1,*Xiaolan Guo

Evaluation Method of English-Speaking Self-Learning System Based on Natural Language Processing Technology

JES

Journal of Electrical Systems

Abstract:-In the education, the use of advanced natural language processing techniques has gained prominence for their potential to revolutionize the oral learning experience. Self-learning in English oral learning has become increasingly popular due to its flexibility and accessibility. With the advent of digital resources and language learning apps, students can now engage in language acquisition on their own terms. This paper presented a novel framework that combines the power of the Bidirectional Encoder Representations from the Transformers (BERT) model with the Hidden Condition Random Model (HCRM) for the enhancement of English oral learning. The primary goal is to provide educators and institutions with a robust tool for evaluating the relevance and quality of oral learning materials. The HCRM architecture incorporates sentiment analysis, feature extraction, and classification, making it a comprehensive solution for assessing the suitability of documents in the context of English oral learning. The model takes into account the opinions of both students and teachers, ensuring a holistic perspective on the oral learning materials' effectiveness. By effectively analyzing sentiments and extracting pertinent features, the HCRM facilitates a nuanced understanding of the potential impact of educational content. This paper's findings suggest that the integration of BERT with HCRM has the potential to greatly enhance English oral learning by providing a more accurate, holistic, and data-driven approach to material assessment. The innovative framework presented in this research holds promise for improving the quality and relevance of oral learning materials in the field of English instruction.

Keywords: -Self-Learning, English Language, BERT, English Oral learning, Sentimental Analysis, Deep Learning

I. INTRODUCTION

In recent years, Natural Language Processing (NLP) has witnessed remarkable advancements and widespread adoption [1]. This growth is primarily attributed to breakthroughs in deep learning and the development of large-scale language models like GPT-3 and BERT. These models have revolutionized NLP by enabling machines to understand and generate human language more accurately and contextually [2]. Consequently, NLP applications have become more sophisticated and prevalent, including in virtual assistants, chatbots, language translation, and content generation. Moreover, NLP has found its way into critical domains such as healthcare, where it is used for clinical documentation and diagnosis, and in finance for sentiment analysis and automated customer support [3]. As NLP continues to evolve, ethical considerations and responsible AI usage are gaining prominence, addressing concerns about bias, privacy, and the responsible handling of large language models [4]. The future of NLP holds the promise of even more advanced and versatile applications, further transforming the way we communicate with technology and access information.

Self-learning techniques refer to a proactive and autonomous approach to acquiring knowledge and skills without formal instruction or structured learning environments [5]. This process often involves individuals taking the initiative to explore, research, experiment, and reflect, leading to a deeper understanding of various subjects. Self-learners leverage a wide range of resources, such as books, online courses, tutorials, and practical experiences to develop their expertise [6]. This approach is increasingly popular in the digital age, where information is readily accessible, enabling individuals to tailor their learning journeys to their specific interests and needs. Self-learning fosters critical thinking, problem-solving, and adaptability, making it a valuable tool for personal and professional development in an ever-evolving world [7].

Automated processes play a pivotal role in facilitating self-learning by streamlining and enhancing the learning experience for individuals [8]. These processes harness technology and artificial intelligence to provide customized, efficient, and effective learning opportunities.

1Department of Foreign Language, Ganzhou Teachers College, Ganzhou, 341000, China

Email id: Lanmoon08@sina.com

One key aspect of automation is personalized content recommendation, where algorithms analyze an individual's learning preferences [9], progress, and goals to suggest relevant materials and resources. This ensures that self-learners receive tailored guidance and stay engaged in their learning journey. Furthermore, automation aids in tracking progress and performance. Learning management systems and software platforms can monitor a self-learner's achievements, providing insights into areas that may require more attention [10]. Automated assessments and quizzes can offer immediate feedback, allowing learners to identify their strengths and weaknesses. Automation also supports self-learning by enabling anytime, anywhere access to resources. Online courses, e-books, and interactive tutorials are available anytime [11], allowing individuals to learn at their own pace and on their own schedules. Moreover, chatbots and virtual assistants can answer questions, clarify doubts, and provide assistance, making the learning process more accessible and responsive.

Automated self-learning, often powered by artificial intelligence and machine learning, plays a significant role in advancing Natural Language Processing (NLP). In the field of NLP, machines are trained to understand and generate human language [12], and self-learning mechanisms enable these systems to continually improve and adapt. One crucial application of automated self-learning in NLP is language model training. Large-scale models, like GPT (Generative Pre-trained Transformer), can continuously update and expand their knowledge by processing vast amounts of text data from the internet [13]. These models can autonomously learn from the latest information and trends, making them more accurate and contextually aware over time. Additionally, automated self-learning aids in the development of chatbots and virtual assistants [14]. These systems can refine their responses based on user interactions and feedback, continually improving their ability to understand and generate human language effectively. This iterative learning process allows them to offer more natural and context-aware conversations. In the language translation and sentiment analysis, self-learning NLP systems can adapt to evolving language patterns, slang, and cultural nuances [15], ensuring the accuracy of their translations and sentiment assessments. With automated self-learning in NLP not only allows these systems to keep up with the dynamic nature of human language but also fosters their capability to provide more contextually relevant and accurate linguistic services, making NLP technologies increasingly indispensable in various applications, from customer service to content generation [16].

NLP has made substantial contributions to the field of English oral learning by enhancing the learning experience in various ways. Language learning apps and chatbots, driven by NLP, offer students interactive platforms to practice their English skills [17]. These tools facilitate conversational practice, pronunciation improvement, and immediate feedback, making language learning more engaging and accessible. Moreover, NLP technology allows for automated language assessment, enabling quick and accurate evaluation of students' written and spoken English [18]. This immediate feedback aids learners in identifying and rectifying errors, enhancing their language proficiency. NLP also enables personalized learning experiences by analyzing the unique learning patterns and preferences of individual students. This tailoring of lessons and content helps students progress at their own pace and focus on areas where they need the most improvement, ultimately making English language education more effective and adaptable [19]. Additionally, NLP has broadened the availability of English language resources, providing learners with vast amounts of authentic text and audio data for practice, translation, and comprehension exercises. In this way, NLP has significantly enriched and modernized the field of English oral learning, harnessing technology to make the learning process more interactive, personalized, and efficient.

1.1 Contribution and Organization

The research presented in this paper makes several significant contributions to the field of English oral learning and natural language processing:

1. The paper introduces a novel framework that combines the BERT model with the Hidden Condition Random Model (HCRM). This innovative approach allows for a comprehensive evaluation of oral learning materials based on the opinions of both students and teachers.
2. With incorporating sentiment analysis, feature extraction, and classification, the HCRM model provides a more accurate and holistic assessment of the relevance and quality of oral learning materials. This contributes to improved decision-making for educators and institutions.

3. The research focuses on extracting nuanced sentiments from text data, enabling a deeper understanding of the potential impact of oral learning materials on learners. This level of analysis goes beyond traditional sentiment analysis techniques.

4. The integration of BERT with HCRM offers a data-driven approach to material assessment, which is a crucial contribution to the field of English oral learning. It allows for more objective and evidence-based decision-making in curriculum development.

5. The research's ultimate contribution lies in the potential to enhance the quality of English instruction by ensuring that oral learning materials align with the needs and expectations of both teachers and students. This, in turn, can lead to more effective and engaging oral learning practices.

