NLP-Explorers at BEA 2026 Shared Task 1: DeBERTa–CatBoost Weighted Ensemble Approach for L1-Specific Vocabulary Difficulty Prediction

Tayyab Latif[1], Asifa Bibi[1], Sabur Butt[2,*], Grigori Sidorov[1], Alexander Gelbukh[1]
[1]Centro de Investigación en Computación (CIC), Instituto Politécnico Nacional, Mexico
[2]Institute for the Future of Education (IFE), Tecnológico de Monterrey, Mexico
[*]Corresponding author: saburb@tec.mx

Abstract

Vocabulary difficulty prediction aims to estimate how difficult a word is for a learner. This is an important problem because word difficulty is shaped not only by the word itself, but also by the learner's background and the context in which the word appears. In this work, we predict continuous difficulty scores for English target words using learner-specific information. Our approach combines a fine-tuned DeBERTa v3 Large model with a CatBoost regressor trained on transformer-based embeddings. The final score is produced through weighted ensembling, where DeBERTa provides the main prediction and CatBoost adds a smaller complementary signal. Our final system achieved RMSE scores of 1.040 for Spanish, 0.992 for German, and 0.882 for Chinese. The results were also stable across multiple runs, showing that the model behaved consistently under small changes in ensemble weight. These findings show that a simple hybrid system can provide reliable performance for vocabulary difficulty prediction. They also suggest that combining strong contextual representations with a lightweight regression model is an effective way to model learner-sensitive word difficulty.

1 Introduction

The prediction of the difficulty of a word to a learner is a significant aspect of language education, as vocabulary knowledge determines reading comprehension, assessment results, and learning materials design (Zhang and Lu, 2025). Practically, though, word difficulty is not fixed. It is not only based on the word itself, but also on the linguistic background of the learner and the context in which the word is introduced (Felice and Skidmore, 2026). This makes vocabulary difficulty prediction both educationally meaningful and technically challenging.

Consequently, there is an increasing interest in constructing NLP models that are capable of predicting learner-sensitive difficulty scores in a consistent and scalable manner (Kapoor et al., 2026). Recent work has also started to examine vocabulary test item difficulty prediction with transformer-based models, showing that such architectures provide a strong recent baseline for this task (Skidmore et al., 2025). Even though transformer-based models have become the standard option in most NLP tasks, their usage in vocabulary difficulty prediction still raises practical concerns (Li et al., 2025). Predicting word difficulty is not only about understanding the general meaning of the word and its context. The system must capture small differences between words and project them to a continuous score that indicates learner difficulty. This complicates the task compared to many traditional text classification problems and implies that the final prediction step should be given as much attention as the transformer model itself (Li et al., 2025).

In this paper, we present our system submission to the BEA 2026 Shared Task on Vocabulary Difficulty Prediction for English Learners (Felice and Skidmore, 2026). Our approach combines DeBERTa v3 Large and CatBoost in a hybrid prediction framework. DeBERTa serves as the main contextual model, while CatBoost uses transformer-based features to provide an auxiliary prediction signal. The final score is obtained through weighted ensembling. This design allows the system to produce a more reliable estimate of vocabulary difficulty than a transformer-only model, especially when the task depends on small lexical differences and learner-specific context.

The rest of the paper is structured in the following way. We describe the data and the evaluation setting first. Then we consider the previous work that is the most related to our work. Then, we describe our methodology, which consists of preprocessing, DeBERTa fine-tuning, the CatBoost component, and the final ensemble strategy. Then

Proceedings of the 21st Workshop on Innovative Use of NLP for Building Educational Applications, pages 1106–1112

we show the results and compare the various models and submitted runs. We end with a discussion of the broader lessons from this study and directions for future research.

2 Shared Task Data and Evaluation

The shared task focuses on vocabulary difficulty prediction for English learners. In this task, the system is given an English target word together with learner-related information, and it must predict a continuous difficulty score for that word. The data are provided separately for three learner groups based on their first language: Spanish, German, and Mandarin Chinese.

