Data Science and Engineering (DSE) Record, Volume 6, issue 1.

Online Platform for English Language Practice and Proficiency

Poosana Thassanavisut [1] and Sakgasit Ramingwong [2]

[1] Master's Degree Program in Data Science, Chiang Mai University, Chiang Mai, Thailand
[2] Department of Computer Engineering, Faculty of Engineering, Chiang Mai University, Chiang Mai, Thailand
poosana_thas@cmu.ac.th

Abstract. This research develops an English learning platform that utilizes a combination of three machine learning models for grammatical topic classification in language assessment. English serves as one of the most crucial languages in today's world, particularly in education, work, and communication. However, learning and developing English language skills remains a significant challenge for many learners, especially in countries where English is not the primary language. Firstly, this independent study aims to address these challenges by developing a comprehensive English learning platform that incorporates advanced machine learning techniques. Secondly, the platform employs three distinct machine learning approaches: facebook/bart-large-mnli, Logistic Regression, and DeBERTa for automated grammatical topic assignment to examination questions. Finally, the empirical results demonstrate that the developed platform effectively enables users to assess their English proficiency according to the CEFR (Common European Framework of Reference for Languages) standards, while providing appropriate skill evaluation across various grammatical topics.

Keywords: TOEIC, Grammar Classification, Machine Learning, NLP, DeBERTa, Proficiency Assessment.

1 Introduction

English is one of the most important languages in the world today. It is widely used in education, work, and communication between countries. Having good English skills gives people many advantages around the world. However, for people who do not speak English as their first language, learning English is still very difficult. This is because they do not have enough chances to use English in daily life and lack good ways to learn the language effectively.

Today, technology and online platforms play an important role in helping people learn English. These platforms are easy to access, offer many different types of content, and allow learners to study continuously. Online learning has changed language education by letting people learn at their own pace and time. However, many current learning platforms do not have good assessment systems that follow international

standards. They also cannot give personalized recommendations that match each learner's skill level.

This research aims to develop an English learning platform that can measure and display users' English language proficiency according to CEFR standards, including recommendations for skill improvement based on test results. The platform will use machine learning systems to classify test questions according to categories commonly found in TOEIC Part 5 (Incomplete Sentences) examinations. The study focuses on TOEIC Part 5 test data collected from Kaggle website, BARRON'S TOEIC Practice Exams, TBX VicTOEIC Mock Test book, and various online learning courses. This platform will provide guidance for practice and development of English language skills, helping to connect standard assessment with personalized learning to make English learning better for users everywhere.

CEFR Level	TOEIC Score (Tannenbaum & Wylie, 2008)		TOEIC Score (Damayanti & Gafur, 2020).
	Listening section	Reading section	Overall score
C2	-	-	-
C1	490	455	945-990
B2	400	385	785-940
B1	275	275	550-780
A2	110	115	225-545
A1	60	60	120-220

Table 1. Mapping between CEFR Proficiency Levels and TOEIC Scores

From Table 1, the CEFR (Common European Framework of Reference for Languages) provides a standard way to measure language proficiency across six levels, from A1 (beginner) to C2 (advanced). The corresponding TOEIC scores show how these international standards relate to widely-used English proficiency tests. For example, learners at B1 level typically score around 275 points in both listening and reading sections, with an overall score range of 550-780 points. This mapping allows the platform to accurately assess users' English proficiency and provide appropriate learning recommendations based on their current level.

2 Literature Review

2.1 Machine Learning-Driven Language Assessment

Machine Learning-Driven Language Assessment [1] uses Machine Learning and Natural Language Processing (NLP) to develop English language tests and assess the

difficulty level of each test question using machine learning algorithms instead of conducting real trials with test takers. The system includes Computer Adaptive Testing (CAT) to adjust test questions according to the test taker's proficiency level. The study results show that scores obtained from this testing method correspond well with Item Response Theory (IRT) analysis and other standard tests. However, there are limitations as the test does not measure writing and speaking skills, which are skills that should be developed in future research.

2.2 Methods for Language Learning Assessment at Scale: Duolingo Case Study

Methods for Language Learning Assessment at Scale: Duolingo Case Study [2] examines how the Duolingo platform applies Educational Data Mining (EDM) and Learning Analytics (LA) techniques to design systems that encourage learner engagement and improve learning outcomes. The researchers used Checkpoint Quizzes as post-lesson assessments to measure learner progress and Review Exercises for reviewing previous lessons. These two system components help Duolingo continuously track user learning performance and enhance the learning experience on the platform more effectively.