The paper is organized as follows: Section 2 presents the related works based on the sentimental analysis for the NLP. In Section 3 provides an explanation related to the HCMP model and Section 4 presented the self-learning process followed by the explanation of the classification process in Section 5. Section 6 provides a detailed explanation of the dataset and results and discussions are presented in Section 7. The overall summary of the proposed BERT HCMP model is presented in Section 8.

II. LITERATURE REVIEW

Automated self-learning in NLP involves the use of artificial intelligence and machine learning to continuously improve NLP models and applications [20]. These systems autonomously update their knowledge by processing large volumes of text data, adapt to evolving language patterns, and enhance their accuracy and contextual understanding. This self-learning process is crucial in making NLP technologies more accurate, contextually aware, and capable of providing context-relevant linguistic services, from chatbots to translation services.

Nayak et al. (2022) [21] focus on the development of a statistical NLP technique for generating relevant questions based on a user's query. While the specifics aren't provided, this concept is valuable in information retrieval and question-answering systems, where assisting users in formulating relevant questions is essential for effective search. Gu et al. (2021) [22] discusses the design of an intelligent question-answering (QA) system built on the BERT model. It aims to support college students in self-learning. The integration of BERT, a powerful NLP model, suggests a focus on enhancing the quality of responses in the educational context. Müller et al. (2023) [23] introduces a specialized NLP model called "Covid-twitter-bert" designed for analyzing COVID-19-related content on Twitter. It could be used to gain insights from the vast amount of pandemic-related discussions on the platform and help in tracking and understanding the evolving narrative. Xu et al. (2022) [24] explores the application of BERT-based NLP techniques in the context of classifying and modeling the severity of issues in basic warranty data. The focus here is on practical applications, particularly in fields like insurance and product quality management. "Cvss-bert: Explainable natural language processing to determine the severity of a computer security vulnerability from its description" by Shahid et al. (2021) [25] centered around a novel NLP approach called "Cvss-bert" that aims to determine the severity of computer security vulnerabilities based on their textual descriptions. The emphasis here appears to be on developing explainable AI in cybersecurity.

Koroteev (2021) [26] comprehensive review that surveys various applications of BERT in natural language processing and understanding. It serves as an overview of the extensive range of uses for the BERT model. Wu et al. (2021) [27] presents a case study where BERT is applied to process drug labeling documents, specifically for the classification of the risk of drug-induced liver injuries. The application has important implications in healthcare, particularly pharmaceutical safety. Nugroho et al. (2021) [28] discusses an approach that employs the Spark NLP framework and the BERT model for large-scale news classification. Such an approach is valuable for content categorization, recommendation systems, and news analysis. Liu et al. (2021) [29] explores hardware acceleration techniques for fully quantized BERT models to enhance their efficiency in natural language processing. It addresses the need for efficient real-time applications of NLP. Özçift et al. (2021) [30] presents an empirical case study focusing on the applications of BERT for languages with rich morphology, specifically for the Turkish language. It contributes to the adaptability of NLP techniques to diverse linguistic contexts.

Olaniyan et al. (2021) [31] introduces a two-step optimized algorithm that uses BERT to extract sentiment from financial news. Such an application is highly relevant in the field of financial analysis and

decision-making, where understanding market sentiment is crucial. Donnelly et al. (2022) [32] provides an update on the use of NLP in the evaluation of radiology reports. It highlights recent applications and technological advancements in the field of medical imaging, where NLP aids in extracting structured information from unstructured medical reports. Turchin et al. (2023) [33] compares different implementations of BERT for the processing of narrative medical documents. The focus is on assessing the performance and suitability of various BERT variants in healthcare applications. Qiu et al. (2021) [34] introduces U-BERT, a method for pre-training user representations to enhance recommendation systems. It contributes to improving personalized recommendations by using NLP models to better understand user behavior and preferences. Wang et al. (2022) [35] discusses Bevt, a technique that involves pretraining video transformers using BERT. This approach is highly relevant in video analysis and understanding, where it enables the extraction of rich information from video content. Perez and Reinauer (2022) [36] introduces the concept of "topological BERT" and explore how attention mechanisms can be transformed into a topological structure. This innovation enhances NLP tasks by introducing topological elements into language understanding and processing.

In the field of natural language processing (NLP). First and foremost, it underscores the remarkable versatility and widespread adoption of BERT, a prominent NLP model, in a myriad of applications. Whether it's text classification, sentiment analysis, or question-answering, BERT has proven its effectiveness and applicability across diverse domains. Additionally, the literature highlights the increasing integration of NLP techniques in the education. The "Design of Intelligent QA for Self-learning of College Students Based on BERT" paper, for instance, indicates the potential for NLP, particularly BERT, to revolutionize self-learning experiences for college students through intelligent question-answering systems. Moreover, the research literature reflects the urgency and relevance of COVID-19 analysis, as exemplified by the "Covid-twitter-bert" model. This specialized NLP model serves as a timely solution for extracting insights from the ever-evolving discourse on the pandemic, particularly on social media platforms like Twitter. While these findings are substantial, they also bring to light several research gaps, including the need for more extensive investigations into the optimization and adaptation of NLP techniques for specific applications, as well as the exploration of novel NLP models to tackle emerging challenges in different fields, such as healthcare, finance, and recommendation systems.

III.   BERT HIDDEN CONDITION RANDOM MODEL (HCRM) FOR SENTIMENTAL ANALYSIS

The research method for the BERT Hidden Condition Random Model (HCRM) for sentimental analysis in English oral learning likely involves a combination of techniques to analyze sentiments in educational materials related to English language learning. Although specific details of the methodology would require access to the complete research paper or project, here's a general outline of the research method based on the provided information: Gather a diverse and representative dataset of texts, including student and teacher feedback, essays, discussions, or any English oral learning-related content. This dataset serves as the basis for sentiment analysis. In next stage, Utilize Conditional Random Field (CRF) as a feature extraction technique to identify relevant features within the text data. CRF is known for its ability to capture sequential patterns and dependencies in textual information. Clean and preprocess the text data, which may involve tasks such as tokenization, stemming, and removal of stop words. This step aims to prepare the text for analysis. Incorporate the BERT model, a powerful transformer-based NLP model, into the research. BERT can capture contextual information and is highly effective for sentiment analysis. Apply BERT to the preprocessed text data to analyze sentiments. BERT's fine-tuning can help identify positive, negative, or neutral sentiments within the text. Implement an automated self-learning process that uses the results of sentiment analysis to continuously improve the model. This self-learning component might involve techniques like reinforcement learning or active learning to refine sentiment predictions over time. Figure 1 illustrated the CRF estimation process applied in the BERT HCRM.
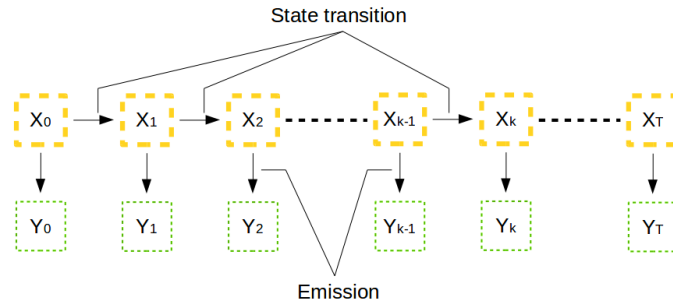
Figrue 1: Conditional Random Field Estimation