Each example includes information about the target word and the prompt shown to the learner. In our system, these fields are not treated as separate inputs. Instead, the selected text components are combined into one input sequence, and the target word is explicitly marked inside the context before the text is passed to the model. This preprocessing step helps the model focus on the target item while still using the surrounding learner-specific context (See Table 1).

The dataset is divided into training, development, and test splits for each language group. In our experiments, each L1 contains 6,091 training instances, 677 development instances, and 748 test instances see Table 2. The training and development sets include the gold difficulty scores, while the test set is used for blind evaluation.

System performance is measured mainly with root mean squared error (RMSE). This is the main metric used to compare systems and determine their ranking. RMSE is defined as:

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2},$$

where y_i is the gold difficulty score, $\hat{y}_i$ is the predicted score, and N is the number of test items. A lower RMSE means better performance because it shows that the predicted scores are closer to the reference scores.

This dataset is useful for the task because it allows the model to learn from both the target word and the learner-specific context. It also enables comparison of how vocabulary difficulty varies across Spanish, German, and Mandarin Chinese learners, rather than treating all learners as a single group.

Field	Description	Example
`item_id`	Unique identifier for each item	`9`
`L1`	Learner first language	`es`
`en_target_word`	English target word	`capacity`
`en_target_pos`	Part of speech of the target word	`noun`
`en_target_clue`	Partial clue for the target word	`c_______`
`L1_source_word`	Source cue / translation	`cabida`
`L1_context`	Learner L1 context	short Spanish prompt
`GLMM_score`	Gold difficulty score in labelled data	`-1.727`

Table 1: Main fields used in the shared task dataset, illustrated with a Spanish labelled example.

3 Literature Review

Recent work on difficulty prediction has moved in two closely related directions: item difficulty modeling from assessment content and text difficulty modeling for language learners.

Kapoor et al. (2026) study item difficulty prediction for reading comprehension items using a repository built from standardized tests. Their work shows that difficulty depends on more than the question text alone. They annotate items with linguistic, test, and context features, and then compare these with LLM-based embeddings. Their results show that the best performance comes from combining these sources of information, while text analysis features and LLM embeddings perform similarly when used on their own. This suggests that interpretable linguistic features still remain useful even when modern embedding methods are available.

A related line of work focuses more directly on text difficulty for second language learners. Zhang and Lu (2025) align English text difficulty with CEFR levels for L2 readers and show that lexical, syntactic, and discoursal features can be used to classify texts effectively. Their findings are especially relevant because they show that text difficulty is not explained by traditional readability formulas alone. Instead, features such as age of acquisition, lexical decision time, mean length of T-unit, academic vocabulary, and complex nominals play an important role. They also show that the features that separate lower proficiency levels are not ex-

item_id	L1	en_target_word	L1_source_word	L1_context
6782	Spanish	luck	suerte	Te deseo suerte con tu nuevo trabajo
6771	German	herd	Herde	Die Viehherde graste auf der Weide
6772	Chinese	harm	伤害，损害	这类项目会损害环境。

Table 2: Illustrative examples showing the dataset format for Spanish, German, and Mandarin Chinese learners.

actly the same as those that matter at higher levels.

Li et al. (2025) examine item difficulty modeling with fine-tuned small and large language models. Their results show that fine-tuned small models such as BERT and RoBERTa remain very competitive, and that careful data augmentation can improve prediction further. In contrast, larger autoregressive models and prompting-based approaches are less effective in their setting, mainly because the task is constrained by small training data and imbalanced difficulty distributions. Their work is important because it shows that stronger results do not always come from using larger models, but from matching the modeling strategy to the structure of the task.

