



Exploring the Potential of Artificial Intelligence -Driven Assessment Tools for ESL Classrooms: Opportunities and Challenges

By

¹AKINTUNDE, Abraham Femi, (Ph.D.), ²OKECHALU, Emmanuel, ³CHUKWUEMEKA, Emeka Joshua (Ph.D.)

¹Department of Arts Education, Faculty of Education, University of Abuja ORCID ID: [0009-0000-3838-2655](https://orcid.org/0009-0000-3838-2655)

²Department of Arts Education, Federal Capital Territory, Secondary Education Board, Area 3, Abuja-Nigeria

³Department of Educational Foundations, Faculty of Education, University of Abuja ORCID ID: [0000-0002-1985-8002](https://orcid.org/0000-0002-1985-8002)



Article History

Received: 11/03/2025

Accepted: 22/03/2025

Published: 24/03/2025

Vol – 3 Issue – 3

PP: - 45-53

Abstract

This paper explores the transformative potential of Artificial Intelligence (AI)-driven assessment tools in English as a Second Language (ESL) classrooms. It provides an overview of core AI technologies, including machine learning, natural language processing (NLP), and deep learning, and their expanding applications in language assessment. The paper examines the evolution of language assessment, highlighting the limitations of traditional methods, and discusses the integration of AI to address these challenges. It delves into specific AI applications in ESL assessment, such as automated essay scoring (AES) employing techniques like Latent Semantic Analysis (LSA) and part-of-speech tagging, and automated spoken language evaluation, emphasizing the crucial roles of acoustic, language, and scoring models. The paper further explores the use of n-grams and intelligent tutoring systems. It analyzes the advantages of AI in ESL assessment, including increased efficiency, objectivity, consistency, and personalized feedback. However, it also addresses the constraints associated with AI integration, such as data privacy concerns, potential biases in algorithms, and the need for robust validation studies. The paper concludes that by strategically embracing AI, ESL classrooms can benefit from more efficient, effective, and fair language assessment systems that empower learners, educators, and institutions. Finally, the paper strongly recommends the establishment of ethical guidelines and standards for AI in language assessment to ensure data privacy, fairness, transparency, and accountability as AI becomes increasingly prevalent in ESL education.

Keywords: Artificial Intelligence, AI technologies, Language Assessment, Second Language and Assessment Tools.

Introduction

Language assessment is vital in various domains, including education, employment, and social integration. It measures individuals' language proficiency, determines their language learning outcomes, and assesses their ability to communicate effectively in multilingual contexts (Brown, 2023). Language assessments are used to make crucial decisions, such as educational placement, employment opportunities, and migration processes. Language assessment has traditionally relied on human evaluators, who assess and score language performance based on standardized criteria. However, this manual assessment approach has limitations, including subjectivity, inter-rater variability, and scalability issues. In recent years, the field of language assessment has witnessed a remarkable transformation with the integration of Artificial

Intelligence (AI) technologies. AI refers to developing computer systems that can perform tasks that typically require human intelligence, such as learning, reasoning, and problem-solving (Wooldridge, 2021). These technologies, including machine learning, natural language processing (NLP), and deep learning, have revolutionized language assessment, opening up new possibilities for more efficient, objective, and innovative evaluation methods (Wooldridge, 2021).

Understanding AI

Artificial Intelligence (AI) is a multidisciplinary field that focuses on developing computer systems that can perform tasks that usually require human intelligence (Zhai & Wibowo, 2023). It encompasses various subfields, including machine learning, natural language processing (NLP), and deep learning. These technologies have found significant

*Corresponding Author: AKINTUNDE, Abraham Femi, (Ph.D.)