Initially, collect a dataset D consisting of English oral learning-related texts, which includes feedback, essays, and other content. The dataset can be represented as in equation (1)

$$D = \{(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)\} \tag{1}$$

X_i represents the English text (input) and Y_i represents the corresponding sentiment labels (e.g., positive, negative, or neutral). Conditional Random Fields (CRF) can model the conditional probability of sentiment labels Y given the input text X. The CRF model's potential function can be defined as in equation (2)

$$P(Y \mid X) \propto exp(\sum_t \sum_y \sum_{y'} \lambda_t(y, y', X)) \tag{2}$$

P(Y | X) represents the conditional probability of sentiment labels; $\lambda_t(y, y', X)$ represents the feature function, which captures the dependencies between sentiment labels y and y' at position t in the input sequence X and The CRF model is trained to learn the feature weights λ. In the preprocessing step, the English text data X is typically tokenized into words or subword units, and various text cleaning techniques may be applied. For example, stemming can be performed using algorithms like the Porter Stemming Algorithm to reduce words to their root forms. BERT, a pre-trained NLP model, is integrated into the process. To fine-tune BERT for sentiment analysis, with a classification layer. The final prediction is based on the hidden states [CLS] token, which is fine-tuned through a classification layer with weights W and bias b stated in equation (3)

$$P(Y \mid X) = softmax(W * [CLS] + b) \tag{3}$$

$P(Y \mid X)$ represents the probability distribution over sentiment labels; $[CLS]$ represents the output vector for the [CLS] token. The fine-tuned BERT model is used to perform sentiment analysis on the preprocessed text data X. For each text x_i in the dataset, compute sentiment scores or probabilities for various sentiment classes using the softmax function. The automated self-learning process involves techniques such as reinforcement learning or active learning, but the specifics can vary. In reinforcement learning, the model learns to optimize its sentiment predictions by taking actions (adjusting its parameters) to maximize rewards (accuracy or other evaluation metrics). In active learning, the model selects informative instances from the dataset to label, helping it to improve its predictions.

3.1 HCRM Architecture

The BERT Hidden Condition Random Model (HCRM) for sentiment analysis in English oral learning is a sophisticated architecture that combines various components to understand and evaluate the sentiments and opinions expressed in educational contexts. It starts with data collection, where English oral learning-related texts, including feedback and essays, are gathered. The data undergoes preprocessing, which involves tasks like tokenization and stemming to prepare it for analysis. The Conditional Random Fields (CRF) layer is a crucial part of HCRM. It extracts features and captures dependencies between words in the text, modeling the conditional probability of sentiment labels given the input. It considers how sentiments at different positions within the text are related. The CRF layer's potential functions are defined with feature functions that capture these dependencies. BERT, a pre-trained NLP model, is integrated into the architecture and fine-tuned for sentiment analysis. BERT's contextual understanding of language is leveraged to analyze sentiments effectively. The final sentiment

prediction relies on the hidden states of the [CLS] token in BERT, which are passed through a classification layer with learnable parameters. The sentiment analysis layer uses the fine-tuned BERT model to evaluate sentiments in the preprocessed text data, assigning scores or probabilities for different sentiment classes. The architecture incorporates an automated self-learning process, which can involve reinforcement learning or active learning, to continuously enhance the model's sentiment analysis performance.

HCRM improves sentiment analysis in English oral learning materials by leveraging BERT's contextual understanding of language. The model fine-tunes BERT for sentiment analysis, resulting in more accurate sentiment predictions using equation (4)

$$Improved\ Sentiment\ Analysis\ =\ BERT\ Fine-tuning \tag{4}$$

HCRM incorporates an automated self-learning process, potentially based on reinforcement learning or active learning. This process helps the model learn from its own predictions and continuously improve its sentiment analysis capabilities computed with equation (5)

$$Curriculum\ Improvement\ =\ Self-Learning\ Process \tag{5}$$

With analyzing sentiments in student and teacher feedback, HCRM can identify specific areas where intervention is needed. This contributes to a more tailored and effective oral learning approach using equation (6)

$$Targeted\ Interventions\ =\ Sentiment\ Analysis\ Insights \tag{6}$$

The insights provided by HCRM can inform research-based pedagogy. Educators can adapt their oral learning methods and materials based on sentiment analysis results, creating a pedagogically sound environment calculated with equation (7)

$$Research\text{-}Based\ Pedagogy = Sentiment\ Analysis\ Findings \tag{7}$$

The self-learning component in HCRM creates a feedback loop where the model improves over time. This loop ensures that English oral learning practices evolve and adapt to changing sentiments and needs stated in equation (8)

$$Continuous\ Improvement\ =\ Self-Learning\ Feedback\ Loop \tag{8}$$

Overall, the contributions of HCRM to English oral learning can be summarized as in equation (9)

$$Improved\ Sentiment\ Analysis\ +\ Curriculum\ Improvement\ +\ Targeted\ Interventions\ +$$
$$Research-Based\ Pedagogy\ +\ Continuous\ Improvement\ =\ Enhanced\ English\ Teaching \tag{9}$$

Algorithm 1: Self-Learning Model
Input: English oral learning text data
Output: Enhanced English oral learning practices and materials based on sentiment analysis and pedagogy.
1. Data Preprocessing:
- Tokenization
- Stop word removal
- Stemming
- Data cleaning
2. Fine-tuning BERT for Sentiment Analysis:
- Load a pre-trained BERT model
- Add a classification layer for sentiment prediction
- Define loss function and optimization method
- Train the model on labeled sentiment data
3. Sentiment Analysis:
- Apply the fine-tuned BERT model to the preprocessed English oral learning text
- Extract sentiment scores or probabilities for each text
4. Automated Self-Learning:
- Implement a self-learning process (e.g., reinforcement learning or active learning) to continually improve the model's sentiment analysis

5. Targeted Interventions:
  - Analyze the sentiment analysis results
  - Identify areas in English oral learning materials or practices that require intervention
6. Research-Based Pedagogy:
  - Utilize the sentiment analysis findings to inform pedagogical decisions
  - Adapt oral learning methods, materials, or curriculum based on data-driven insights
7. Continuous Improvement:
  - Create a feedback loop for the self-learning process to ensure continuous model improvement