This shared task also builds on a longer line of work on lexical complexity and simplification. Earlier shared tasks, such as SemEval 2016 on Complex Word Identification and the BEA 2018 multilingual CWI task, treated the problem mainly as deciding whether a word is difficult for a reader (Paetzold and Specia, 2016; Yimam et al., 2018). Later tasks moved closer to graded prediction, especially SemEval 2021 Task 1 on Lexical Complexity Prediction, while the BEA 2024 MLSP shared task connected lexical complexity with multilingual simplification pipelines (Shardlow et al., 2021, 2024). Compared with these earlier settings, the present task focuses more directly on learner-specific vocabulary difficulty by taking the learner's first language and contextual information into account.

Taken together, these studies suggest that difficulty prediction benefits from three broad ideas: the use of learner- or task-relevant linguistic features, the inclusion of contextual information when available, and the careful choice of modeling architecture rather than relying only on model size. Our work follows this direction by combining a strong contextual encoder with a separate regression component, aiming to capture both contextual meaning and structured variation in the prediction of vocabulary difficulty. In particular, the transformer model provides the main contextual representation of the target word, while the regression branch offers an additional signal derived from dense embedding features. This design keeps the system simple, but still allows it to combine information from complementary sources.

4 Methodology

Our system combines a fine-tuned transformer model with a second-stage regression model built on dense embedding features. The main predictive signal comes from DeBERTa v3 Large (He et al., 2021), which is fine-tuned separately for each learner group. In parallel, we extract feature representations from DeBERTa and XLM-RoBERTa (Conneau et al., 2020) and use them as input to a CatBoost regressor (Prokhorenkova et al., 2018). The final prediction is obtained by combining the transformer score and the CatBoost score through weighted averaging. Figure 1 shows a high-level view of the workflow.

4.1 Preprocessing and Feature Extraction

The input to the transformer is created by combining several text fields defined in the model configuration file. During preprocessing, these fields are merged into a single sequence in a fixed order, separated by the tokenizer's separator token when available. The target word is explicitly marked inside the context by surrounding it with special tags, <t> and </t>. This design means that the model does not process the target word in isolation. The remaining unused columns are removed during preprocessing.

For both models, we used an identical set of input features: the L1 source word, L1 context, English part of speech, English clue, and English target word. To support the second-stage regression model, we extract dense vector representations from two pretrained transformers: DeBERTa v3 Large and XLM-RoBERTa Large. The embedding script loads a tokenizer and base transformer model for each encoder, applies tokenization with padding

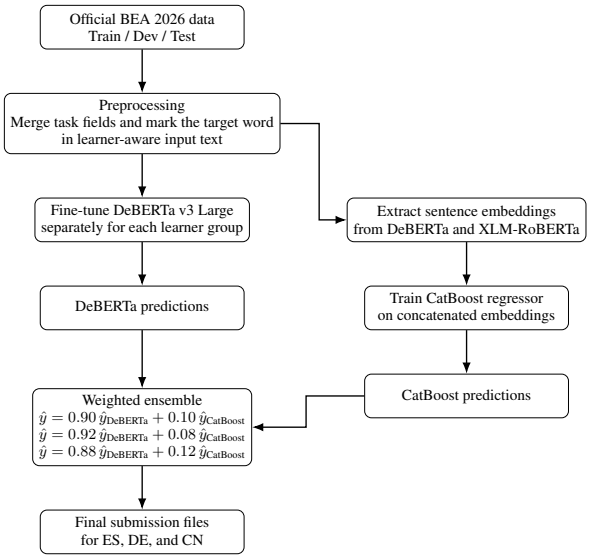

Figure 1: Workflow of the proposed system. DeBERTa v3 Large serves as the main prediction model, while a CatBoost regressor is trained on sentence embeddings extracted from DeBERTa and XLM-RoBERTa. The final prediction is obtained through weighted ensembling. The figure reports the main ensemble setting, 0.90/0.10, together with the two additional weight configurations, 0.92/0.08 and 0.88/0.12, explored in our submitted runs.

and truncation, and then computes a sentence-level representation using mean pooling over the last hidden states.

4.2 Experimental Design

The transformer component is trained using the Hugging Face Trainer API.[1] Tokenization is performed after preprocessing, with truncation applied to the combined input text. Training is carried out with epoch-based evaluation, epoch-based checkpointing, and selection of the best model at the end of training.