2.3 A Study on the Effectiveness of Learning Strategies in English

A Study on the Effectiveness of Learning Strategies in English [3] investigates learning behaviors in developing English language skills. The study found that successful learners tend to develop active learning strategies and have learning patterns, strategy preferences, and language use patterns that differ from typical learners. The research also revealed that speaking skills are the most important skill for developing English language proficiency.

3 Data and Methodology

3.1 Data

This study uses TOEIC Part 5 (Incomplete Sentences) test data, which consists of multiple-choice questions with 4 options. A total of 3,625 questions were collected from online sources (Kaggle), online courses, and TOEIC practice books. The data is divided into 4 datasets

1) **toeic_test.json:** Data from Kaggle containing 3,625 questions without topic labels.

2) **toeic_label_test.json:** Data from TOEIC books and online courses with grammar topic labels, containing 180 questions.

3) **toeic_label_balanced_augmented.json:** Data that has been augmented and balanced to have similar numbers of examples in each topic, covering 18 grammar categories.

4) **toeic_test_manual_label.json:** Data with manually labeled topics, containing 1,060 questions from the Kaggle dataset for model evaluation.

The classification follows 18 grammatical topics such as adjective, adverb, verb patterns, subject-verb agreement, and preposition.

Before using the data, initial data cleaning was performed, including correcting misspelled topic names and removing questions without topics. Three input formats were created for model training: question text only, question with all 4 choices, and question with choices and correct answer. The platform provides 45 questions with a 60-minute time limit and displays scores, estimated TOEIC Reading scores using the formula

$$\text{Estimated TOEIC Score} = \left(\frac{45}{\text{Correct Answers}}\right) \text{x} 495$$

Estimated CEFR levels based on TOEIC score conversion, topics scoring below 40% for improvement recommendations, score summaries by topic, and answer revealing.

3.2 Methodology

This research applies machine learning techniques for text classification, divided into unsupervised and supervised learning approaches.

3.2.1 Unsupervised Learning: Zero-shot Classification

The unsupervised approach uses Zero-shot Classification with the facebook/bart-large-mnli model, which was trained on MNLI dataset and can effectively evaluate relationships between text and specified topics without requiring additional training data. TOEIC Part 5 questions are converted to text containing questions and choices as model input, and the model selects the topic with the highest probability.

3.2.2 Supervised Learning

1) **Logistic Regression:** Logistic Regression is a linear classification model that converts text to numerical vectors using TF-IDF technique, with class_weight='balanced' to address class imbalance issues.

2) **DeBERTa:** DeBERTa is a deep language model developed by Microsoft Research that incorporates disentangled attention and enhanced mask decoder concepts to improve understanding of word relationships in sentences. The pre-trained microsoft/deberta-v3-base model was fine-tuned with grammar topic data.

3.2.3 Training Configuration

Both supervised models were trained using two datasets: original data from TOEIC books (train_small) and balanced augmented data (train_aug). Three input formats were compared for all models: question only, question with choices, and question with choices and answer. Training settings for DeBERTa included truncation for text longer than 128 tokens with padding to ensure equal text length, suitable for deep learning and mini-batch processing.

4 Result

4.1 Impact of Input Context on Model Performance

Adding contextual information significantly improves classification accuracy across all models. As shown in Table 2, the progression from question-only to question+choices+answer consistently enhances performance. For example, DeBERTa with augmented data improves from 45% accuracy (question-only) to 60% (question+choices) to 64% (question+choices+answer). This pattern demonstrates that additional context helps models better understand grammatical structures and question intent.

4.2 Effect of Training Data Size and Quality

Data augmentation substantially impacts model performance, particularly for complex models. DeBERTa shows dramatic improvement from train_small to train_aug datasets, jumping from 16-17% accuracy to 45-64% accuracy. Traditional models like Logistic Regression also benefit from augmented data but to a lesser extent, improving from 8-22% to 12-27% accuracy, indicating that transformer-based models are more sensitive to training data quality and quantity.