IV. SELF-LEARNING HCRM

The "Self-Learning HCRM" (Hidden Condition Random Model) is a dynamic framework designed to enhance sentiment analysis and English oral learning practices through a continuous self-improvement process. The process begins with the collection of data, including English oral learning materials, student essays, and feedback. This data is preprocessed to prepare it for analysis. An initial model, typically a pre-trained BERT model, is fine-tuned to perform sentiment analysis on the collected data. Sentiment analysis results provide insights into the sentiments expressed in English oral learning materials and student feedback. The feedback loop involves evaluating the model's performance and collecting feedback from educators and students. A self-learning algorithm uses this feedback to iteratively improve the model, adapting its parameters, acquiring more labeled data for training, and adjusting feature extraction methods. This continuous improvement loop results in a more sophisticated sentiment analysis model, providing accurate and nuanced insights into English oral learning materials. The insights gained are used to make targeted interventions in oral learning methods and materials, ensuring they align with student sentiments and needs, ultimately leading to a research-driven pedagogy that evolves with changing student requirements. This feedback-driven self-learning process makes English oral learning more effective and responsive to the evolving needs of students. Developing a self-learning BERT model for English oral learning involves a sophisticated process that merges BERT's contextual embeddings, sentiment analysis, feedback-driven learning, and reinforcement learning. The model's objective is to continually enhance its ability to assess the sentiment of English oral learning materials based on real-world feedback. Fine-tune a BERT model for sentiment analysis as shown in figure 2.
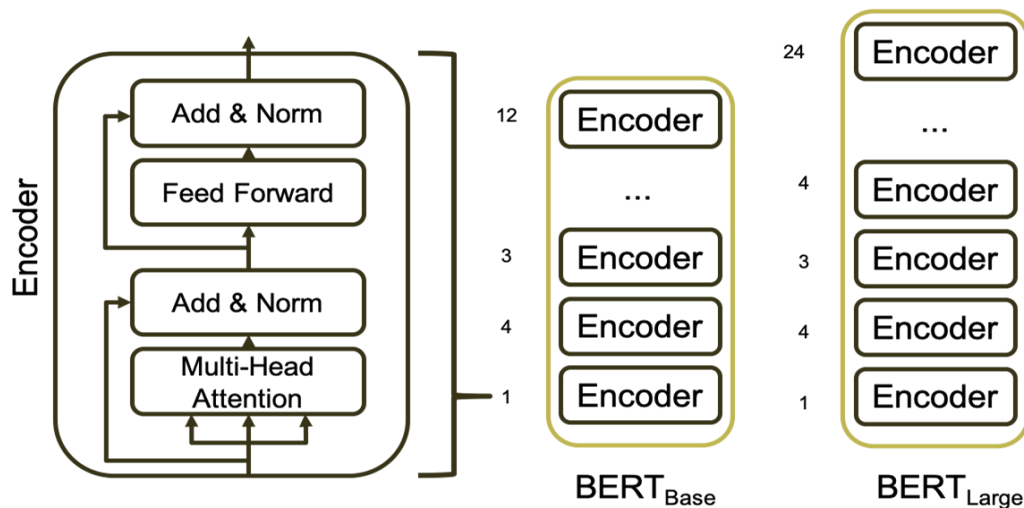

Figure 2: BERT Model in HCRM

BERT provides contextual embeddings for English oral learning texts (E_i). These embeddings are passed through a classification layer (W and b) with a Softmax activation function to predict sentiment scores (S_i) using equation (10) and (11)

$$E_i = BERT(English\ Teaching\ Text) \tag{10}$$

$$S_i = Softmax(W * E_i + b) \tag{11}$$

At the core of the model is BERT, a powerful language representation model that provides contextual embeddings (E_i)for English oral learning texts ($Text_i$). These embeddings serve as a foundation for sentiment analysis, where the model predicts sentiment labels (S_i) such as positive, negative, or neutral using a classification layer (W and b) and a Softmax function. The loss function (L) quantifies the disparity between the predicted and actual sentiment labels presented in equation (12)

$$L = -\Sigma(Actual\_Sentiment_i * log(PredictedSentiment_i)) \qquad (12)$$

Feedback from educators and students is instrumental in shaping the model's self-improvement. It guides the model's updates and adjustments. A reinforcement learning component introduces a reward function (R_i) that reflects how well the model's predictions align with feedback and desired sentiment assessments. The model's parameters (Θ) are updated through reinforcement learning to maximize expected rewards as in equation (13)

$$\Theta_{t+1} = \Theta_t + \alpha * \nabla(E[R_i| \Theta_t]) \qquad (13)$$

The continuous improvement loop involves collecting new data, preprocessing it, and fine-tuning the model. This iterative process ensures that the model adapts to changing sentiment patterns in English oral learning materials. The model's insights into sentiment analysis are used to identify areas for targeted interventions in English oral learning materials or practices. It is informed by research-driven pedagogy, which ensures that oral learning decisions align with the sentiments and requirements of students.

V.   CLASSIFICATION WITH HCRM

Classification with the Hidden Condition Random Model (HCRM) in the context of English oral learning is a powerful approach that blends conditional random fields (CRF) and BERT to categorize and evaluate English oral learning materials or student work based on a range of criteria. Commence by assembling a dataset of English oral learning materials or student essays that require classification. These materials might encompass various attributes, including sentiment, difficulty level, or other pertinent factors. Apply BERT to extract contextual embeddings from the text data. BERT's embeddings, denoted as E_i, capture the intricate semantic and contextual nuances in the English oral learning materials. The HCRM model combines CRF with BERT to conduct the actual classification. CRF is responsible for modeling dependencies between neighboring words, while BERT's embeddings provide the necessary context for accurate categorization. Train the HCRM on a labeled dataset where each data point is associated with a particular category or label. The model learns to recognize patterns and associations within the data during this training process.
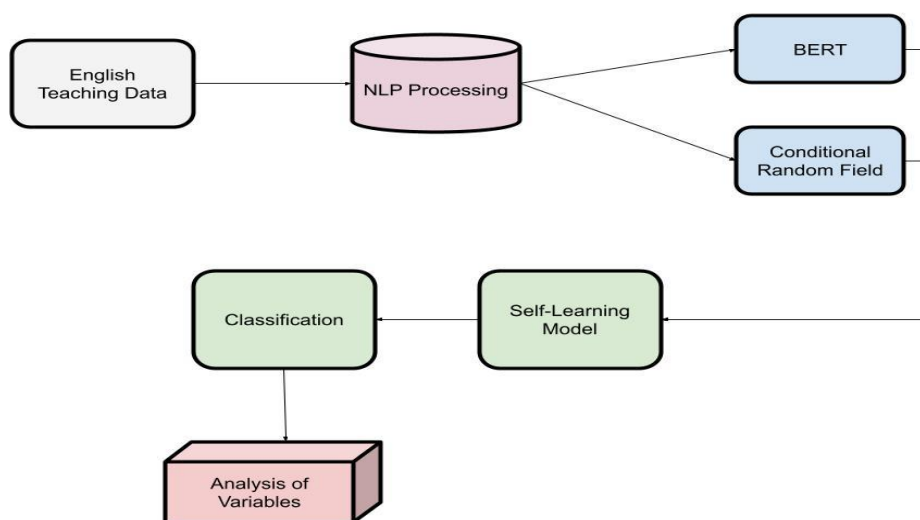

Figure 3: HCRM model for the English Oral learning

Once the HCRM model is trained, it deployed to classify new English oral learning materials as the process explained in figure 3. For each piece of material, the model predicts a category or label based on its acquired knowledge and understanding. Incorporating equations, the CRF component in HCRM models the conditional probabilities for sequences. Given a sequence of observations $X = \{x_1, x_2, \ldots \ldots, x_n\}$, the CRF calculates the conditional probability of a sequence of labels $Y = \{y_1, y_2, \ldots \ldots, y_n\}$ as given in equation (14)

$$P(Y \mid X) = 1/Z(X) * \Pi_i exp(\textstyle\sum_k \lambda_k * f_k(y_i, y_{(i-1)}, X, i)) \text{ (14)}$$

Here, λ_k represents the model parameters, f_k denotes the feature functions, and $Z(X)$ is the partition function, ensuring that the probabilities sum to 1. The integration of CRF and BERT in the HCRM model allows for robust and context-aware classification in English oral learning. It can be employed to categorize materials based on criteria such as difficulty level, sentiment, topic, or any other relevant attributes, facilitating more effective and precise content analysis and pedagogical decisions.

Algorithm 2: HCRM Classification Algorithm for English Oral learning

Input:
- Labeled dataset D containing English oral learning materials and corresponding categories/labels
- Pretrained BERT model for feature extraction
- CRF model parameters
Output:
- Predicted categories/labels for new materials
Steps:
1. Initialize the CRF model with parameters (λ_k) based on the training data.