DeBERTa, short for Decoding-enhanced BERT with Disentangled Attention, is a transformer-based encoder that models content and positional information separately. This design helps the model capture fine contextual differences more precisely, which is especially useful in vocabulary difficulty prediction where small lexical and contextual cues matter. We trained the DeBERTa model for each language using 6 epochs with a batch size of 16, a learning rate of 2e-5, and a warmup ratio of 0.1. The weight decay was set to 0.1 for Spanish and Chinese, and 0.0 for German. We kept these settings because they provided a stable training process and gave the strongest development performance in our experiments, while still keeping the setup simple and consistent across the three learner groups.

In our final submission, DeBERTa v3 Large serves as the main neural model. Its role is to learn a direct mapping from the learner-aware input representation to the target difficulty score. Since the model is trained separately for each learner group, it can adapt to differences across Spanish, German, and Mandarin Chinese learners without mixing the three data distributions during inference. CatBoost is a gradient boosting regressor designed to model continuous values from structured feature representations. In our system, it serves as a second-stage model over embedding-based features and provides an auxiliary prediction that complements the main transformer output. Additionally, the CatBoost stage uses concatenated DeBERTa and XLM-RoBERTa embeddings as its input features. For each learner group, the code loads the embedding arrays, concatenates the two representation spaces, and fits a CatBoostRegressor using the training labels from the official training split. The model is trained with RMSE as the loss function, 2000 boosting iterations, a learning rate of 0.03, and a tree depth of 6. The CatBoostRegressor operates on fixed embedding vectors and contributes an auxiliary score that complements the main DeBERTa output. This is useful because the final regression problem may involve nonlinear interactions in the representation space that are not fully exploited by a standard transformer regression head.

The final prediction is obtained by combining the DeBERTa score and the CatBoost score through weighted averaging. Before combining the two files, the code checks that the item_id ordering matches exactly. The final prediction is then computed as:

$$\hat{y}_{\text{final}} = \alpha\,\hat{y}_{\text{DeBERTa}} + (1-\alpha)\,\hat{y}_{\text{CatBoost}}, \quad (1)$$

where α is the weight assigned to the DeBERTa prediction. For the main run, we used:

$$\hat{y}_{\text{final}} = 0.90\,\hat{y}_{\text{DeBERTa}} + 0.10\,\hat{y}_{\text{CatBoost}}. \quad (2)$$

We also generated two additional runs using nearby weights, namely 0.92/0.08 and 0.88/0.12, to test whether small changes in the ensemble balance would improve stability or final performance (See Table 4). The script uploaded here corresponds to

[1] https://huggingface.co/docs/transformers/main_classes/trainer

the run with weights 0.88 and 0.12, but the same logic was used for all three submitted runs.

5 Results

This section presents the performance of the models explored in our experiments and compares them with the final ensemble system. We first report the development results of the main transformer models that were considered during model selection. We then present the official results of the three submitted ensemble runs, which were obtained by combining DeBERTa v3 Large and CatBoost with slightly different weights.

5.1 Transformer Models

Table 3 compares the three main transformer settings that were examined during development: XLM-RoBERTa Base, XLM-RoBERTa Large, and DeBERTa v3 Large. The results show a clear and consistent pattern across all three learner groups. XLM-RoBERTa Large improved substantially over XLM-RoBERTa Base, which suggests that increased model capacity was useful for this task. However, the strongest results were obtained with DeBERTa v3 Large in every language.

The advantage of DeBERTa was especially clear in Chinese, where it reached the lowest RMSE among the three transformer settings. The same trend also appeared in Spanish and German. This consistent improvement made DeBERTa the strongest candidate for the main prediction branch of the final system.

