating the accuracy and F1 score.

Language	Accuracy	F1-Score (4.s.f)
English	71.75%	0.7483
French	76.75%	0.8239
Japanese	73.5%	0.80

Table 3: Accuracy and F1-Score for Subtask 2 models

Compared to subtask 1 models, we see that subtask 2 models have an accuracy issue in correctly verifying whether or not a dataset contains evidence statements, with the French being an exception, where its evidence verification model performed slightly better than its promise model. The same can be partially said about the F1 scores, as the performance of all models dips compared to the results seen in subtask 1. In particular, we found that the English subtask 2 model performed the worst among all models submitted for this task. However, considering that the scores ranged from 0.74 to around 0.82, our models perform relatively well in detecting both majority and minority classes.

5.2 Multilingual Model Performance

5.2.1 Subtask 1 Results

Comparing the multilingual model's performance across the three languages with the monolingual model, we see that it yields similar results in accu-

Language	Accuracy	F1-Score (4.s.f)
English	78.25%	0.8581
French	77%	0.8585
Japanese	92.25%	0.9593

Table 4: Accuracy and F1-Score for Subtask 1 with Multilingual Model

racy and F1 Score. In terms of figures, the model slightly underperformed in English and Japanese, by around 2% in accuracy; it performed better with French data compared to the CamemBERT model. Overall, the multilingual model for subtask 1 produced results comparable to those of monolingual models.

5.2.2 Subtask 2 Results

For subtask 2, the multilingual model noticeably demonstrated an overall higher performance in classifying evidence status in the data when directly compared with monolingual models. Unlike the results we've seen in the subtask 1, it demonstrated better performances in both areas of accuracy and F1-scores for English and Japanese against the monolingual model created from the baseline BERT and TohokuBERT. However, when compared to CamemBERT, the multilingual model's performance was significantly worse, by around 9 points, and was the worst-performing metric across all models and tasks.

Language	Accuracy	F1-Score (4.s.f)
English	73.75%	0.7870
French	67.25%	0.7745
Japanese	75%	0.8148

Table 5: Accuracy and F1-Score for Subtask 2 with Multilingual Model

5.3 Monolingual Models or Multilingual Model

Comparing the results of monolingual models against multilingual model across both subtasks, we see that by performance metrics alone we found that each strategy offered different advantages in classification task with ESG datasets. For subtask 1, we found that the monolingual approach outperformed in promise verification. For subtask 2, multilingual model was able to handle the detection of evidences better, with the exception of French data. This result is relatively supportive of past research that have shown multilingual model's pretraining leading to favorable metrics in complex classification/retrieval tasks such is the case with evidence verifications (Conneau et al., 2020; Ranaldi et al., 2025). Regardless, as we observed both strategy's overall performance to be comparable, we found that using either approach can lead to competitive results.

5.4 Competition Ranking

Our official ranking in the competition is as follows: 10th out of 11 for English, 3rd out of 4 for French, and 2nd out of 3 for Japanese. Based on ranking alone, our model's performance was not the best in the competition for the three languages. However, we note that the PromiseEval submission on Kaggle (in all languages) for the evaluation phase on Kaggle was not segregated by task. Instead, participants were required to submit a copy of the evaluation dataset where we had replaced the values for the labels related to the tasks we did (for our case, promise_status and evidence_status), while keeping other labels unedited. This means that our submission was evaluated for all subtasks (1 through 4) available on PromiseEval rather than just the subtasks we actually worked on, making it difficult to assess our models' performance for subtask 1 and 2 compared to other participants.

6 Conclusion

In this paper, we introduce our results for the Semeval task. Our scores on evaluation data show that models fine-tuned on general corpora can obtain competitive results, especially for English. BERT performed well on promise identification, and less well on evidence identification.

However, a more thorough comparison between multilingual and language-specific BERT is expected to yield more definitive answers as to which methodology is superior overall. While we addressed label imbalance by adjusting class weights during training, our model may still be more biased towards 'yes' data. As such, techniques such as Synthetic Minority Over-sampling Technique (SMOTE), undersampling the majority class, or implementing alternative loss functions such as Focal Loss may further improve model robustness, especially in tasks with severe imbalance in the training dataset.

The introduction and successful implementation of the Promise Verification tasks are expected to have impacts on various fronts, including enhanced

accountability and transparency, empowerment of stakeholders, policy and regulation shaping, and increased social and environmental impact. For example, by providing a systematic and scalable approach to verify promises, this task could significantly improve the transparency of organizations and public figures, compelling them to adhere more closely to their commitments. This, in turn, could foster greater trust and credibility among stakeholders and the general public. Additionally, equipped with data and insights from the Promise Verification task, stakeholders, including consumers, investors, and the general public, could make more informed decisions based on the verifiable actions and commitments of organizations and leaders. Ultimately, by holding entities accountable for their promises, especially those related to ESG commitments, this task could contribute to more tangible progress in addressing environmental issues, promoting social justice, and ensuring ethical governance, leading to a more sustainable and equitable world.