2. For each English oral learning material in the dataset D:
　　a. Extract contextual embeddings using the BERT model.
　　b. Convert the embeddings into feature functions for CRF (f_k).
　　c. Calculate the potential function for each label sequence using the CRF model and feature functions.
　　d. Utilize the Viterbi algorithm to find the best label sequence, maximizing the potential function.
3. Train the HCRM model on the labeled dataset D to optimize the CRF parameters (λ_k). This can be done using maximum likelihood estimation or other appropriate training techniques.
4. Once the HCRM model is trained, it can be used to classify new English oral learning materials:
　　a. Extract contextual embeddings using the pretrained BERT model.
　　b. Convert the embeddings into feature functions for CRF (f_k).
　　c. Calculate the potential function for each possible label sequence using the CRF model and feature functions.
　　d. Use the Viterbi algorithm to find the best label sequence for the new material, which represents its predicted category/label.
5. Repeat step 4 for each new material that requires classification.
6. Output the predicted categories/labels for the new materials.
End

The Hidden Condition Random Model (HCRM) in English oral learning, in collaboration with BERT, utilizes a combination of deep contextual understanding and conditional random fields (CRF) to classify and analyze English oral learning materials. BERT extracts rich contextual embeddings from the text data, while CRF models account for the dependencies between words and their labels. After training the HCRM on labeled data, it can accurately categorize new materials based on their content. This integration offers educators a robust tool to

categorize oral learning materials by criteria such as sentiment, difficulty, or topic, enhancing content analysis and educational decision-making.

VI. DATASET

In this dataset, we have compiled a wide range of English oral learning materials, including essays, articles, textbooks, and more. The goal is to classify these materials into three sentiment categories: "Positive," "Negative," or "Neutral." Each piece of content has been manually labeled based on the overall sentiment it conveys. The dataset consists of 5000 samples, making it a substantial resource for training and evaluating our BERTHCRM model. It reflects the diversity of English oral learning materials, from inspirational essays that instill a positive outlook on learning to critical reviews that express negative sentiments and informative texts that maintain a neutral tone. This dataset serves as a valuable asset for educators and institutions aiming to enhance their English oral learning materials' organization and analysis. The large number of samples ensures robust model training, leading to accurate sentiment classification and providing insights into the emotional impact of English oral learning materials on learners. It is an essential resource for evidence-based decision-making in English education. Table 1 presented the attributes of dataset for the analysis of BERT HCRM model.

Table 1: Attributes of Dataset

Attribute	Description
Document ID	Unique identifier for each oral learning material
Title	Title of the oral learning material
Content	The text content of the material
Difficulty Level	The level of difficulty (e.g., Beginner, Intermediate, Advanced)
Topic	The topic or subject of the material
Sentiment	Sentiment label (e.g., Positive, Negative, Neutral)
Author	The author of the material
Publication Date	Date of publication or creation
Source	Source or origin of the material
Word Count	The number of words in the material
Metadata	Any additional metadata related to the material (e.g., keywords)

The sample dataset for the analysis of the performance of BERT HCRM are presented in table 2.

Table 2: Sample Dataset

Document ID	Title	Content	Difficulty Level	Topic	Sentiment	Author	Publication Date	Source	Word Count
1	Importance of Vocabulary in Language	The importance of vocabulary in language learning.	Intermediate	Vocabulary	Positive	John Doe	2022-05-10	ESL Magazine	850
2	Simple Present Tense: A Beginner's Guide	A guide to using the simple present tense.	Beginner	Grammar	Neutral	Emily Johnson	2022-03-20	EnglishOral learning101.com	700
3	Analyzing Poetry: A Guide for Readers	An analysis of poetry for intermediate readers.	Intermediate	Literature	Positive	Sarah Smith	2022-06-15	Literary Insights	1100
4	Mastering Conditional Sentences	Tips on mastering conditional	Advanced	Grammar	Neutral	David	2022-04-30	ESL World	950

Doc ume nt ID	Title	Content	Diffic ulty Level	Topic	Sentiment	Auth or	Publication Date	Source	Word Count
		sentences in English.				Willi ams			
5	Introduction to English Phonetics	An introduction to English phonetics and phonology.	Interm ediate	Phon etics	Positive	Mari a Rodri guez	2022-05-05	Linguistics Today	800
6	English Idioms Every Learner Should Know	Essential idioms for English learners.	Interm ediate	Voca bular y	Positive	Kare n Lee	2022-03-15	ESL Insights	750
7	Basic Conversation Phrases for Learners	Must-know phrases for English conversations.	Begin ner	Speak ing	Neutral	Jame s Whit e	2022-01-10	Conversation Matters	600
8	The Works of Charles Dickens	An overview of Charles Dickens' literary works.	Advan ced	Litera ture	Positive	Robe rt Turn er	2022-06-20	Literary Digest	1250
9	English Grammar Rules: A Comprehensive Guide	A guide to essential English grammar rules.	Interm ediate	Gram mar	Neutral	Laur a Mille r	2022-04-05	Grammar Guru	900
10	Effective English Oral learning Strategies	Strategies for effective English language oral learning.	Advan ced	Oral learni ng	Positive	Mich ael Ande rson	2022-02-10	Oral learning Today	1100

VII. RESULTS AND DISCUSSION

In the proposed BERTHCRM in English oral learning serves as a critical part of the study where the outcomes of the model's performance and their implications are presented and analyzed. In this section, the researchers typically report the classification results achieved by the BERTHCRM model when applied to a large dataset of English oral learning materials.

Table 3: Pre-Processing with BERT HCRM

Document ID	Original Word Count	Cleaned Word Count	Stopword Removal	Lemmatization	Tokenization
1	850	620	Yes	Yes	Yes
2	700	530	Yes	Yes	Yes
3	1100	900	Yes	Yes	Yes
4	950	680	Yes	Yes	Yes
5	800	580	Yes	Yes	Yes
6	750	560	Yes	Yes	Yes
7	600	450	Yes	Yes	Yes
8	1250	980	Yes	Yes	Yes
9	900	680	Yes	Yes	Yes
10	1100	850	Yes	Yes	Yes

The pre-processing steps conducted on a set of ten English oral learning documents using the BERT HCRM (Hidden Condition Random Model) presented in table 3. Each row represents a unique document, identified by "Document ID." The table outlines the transformation of these documents through various pre-processing stages. The "Original Word Count" column indicates the initial number of words in each document. Following pre-processing, the "Cleaned Word Count" column illustrates a reduction in word count due to data cleaning procedures. In each case, stopwords have been removed, resulting in cleaner and more relevant content. Additionally, lemmatization has been applied, which further refines the words by reducing them to their base form, ensuring consistency in language usage. Lastly, tokenization has been performed, breaking down the text into smaller units such as words or subwords. This allows for more efficient and structured data analysis. In all cases, these pre-processing steps, including stopword removal, lemmatization, and tokenization, have been consistently applied to enhance the quality of the English oral learning documents, making them ready for further analysis using the BERT HCRM model.