Model	L1	RMSE ↓	Pearson ↑
XLM-RoBERTa Base	Spanish	1.380	0.708
XLM-RoBERTa Base	German	1.383	0.710
XLM-RoBERTa Base	Chinese	1.227	0.711
XLM-RoBERTa Large	Spanish	1.161	0.828
XLM-RoBERTa Large	German	1.153	0.824
XLM-RoBERTa Large	Chinese	1.061	0.820
CatBoost	Spanish	1.737	0.423
CatBoost	German	1.698	0.375
CatBoost	Chinese	1.464	0.499
DeBERTa v3 Large	Spanish	**1.021**	**0.854**
DeBERTa v3 Large	German	**1.059**	**0.838**
DeBERTa v3 Large	Chinese	**0.925**	**0.841**

Table 3: Development set results of the main transformer models across the three learner groups. Lower RMSE and higher Pearson indicate better performance.

5.2 Final Ensemble

We submitted three runs with nearby weight settings. The main run used a weight of 0.90 for DeBERTa and 0.10 for CatBoost. Two additional runs, 0.92/0.08 and 0.88/0.12, were included to test whether a small change in the ensemble balance would improve the final performance. The results are shown in Table 4.

Run	Weights (D/C)	Spanish	German	Chinese
Run 1	0.90 / 0.10	**1.040**	**0.992**	0.883
Run 2	0.92 / 0.08	1.040	0.993	**0.882**
Run 3	0.88 / 0.12	1.041	0.992	0.885
Best Model	- / -	0.903	0.885	0.776

Table 4: Official RMSE results of the three submitted ensemble runs, compared with the best model submitted in the shared task. D/C denotes the weights assigned to DeBERTa and CatBoost, respectively. Lower values are better.

The final ensemble results were very close across the three runs. This is consistent with the development results, where DeBERTa was already the strongest standalone model.

6 Discussion

The gap between the standalone DeBERTa model and the final ensemble is much smaller. This is also an important finding. It suggests that most of the predictive signal already comes from DeBERTa, while CatBoost provides a limited but useful correction. The small differences across the three submitted ensemble runs support the same interpretation. If the gains had been random, the relative order of the runs would likely have varied more sharply across languages. Instead, the scores remain very close, which points to a stable system with a fairly fixed error profile (See Table 4). Overall, the results show that the final system benefited more from a strong contextual encoder than from aggressive reweighting in the ensemble stage. The submitted runs confirmed that the proposed hybrid approach was stable across languages, with the strongest final performance observed for Chinese and the most difficult condition appearing in Spanish.

A likely explanation for the gap between the weaker models and the final model is the way they represent lexical information. Vocabulary difficulty prediction is not only a matter of general sentence meaning. The model must detect subtle differences

in target-word usage, clue structure, and learner-specific context, then map that information to a continuous score. Smaller or less task-suited encoders may capture broad semantic information, but they are more likely to smooth over the fine distinctions that matter in this setting. DeBERTa appears to reduce this problem, and the CatBoost component adds a second decoding layer over dense embeddings, which helps recover some structured patterns that may be missed by a simple transformer regression head.

From a hypothesis-testing perspective, several item-level factors are worth examining in future work. The first is *word length*, since longer words may be more morphologically complex and therefore harder both for learners and for the model. The second is *frequency*, because rare words are usually more difficult and may also be less well represented in pretrained models. The third is *part of speech*, as verbs, adjectives, and abstract nouns may not behave in the same way. A fourth factor is *context dependence*: some words can be understood from the clue alone, while others require close reading of the full prompt. These factors are promising because they offer a concrete way to test whether the model's residual errors are concentrated in specific lexical categories rather than distributed randomly across the dataset.