applications in language assessment, transforming how assessments are conducted and enabling new possibilities for more efficient and accurate evaluation (Chapelle & Douglas, 2017). Machine learning is a branch of AI that allows computer systems to learn from data and improve their performance over time without being explicitly programmed. In language assessment, machine learning algorithms can be trained on large datasets of language samples, enabling them to identify patterns, linguistic features, and performance indicators automatically. It enables the development of AI models that can evaluate language proficiency and provide reliable scores (Zhai & Wibowo, 2023). Natural Language Processing (NLP) focuses on the interaction between computers and human language. It involves the development of algorithms and techniques to understand, analyze, and generate natural language text or speech. NLP has been instrumental in language assessment, enabling the processing and analysis of written essays, spoken responses, and other language samples. NLP techniques, such as sentiment analysis, syntactic parsing, and discourse analysis, can provide valuable insights into language usage and proficiency (Dodigovic, 2020). Deep learning is a subset of machine learning that utilizes artificial neural networks to model and simulate the human brain's structure and function. Deep learning algorithms can process and analyze large amounts of language data, allowing for more complex assessments. This technology has been particularly effective in speech recognition, language generation, and complex language understanding (Yu & Deng, 2015; Bello et al., 2024).

Language Assessment

Language assessment involves evaluating individuals' language skills in various areas, including reading, writing, listening, and speaking (Akintunde, 2024; Akintunde & Abdallah, 2024). Traditional language assessment methods often rely on standardized tests designed to measure language proficiency based on predetermined criteria. These tests are typically administered and scored by human assessors, who assess the quality and accuracy of language performance (Brown, 2023). However, traditional assessment methods have inherent limitations. They can be time-consuming, subjective, and prone to inter-rater variability. Human assessors may have different interpretations of assessment criteria, leading to consistency in scoring. Additionally, manual assessment processes can be resource-intensive and need more scalability, making it challenging to evaluate many language samples efficiently (Phongsirikul, 2018).

Integration of AI in Language Assessment

Integrating AI technologies in language assessment has brought significant advancements and benefits to the field. AI-based language assessment systems can automate and streamline assessment processes, providing more objective, reliable, and scalable evaluations (Dodigovic, 2020). By leveraging machine learning, NLP, and deep learning techniques, AI models can analyze language samples, identify linguistic features, and provide automated scoring and feedback (Farhady, 2019). Automated scoring systems powered by AI can evaluate written essays, spoken responses,

or other language samples. These systems can analyze linguistic aspects, including grammar, vocabulary, coherence, and argumentation, and provide objective and consistent evaluations. Automated scoring saves time and enables immediate feedback for learners, facilitating their language development and improvement (Farhady, 2019). Intelligent tutoring systems, another application of AI in language assessment, provide personalized instruction and feedback to language learners. These systems can adapt to learners' needs, identify their strengths and weaknesses, and tailor instruction accordingly. Intelligent tutoring systems can provide interactive and engaging learning experiences using AI technologies, enhancing learners' language skills (Shi et al., 2018). Natural language processing techniques have also been integrated into language assessment to analyze and understand language samples. Sentiment analysis, for example, can assess the emotional tone or sentiment expressed in written essays or spoken responses. Textual complexity and readability analysis can evaluate the difficulty level of written texts, ensuring appropriate language levels for learners (Chapelle & Chung, 2010). The integration of AI in language assessment has transformed the field, offering automated scoring and evaluation of intelligent tutoring system.

Automated Scoring of Writing

Automated essay scoring (AES) leverages AI algorithms to evaluate and score written responses, offering a scalable solution for handling large volumes of essays while ensuring quick and consistent feedback. This application of AI in language assessment has gained significant attention due to its ability to enhance efficiency, reduce human bias, and provide timely evaluations of writing proficiency (Ramesh & Sanampudi, 2021). AES systems utilize natural language processing (NLP) techniques to analyze key linguistic features, such as grammar, vocabulary, coherence, and argumentation. Machine learning models, trained on extensive datasets of manually scored essays, predict scores for new, unseen responses, aligning their assessments with established scoring rubrics. However, ensuring the reliability and validity of AI-based scoring remains crucial. Research continues to compare AI-generated scores with those assigned by human assessors, assessing their correlation to determine accuracy and consistency (Kumar & Boulanger, 2020). Studies also examine how well AI scoring aligns with standardized assessment frameworks to uphold fairness in evaluation (Kaldaras & Haudek, 2022).