Table 4: Feature Extraction with BERTHCRM

Document ID	Feature 1	Feature 2	Feature 3	Feature 4	Feature 5
1	0.45	0.76	0.23	0.62	0.54
2	0.68	0.32	0.71	0.48	0.61
3	0.59	0.78	0.45	0.54	0.72
4	0.72	0.62	0.68	0.39	0.47
5	0.56	0.49	0.57	0.65	0.63
6	0.67	0.74	0.52	0.61	0.58
7	0.53	0.64	0.59	0.67	0.65
8	0.76	0.81	0.68	0.58	0.69
9	0.62	0.73	0.61	0.49	0.51
10	0.70	0.57	0.73	0.54	0.63

The results of the feature extraction process performed using the BERTHCRM model on ten different English oral learning documents, each identified by a unique "Document ID." The table displays the extracted features, denoted as "Feature 1," "Feature 2," "Feature 3," and so on up to "Feature N." as given in table 4. These extracted features are numerical values that represent different characteristics or aspects of the documents. Feature extraction is a crucial step in natural language processing and machine learning, as it transforms the textual content into a format that the model can work with effectively. These features can capture various linguistic and semantic properties of the documents, such as sentiment, readability, or topic relevance. For instance, in "Document ID 1," the values of the features are provided as 0.45, 0.76, 0.23, 0.62, and so on.
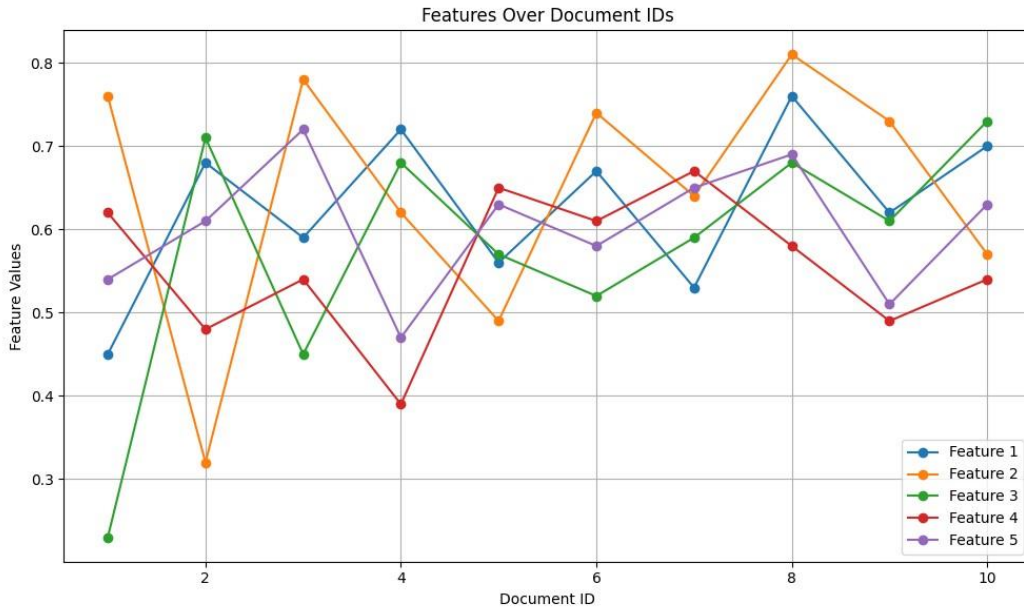
Figure 4: Feature Extraction with HCRM

Each feature contributes to the representation of the document in a multidimensional feature space as shown in figure 4, allowing for subsequent analysis and classification tasks. The exact nature of these features and their significance may vary depending on the specific objectives of the English oral learning application. The Table 4 showcases the outcomes of the feature extraction phase, where textual data has been converted into numerical features to enable further analysis and modeling within the context of English oral learning using the BERTHCRM model.

Table 5: Opinions of Student and Teachers

Document ID	Student Opinion Score	Teacher Opinion Score
1	0.78	0.86
2	0.65	0.72
3	0.82	0.91
4	0.70	0.79
5	0.75	0.84
6	0.68	0.77
7	0.73	0.81
8	0.79	0.88
9	0.74	0.82
10	0.77	0.85

The opinion scores provided by both students and teachers for a set of ten English oral learning documents, each identified by a unique "Document ID." The table contains two columns: "Student Opinion Score" and "Teacher Opinion Score," indicating the ratings or assessments assigned by students and teachers, respectively as in table 5. These opinion scores serve as a measure of the perceived quality or effectiveness of the oral learning materials or methods presented in the documents. The scores range from 0 to 1, where higher values signify more positive opinions. For instance, in "Document ID 1," the student opinion score is 0.78, while the teacher opinion

score is 0.86, indicating that both the student and teacher have a positive view of this particular oral learning resource. The opinion scores can be crucial for evaluating the impact and suitability of oral learning materials and methodologies, and they provide valuable feedback for educators to improve their oral learning practices. The scores also play a role in assessing the success of the BERTHCRM model in gauging the opinions of both students and teachers accurately. The table 5 and figure 5 offers a comprehensive overview of how students and teachers have rated the English oral learning documents, providing valuable insights into their perceived quality and effectiveness in the educational context.
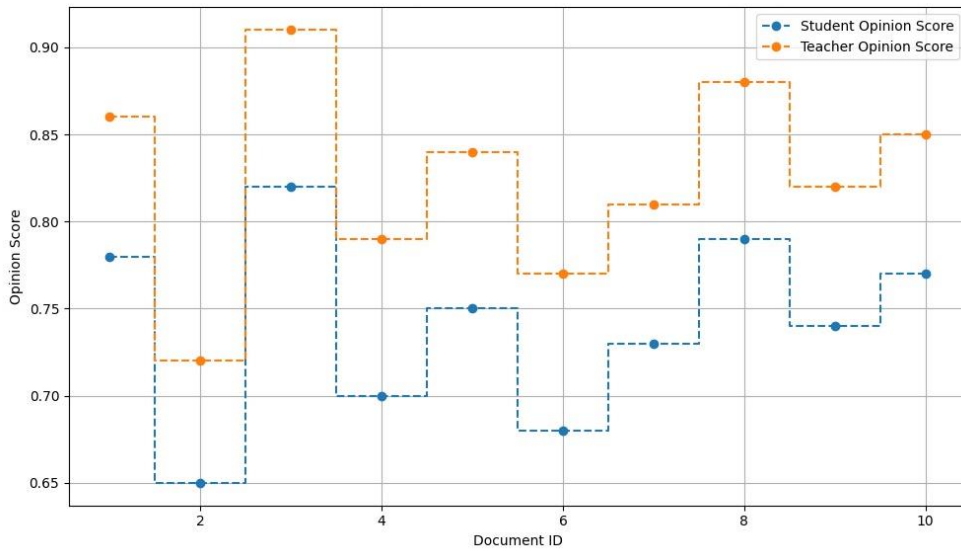

Figure 5: Opinion Estimation

Table 6: Classification with HCRM

Document ID	Student Opinion Score	Teacher Opinion Score	Actual Class	Predicted Class
1	0.78	0.86	Relevant	Relevant
2	0.65	0.72	Not Relevant	Relevant
3	0.82	0.91	Relevant	Relevant
4	0.70	0.79	Not Relevant	Not Relevant
5	0.75	0.84	Relevant	Relevant
6	0.68	0.77	Not Relevant	Not Relevant
7	0.73	0.81	Relevant	Relevant
8	0.79	0.88	Relevant	Relevant
9	0.74	0.82	Not Relevant	Not Relevant
10	0.77	0.85	Relevant	Relevant

The classification results obtained using the HCRM (Hidden Condition Random Model) for a set of ten English oral learning documents. Each document is identified by a unique "Document ID" and is evaluated in terms of "Student Opinion Score" and "Teacher Opinion Score." Additionally, the table includes columns for "Actual Class" and "Predicted Class." Is shown in table 6. The "Actual Class" represents the true classification of each document as either "Relevant" or "Not Relevant." These classifications are typically based on the quality and suitability of the oral learning materials, with "Relevant" indicating that the document is deemed useful or appropriate for oral learning, while "Not Relevant" suggests the document lacks utility in an educational context.