References

Thomas Bayes. 1968. Naive bayes classifier. *Article Sources and Contributors*, pages 1–9.

Bahzad Charbuty and Adnan Abdulazeez. 2021. Classification based on decision tree algorithm for machine learning. *Journal of applied science and technology trends*, 2(01):20–28.

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.

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.

Quinn Curtis, Jill Fisch, and Adriana Z. Robertson. 2021. Do esg mutual funds deliver on their promises? *Michigan Law Review*, 120(3):393–450.

Daryna Dementieva, Mikhail Kuimov, and Alexander Panchenko. 2022. Multiverse: Multilingual evidence for fake news detection.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Jose Garrido Ramas, Giorgio Pessot, Abdalghani Abujabal, and Martin Rajman. 2021. Identifying and resolving annotation changes for natural language understanding. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Papers*, pages 10–18, Online. Association for Computational Linguistics.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization.

Thorsten Joachims et al. 1999. Transductive inference for text classification using support vector machines. In *Icml*, volume 99, pages 200–209.

Emmanuelle Kelodjoue, Jérôme Goulian, and Didier Schwab. 2022. Performance of two French BERT models for French language on verbatim transcripts and online posts. In *Proceedings of the 5th International Conference on Natural Language and Speech Processing (ICNLSP 2022)*, pages 88–94, Trento, Italy. Association for Computational Linguistics.

Ting-Ting Li, Kai Wang, Toshiyuki Sueyoshi, and Derek D. Wang. 2021. Esg: Research progress and future prospects. *Sustainability*, 13(21).

Yihua Liao and V Rao Vemuri. 2002. Use of k-nearest neighbor classifier for intrusion detection. *Computers & security*, 21(5):439–448.

Cedric Lothritz, Kevin Allix, Bertrand Lebichot, Lisa Veiber, Tegawendé F. Bissyandé, and Jacques Klein. Comparing MultiLingual and multiple MonoLingual models for intent classification and slot filling. In *Natural Language Processing and Information Systems*, pages 367–375. Springer International Publishing.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah, and Benoît Sagot. 2020. CamemBERT: a tasty French language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219, Online. Association for Computational Linguistics.

Stefan Pasch and Daniel Ehnes. 2022. Nlp for responsible finance: Fine-tuning transformer-based models for esg. In *2022 IEEE International Conference on Big Data (Big Data)*, pages 3532–3536.

Konstantinos Petridis, Ioannis Tampakoudis, George Drogalas, and Nikolaos Kiosses. 2022. A support vector machine model for classification of efficiency: An application to ma. *Research in International Business and Finance*, 61:101633.

Leonardo Ranaldi, Barry Haddow, and Alexandra Birch. 2025. Multilingual retrieval-augmented generation for knowledge-intensive task.

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.

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.

Yunkyung Min Shi Bowen. 2023. Analyzing esg news articles and research papers through lda (latent dirichlet allocation).

Naoki Shibayama and Hiroyuki Shinnou. 2021. Construction and evaluation of Japanese sentence-BERT models. In *Proceedings of the 35th Pacific Asia Conference on Language, Information and Computation*, pages 731–738, Shanghai, China. Association for Computational Lingustics.

Dominik Stammbach, Nicolas Webersinke, Julia Bingler, Mathias Kraus, and Markus Leippold. 2023. Environmental claim detection. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1051–1066, Toronto, Canada. Association for Computational Linguistics.

Simon Tong and Daphne Koller. 2001. Support vector machine active learning with applications to text classification. *Journal of machine learning research*, 2(Nov):45–66.

Abhishek Velankar, Hrushikesh Patil, and Raviraj Joshi. 2022. *Mono vs Multilingual BERT for Hate Speech Detection and Text Classification: A Case Study in Marathi*, page 121–128. Springer International Publishing.

Tong Zhang. 2004. Solving large scale linear prediction problems using stochastic gradient descent algorithms. In *Proceedings of the twenty-first international conference on Machine learning*, page 116.

Ye Zhang and Byron Wallace. 2015. A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification. *arXiv preprint arXiv:1510.03820*.

Zhixue Zhao and Nikolaos Aletras. 2024. Comparing explanation faithfulness between multilingual and monolingual fine-tuned language models.