7 Conclusion

In this paper, we have shown that combining DeBERTa v3 Large with a CatBoost regressor trained on embeddings extracted from DeBERTa v3 Large can provide good estimates of the difficulty of target English words and incorporate additional user-specific data into consideration. Among the transformers examined in the course of our study, DeBERTa v3 Large showed the best performance on its own, outperforming XLM-RoBERTa Base and XLM-RoBERTa Large in Spanish, German, and Chinese. The presented ensemble confirmed the stability of a simple linear combination of an effective contextual language model and a relatively lightweight machine learning model. We showed that the effectiveness of the system is largely dependent on the system's ability to capture subtle lexical and contextual information about the given input. Second, the improvement comes not merely from increasing the size of the system or adding layers, but from the careful integration of complementary signals in the process. In our case, the main power was provided by DeBERTa, whereas CatBoost added a minor correction that improved predictions. The small gap between the submitted runs shows that the system was relatively stable in spite of changing ensemble weights.

However, there are also several important limitations. While Chinese was the most successfully predicted language, Spanish was the most difficult one, indicating that the link between the learner, contextual information, and word difficulty may be differently tractable across learner populations. An important direction of future research would consist of analyzing errors at the level of individual items in order to find out whether additional explicit lexical features help in the predictions.

Acknowledgments

This research was funded by the Challenge-Based Research Funding Program, Grant No. I030-IFE002-C2-T1-E, of the Institute for the Future of Education, Tecnológico de Monterrey, Mexico. The work was also carried out with partial support from Grant No. 20260626 (G.S.) awarded by the Secretaría de Investigación y Posgrado (SIP) of Instituto Politécnico Nacional, Mexico.

References

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. Association for Computational Linguistics.

Mariano Felice and Lucy Skidmore. 2026. Findings of the bea 2026 shared task on vocabulary difficulty prediction for english learners. In *Proceedings of the 21st Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2026)*, San Diego, California. Association for Computational Linguistics.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced BERT with disentangled attention. In *International Conference on Learning Representations*.

Radhika Kapoor, Sang T. Truong, Nick Haber, Maria Araceli Ruiz-Primo, and Benjamin W. Domingue. 2026. Prediction of item difficulty for reading comprehension items by creation of annotated item repository. *arXiv preprint arXiv:2502.20663*.

Ming Li, Hong Jiao, Tianyi Zhou, Nan Zhang, Sydney Peters, and Robert W Lissitz. 2025. Item difficulty modeling using fine-tuned small and large language models. *Educational and Psychological Measurement*, 85(6):1065–1090.

Gustavo Paetzold and Lucia Specia. 2016. SemEval 2016 task 11: Complex word identification. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 560–569, San Diego, California. Association for Computational Linguistics.

Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin. 2018. Catboost: unbiased boosting with categorical features. In *Advances in Neural Information Processing Systems*.

Matthew Shardlow, Fernando Alva-Manchego, Riza Batista-Navarro, Stefan Bott, Saul Calderon Ramirez, Rémi Cardon, Thomas François, Akio Hayakawa, Andrea Horbach, Anna Huelsing, Yusuke Ide, Joseph Marvin Imperial, Adam Nohejl, Kai North, Laura Occhipinti, Nelson Peréz Rojas, Nishat Raihan, Tharindu Ranasinghe, Martin Solis Salazar, and 3 others. 2024. The BEA 2024 shared task on the multilingual lexical simplification pipeline. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 571–589, Mexico City, Mexico. Association for Computational Linguistics.

Matthew Shardlow, Richard Evans, Gustavo Paetzold, and Marcos Zampieri. 2021. SemEval-2021 task 1: Lexical complexity prediction. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*. Association for Computational Linguistics.

Lucy Skidmore, Mariano Felice, and Karen Dunn. 2025. Transformer architectures for vocabulary test item difficulty prediction. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*. Association for Computational Linguistics.

Seid Muhie Yimam, Chris Biemann, Shervin Malmasi, Gustavo Paetzold, Lucia Specia, Sanja Štajner, Anaïs Tack, and Marcos Zampieri. 2018. A report on the complex word identification shared task 2018. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 66–78, New Orleans, Louisiana. Association for Computational Linguistics.

Xiaopeng Zhang and Xiaofei Lu. 2025. Aligning linguistic complexity with the difficulty of english texts for l2 learners based on cefr levels. *Studies in Second Language Acquisition*, 47(5):1407–1434.