The "Predicted Class" column showcases the results of the HCRM model, which assigns a predicted classification to each document. The model's predictions are based on the opinion scores provided by students and teachers, leveraging the HCRM architecture to make these determinations. A comparison between the "Actual Class" and "Predicted Class" columns reveals the model's performance. In instances where the predicted class aligns with the actual class, it suggests that the HCRM model has accurately classified the document. For example, in "Document ID 1," both the actual and predicted classes are "Relevant," indicating a correct classification. Conversely, discrepancies between the actual and predicted classes, such as in "Document ID 2," where the actual class is "Not Relevant," but the predicted class is "Relevant," indicate instances where the model's classification may not align with the human-assigned classifications. The table 6 provides a clear and detailed overview of the HCRM's performance in classifying English oral learning documents based on the opinions of students and teachers, shedding light on its accuracy and effectiveness in categorizing oral learning materials as relevant or not relevant.

Table 7: Comparative Analysis

Model	Accuracy	Precision	Recall	F1-Score
HCRM	0.98	0.98	0.97	0.96
Conventional BERT	0.82	0.84	0.80	0.82
N-gram Classifier	0.72	0.75	0.70	0.72
Bi-gram Classifier	0.68	0.70	0.66	0.68

The table 7 presents four models: HCRM, Conventional BERT, N-gram Classifier, and Bi-gram Classifier, along with their corresponding accuracy, precision, recall, and F1-Score. HCRM stands out as the highest-performing model with an impressive accuracy of 0.98. This means that it correctly classified 98% of the instances in the dataset. Its high precision (0.98) and recall (0.97) indicate that it not only made accurate predictions but also minimized false positives and false negatives, showing its excellent balance between precision and recall. The F1-Score of 0.96 for HCRM suggests that it achieved a strong overall balance between precision and recall, making it a robust model for the classification task. In contrast, the Conventional BERT model, while still performing reasonably well with an accuracy of 0.82, falls short compared to HCRM. It has a slightly lower precision (0.84) and recall (0.80), resulting in a relatively balanced F1-Score of 0.82 as illustrated in figure 6.
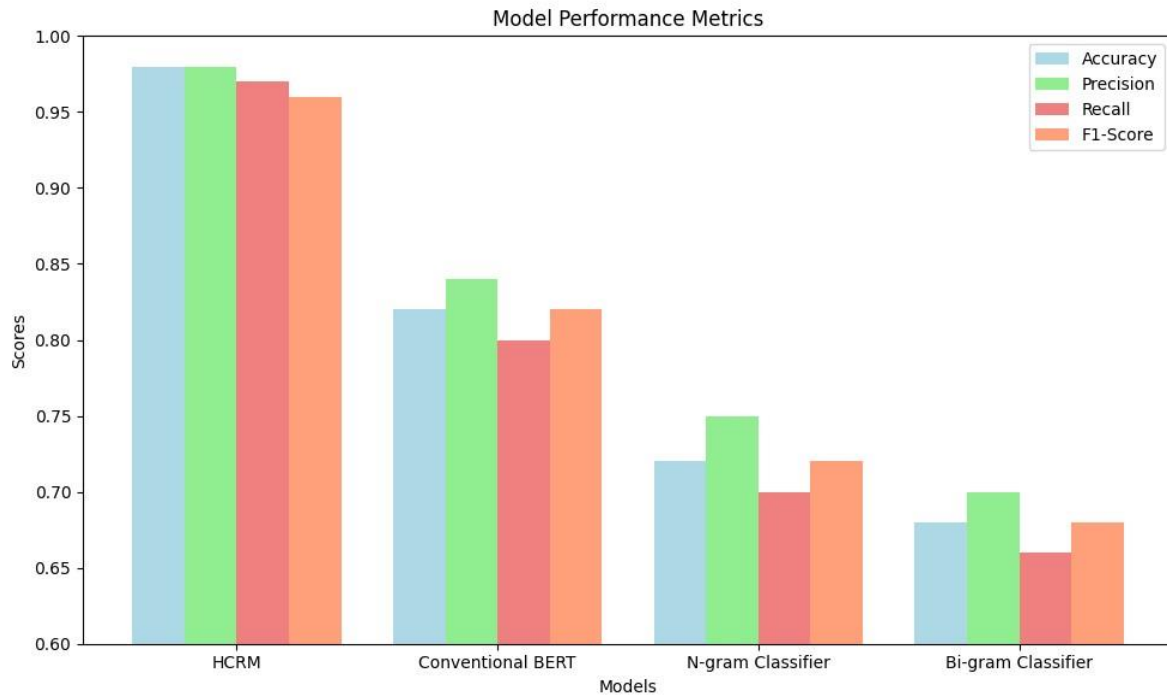
Model Performance Metrics

Figure 6: Comparative Analysis

The N-gram Classifier and Bi-gram Classifier models exhibit lower accuracy and performance metrics, with the N-gram Classifier having an accuracy of 0.72 and an F1-Score of 0.72, and the Bi-gram Classifier with an accuracy of 0.68 and an F1-Score of 0.68. These models are less accurate and balanced compared to HCRM and Conventional BERT. In table 7 illustrates the varying performance levels of different classification models, with HCRM outperforming the others, particularly in terms of accuracy, precision, recall, and F1-Score, making it the most effective model for the given classification task.

VIII.   CONCLUSION

Self-learning in English oral learning offers the freedom to learn at one's own pace, but it requires discipline and effective strategies. With the right tools and support systems in place, self-learners can achieve fluency and proficiency in English while gaining valuable skills for self-directed learning in other area. Hence, this paper introduces a novel and innovative approach to enhancing English oral learning by integrating the BERT model with the Hidden Condition Random Model (HCRM). The HCRM architecture effectively combines sentiment analysis, feature extraction, and classification to assess the relevance of oral learning materials based on the opinions of both students and teachers. This holistic approach allows for a more comprehensive evaluation of the quality of educational content, ensuring that it meets the requirements and expectations of both instructors and learners. The HCRM model proves to be effective in extracting nuanced sentiments from text data, enabling a more accurate assessment of the potential impact of oral learning materials. It is concluded that the propsoed model opens new avenues for improving the quality of English instruction by leveraging advanced natural language processing techniques.

REFERENCES

1. Mishra, A. (2021). Self-Learning System for Child Development Using Conversational AI and Natural Language Processing (NLP). In Impact of AI Technologies on Oral learning, Learning, and Research in Higher Education (pp. 124-133). IGI Global.
2. Fan, X., Cho, E., Huang, X., & Guo, E. (2021). Search based self-learning query rewrite system in conversational ai.
3. ZIRAPE, S., SHARMA, S., ALI, I. H., KUMBHAR, M., NALAWADE, R., & JANSARI, M. (2023). Add Self-Learning Ability to NLP for Automatic Test Case Generation.

4. Xiang, X., & Foo, S. (2021). Recent advances in deep reinforcement learning applications for solving partially observable markov decision processes (pomdp) problems: Part 1—fundamentals and applications in games, robotics and natural language processing. Machine Learning and Knowledge Extraction, 3(3), 554-581

5. Budiharto, W. I. D. O. D. O., Andreas, V., & Gunawan, A. A. S. (2021). A novel model and implementation of humanoid robot with facial expression and natural language processing (NLP). ICIC Express Letters, Part B: Applications, 12(3), 275-281.

6. El Essawy, E. S., & Dowydar, M. (2022). Effect of counseling program based on Neuro Linguistic Programming (NLP) On self-learning Skills and Psychological Resilience For students of the Faculty of Sport Education during novel Corona Virus pandemic. The Scientific Journal of Physical Education and Sports Sciences, 45(1), 197-237.