4.3 Model Architecture Comparison

Model	question	question + choices	question + choices + answer
facebook/bart-large-mnli	-	0.06	
Logistic Regression train_small)	0.08	0.18	0.22
Logistic Regression (train_aug)	0.12	0.22	0.27
DeBERTa (train_small)	0.16	0.16	0.17
DeBERTa (train_aug)	0.45	0.60	0.64

Table 2. Accuracy Comparison Across Models

Model	question	question + choices	question + choices + answer
facebook/bart-large-mnli	-	0.04	
Logistic Regression train_small)	0.05	0.15	0.18
Logistic Regression (train_aug)	0.06	0.17	0.21
DeBERTa (train_small)	0.03	0.03	0.03
DeBERTa (train_aug)	0.19	0.39	0.47

Table 3. Macro F1-Score Comparison Across Models

DeBERTa with augmented data achieves the highest performance (64% accuracy, 0.47 F1-score), significantly outperforming traditional Logistic Regression (27% accuracy, 0.21 F1-score). Zero-shot classification shows limited effectiveness (6% accuracy, 0.04 F1-score), highlighting the importance of domain-specific fine-tuning for grammar topic classification.

4.4 Model Performance with Limited Training Data

Complex models require sufficient training data to realize their potential. DeBERTa with small training data performs poorly (16-17% accuracy) and shows high class bias, while simpler models like Logistic Regression maintain relatively stable performance regardless of data size. This suggests that transformer-based models need careful consideration of data requirements for practical implementation.

5 Implementation

After applying the best-performing model to assign grammar topics to the dataset of interest, the labeled data was stored in a database to serve as English proficiency assessment tests for platform users. When users complete the 45-question test within 60 minutes, the system provides comprehensive multi-dimensional assessment results.

The platform calculates estimated TOEIC Part Reading scores based on correct answers and converts these scores to corresponding CEFR levels. It displays scores categorized by all 18 grammar topics and identifies topics where users scored below 40%, which are considered weak areas requiring additional practice. For example, if a user scores poorly on Subject-Verb Agreement and Preposition topics, the system displays recommendations such as "Should practice more on: Subject-Verb Agreement and Preposition" based on actual learner performance. The platform also provides complete answer explanations for immediate review.

The platform's capabilities extend beyond typical scoring or general answer explanations, focusing on in-depth analysis to provide personalized feedback - a strength that general platforms or rule-based systems cannot achieve. While conventional systems may only show which questions were answered incorrectly, they cannot identify specific grammatical weaknesses. Additionally, many general platforms lack systematic TOEIC to CEFR score comparison and cannot provide accurate, systematic topic recommendations for improvement.

6 Conclusion

This study investigated machine learning approaches for classifying grammatical topics in TOEIC Part 5 questions using both unsupervised and supervised learning methods. The zero-shot classification approach using pre-trained models without domain-specific training achieved limited performance with only 6% accuracy and 0.04 macro F1-score, indicating that general language models require task-specific fine-tuning for effective grammar topic classification. Logistic Regression, representing traditional machine learning approaches, showed modest performance improvements when trained with augmented data and enhanced input formats, achieving up to 27% accuracy and 0.21 macro F1-score. However, its TF-IDF representation limits its ability to capture complex grammatical relationships. DeBERTa demonstrated superior performance when properly fine-tuned with augmented data, achieving 64% accuracy and 0.47 macro F1-score with complete input format (question+choices+answer), indicating that transformer-based models have significant potential for TOEIC grammar topic classification when provided with sufficient and diverse training data.

Future research should focus on expanding the labeled dataset, particularly for topics with limited examples such as causative verbs, future time clause, and conditional sentences, to enable comprehensive learning and reduce class imbalance issues. Additionally, labeling methodology should be improved through expert validation to ensure data quality and consistency. Implementing these recommendations would enhance model accuracy, coverage, and reliability in classifying TOEIC question topics, leading to more effective automated English proficiency evaluation systems that provide personalized learning recommendations based on detailed grammatical analysis.

References

1. Settles, B., LaFlair, G., & Hagiwara, M. (2020). Machine learning–driven language assessment. Transactions of the Association for Computational Linguistics, 8, 247–263. https://doi.org/10.1162/tacl_a_00310
2. Portnoff, L., Gustafson, E., Rollinson, J., & Bicknell, K. (2021). Methods for language learning assessment at scale: Duolingo case study. In *Proceedings of the 14th International Conference on Educational Data Mining (EDM)*.
3. Khansir, A. A., Dehkordi, F., & Mirzaei, M. (2023). A study on the effectiveness of learning strategies in English. *Indian Journal of Language and Linguistics*, *4*(2), 18–31. https://doi.org/10.54392/ijll2323