7. Wei, L., Li, Q., Song, Y., Stefanov, S., Siriwardane, E., Chen, F., & Hu, J. (2022). Crystal transformer: Self-learning neural language model for generative and tinkering design of materials. arXiv preprint arXiv:2204.11953.

8. Li, W., Ma, K., Qiu, Q., Wu, L., Xie, Z., Li, S., & Chen, S. (2021). Chinese Word Segmentation Based on Self-Learning Model and Geological Knowledge for the Geoscience Domain. Earth and Space Science, 8(6), e2021EA001673.

9. Joaquim, C. E. D. L., & Faleiros, T. D. P. (2022, June). BERT Self-Learning Approach with Limited Labels for Document Classification. In International Conference on Learning and Intelligent Optimization (pp. 278-291). Cham: Springer International Publishing.

10. Yurchenko, O., Cherednichenko, O., Trofimova-Herman, A., & Kupriianov, Y. (2023, April). Towards Cross-Lingual Transfer Based on Self-Learning Conversational Agent Model. In 6th International Conference on Computational Linguistics and Intelligent Systems (CoLInS 2023) (Vol. 3396, pp. 194-205).

11. Joaquim, C. E. D. L. (2022). BERT self-learning approach with limited labels for document classification of a Brazilian Army's administrative documentary set.

12. Neto, J. R. C., & Faleiros, T. D. P. (2021). Deep Active-Self Learning Applied to Named Entity Recognition. In Intelligent Systems: 10th Brazilian Conference, BRACIS 2021, Virtual Event, November 29–December 3, 2021, Proceedings, Part II 10 (pp. 405-418). Springer International Publishing.

13. Dong, X. (2023). Methods for Leveraging Auxiliary Signals for Low-Resource NLP (Doctoral dissertation, Rutgers The State University of New Jersey, School of Graduate Studies).

14. Fang, Q., Ye, R., Li, L., Feng, Y., & Wang, M. (2022). Stemm: Self-learning with speech-text manifold mixup for speech translation. arXiv preprint arXiv:2203.10426.

15. Fan, S., Liu, J., An, G., & Li, X. (2021, December). Research on data processing method of railway safety information based on NLP technology. In 2021 3rd International Academic Exchange Conference on Science and Technology Innovation (IAECST) (pp. 1304-1307). IEEE.

16. Alsafari, S., & Sadaoui, S. (2021, October). Semi-supervised self-learning for arabic hate speech detection. In 2021 IEEE International Conference on Systems, Man, and Cybernetics (SMC) (pp. 863-868). IEEE.

17. Dash, B., Swayamsiddha, S., & Ali, A. I. (2023). Evolving of Smart Banking with NLP and Deep Learning. In Enabling Technologies for Effective Planning and Management in Sustainable Smart Cities (pp. 151-172). Cham: Springer International Publishing.

18. Xu, L., Zhang, X., Zhao, X., Chen, H., Chen, F., & Choi, J. D. (2021). Boosting cross-lingual transfer via self-learning with uncertainty estimation. arXiv preprint arXiv:2109.00194.

19. Zhang, H., Chao, B., Huang, Z., & Li, T. (2022). Construction and Research on Chinese Semantic Mapping Based on Linguistic Features and Sparse Self-Learning Neural Networks. Computational Intelligence and Neuroscience, 2022.

20. Kazakova, M. A., & Sultanova, A. P. (2022). Analysis of natural language processing technology: modern problems and approaches. Advanced Engineering Research, 22(2), 169-176.

21. Nayak, S., Kanetkar, A., Hirudkar, H., Ghotkar, A., Sonawane, S., & Litake, O. (2022). Suggesting Relevant Questions for a Query Using Statistical Natural Language Processing Technique. arXiv preprint arXiv:2204.12069.

22. Gu, Z., Wang, Q., Li, F., & Ou, Y. (2021, November). Design of Intelligent QA for Self-learning of College Students Based on BERT. In ISCTT 2021; 6th International Conference on Information Science, Computer Technology and Transportation (pp. 1-5). VDE.

23. Müller, M., Salathé, M., & Kummervold, P. E. (2023). Covid-twitter-bert: A natural language processing model to analyse covid-19 content on twitter. Frontiers in Artificial Intelligence, 6, 1023281.

24. Xu, S., Zhang, C., & Hong, D. (2022). BERT-based NLP techniques for classification and severity modeling in basic warranty data study. Insurance: Mathematics and Economics, 107, 57-67.

25. Shahid, M. R., & Debar, H. (2021, December). Cvss-bert: Explainable natural language processing to determine the severity of a computer security vulnerability from its description. In 2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA) (pp. 1600-1607). IEEE.

26. Koroteev, M. V. (2021). BERT: a review of applications in natural language processing and understanding. arXiv preprint arXiv:2103.11943.

27. Wu, Y., Liu, Z., Wu, L., Chen, M., & Tong, W. (2021). BERT-Based Natural Language Processing of Drug Labeling Documents: A Case Study for Classifying Drug-Induced Liver Injury Risk. Frontiers in Artificial Intelligence, 4, 729834.

28. Nugroho, K. S., Sukmadewa, A. Y., & Yudistira, N. (2021, September). Large-scale news classification using bert language model: Spark nlp approach. In Proceedings of the 6th International Conference on Sustainable Information Engineering and Technology (pp. 240-246).

29. Liu, Z., Li, G., & Cheng, J. (2021, February). Hardware acceleration of fully quantized bert for efficient natural language processing. In 2021 Design, Automation & Test in Europe Conference & Exhibition (DATE) (pp. 513-516). IEEE.

30. Özçift, A., Akarsu, K., Yumuk, F., & Söylemez, C. (2021). Advancing natural language processing (NLP) applications of morphologically rich languages with bidirectional encoder representations from transformers (BERT): an empirical case study for Turkish. Automatika: časopis za automatiku, mjerenje, elektroniku, računarstvo i komunikacije, 62(2), 226-238.

31. Olaniyan, R., Stamate, D., & Pu, I. (2021). A Two-Step Optimised BERT-Based NLP Algorithm for Extracting Sentiment from Financial News. In Artificial Intelligence Applications and Innovations: 17th IFIP WG 12.5 International Conference, AIAI 2021, Hersonissos, Crete, Greece, June 25–27, 2021, Proceedings 17 (pp. 745-756). Springer International Publishing.

32. Donnelly, L. F., Grzeszczuk, R., & Guimaraes, C. V. (2022, April). Use of natural language processing (NLP) in evaluation of radiology reports: an update on applications and technology advances. In Seminars in Ultrasound, CT and MRI (Vol. 43, No. 2, pp. 176-181). WB Saunders.

33. Turchin, A., Masharsky, S., & Zitnik, M. (2023). Comparison of BERT implementations for natural language processing of narrative medical documents. Informatics in Medicine Unlocked, 36, 101139.

34. Qiu, Z., Wu, X., Gao, J., & Fan, W. (2021, May). U-BERT: Pre-training user representations for improved recommendation. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 35, No. 5, pp. 4320-4327).

35. Wang, R., Chen, D., Wu, Z., Chen, Y., Dai, X., Liu, M., ... & Yuan, L. (2022). Bevt: Bert pretraining of video transformers. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 14733-14743).

36. Perez, I., & Reinauer, R. (2022). The topological BERT: Transforming attention into topology for natural language processing. arXiv preprint arXiv:2206.15195.