AES models can be categorized as prompt-specific or generic (Chapelle & Chung, 2010). Prompt-specific models require a large sample of responses for each essay prompt to develop tailored scoring parameters, enabling a more detailed evaluation. Generic models, in contrast, apply fixed linguistic and structural criteria across various prompts, often relying on surface-level features such as grammar errors, punctuation, sentence complexity, and word frequency. While these models enhance efficiency, they must be carefully designed to balance predictive accuracy with construct validity. The development process involves identifying relevant linguistic variables, analyzing training essays, and selecting the most

*Corresponding Author: AKINTUNDE, Abraham Femi, (Ph.D.)

predictive features. Although essay length often serves as a strong predictor of writing proficiency, overreliance on it can compromise validity, prompting model developers to refine their algorithms to prevent unintended biases (Farhady, 2019). Ultimately, effective AES systems result from collaboration between computational linguists and content experts, ensuring both scoring accuracy and meaningful assessment outcomes (Chapelle & Chung, 2010).

Surface Features and Grammar Checkers

Project Essay Grade (PEG) was the earliest AES system, produced by Ellis Page in the late 1960's (Page, 2003). At the time, hand-written essays had to be entered manually into a mainframe computer for scoring. Page believed that writing skills could be measured indirectly through proxy traits (Farhady, 2019), so the computer scoring algorithm focused on quantifying surface linguistic features of a block of text (e.g. essay length, average word length, part of speech, count of punctuation, counts of grammatical function words such as prepositions and pronouns, lexical variation, etc.), without reference to semantic content of the essay. Using multiple regression, PEG produced a holistic score that correlated with human scores as high as $r = 0.87$ (Farhady, 2019). In 2002, the PEG system was acquired by Measurement, Inc. and developed into a web-based writing practice program. Between 300 and 500 intrinsic characteristics of writing are now linked to proxy traits like content, word variety, grammar, text complexity, sentence variety, fluency, diction, etc. Though, recent developments of PEG have included incorporating new parsers and improving classification schemes, a continuing criticism of this style of analysis is its immunity to the semantics, or meaning, of the passage (Shi et al., 2018).

Latent Semantic Analysis

Pearson's scoring engine, Intelligent Essay Assessor (IEA), evaluates meaning using a natural language processing technique called Latent Semantic Analysis (LSA) (Foltz, Streeter, Lochbaum and Landauer, 2013). With LSA, each word, sentence, and passage becomes a vector in relation to a multidimensional semantic space. Foltz et al. (2013, p. 78) provide the following examples:

Surgery is often performed by a team of doctors.

On many occasions, several physicians are involved in an operation.

Although the two sentences contain no words in common, their meanings are approximately the same based on the contexts of the words that comprise them. For example, the words "physicians" and "doctors" appear in similar contexts in English. In an LSA vector space, these two sentences would describe effectively the same "vector" because their underlying meaning is the same. Content-based scoring is enabled by a "background" model using an enormous corpus to evaluate a newly submitted essay or summary. When scoring prompt-specific traits, IEA compares the incoming essay with all known scored essays from the training set, and

determines the new essay's vector proximity (cosine) to other known score vectors in the semantic space.

LSA variables are not only used to predict content, word choice and task completion, but also organizational traits such as sentence fluency and essay coherence. If sentences follow one another logically, the sentence vectors will be in close proximity. The same is true at the paragraph level. Thus, proximity of sentence vectors in a paragraph can be used to predict expert human judgments of, for example, writing coherence. IEA supports numerous assessments, including the Pearson Test of English - Academic (Pearson, 2009) and the English for Professionals Exam (Pearson 2013), as well as writing practice tools such as WriteToLearn and Summary Street. In addition to LSA, IEA uses n-grams and grammar, usage, and mechanics (GUM) to achieve correlations with human judgments between 0.80 and 0.86 (Landauer, Laham & Foltz, 2013; Pearson, 2013).

Part of Speech Categorizing

The Educational Testing Service (ETS) e-rater system draws heavily on part of speech tagging for scoring capabilities (Attali & Burstein, 2016). This approach assigns part of speech (POS) tags to each word in the text, e.g. noun, verb, etc. In e-rater, three modules operate on each submitted essay: a syntactic parser, a discourse analyzer, and a topical analysis module that operates in a vector space (Burstein, 2003). From the output of these modules, e-rater identifies a set of micro-features including the presence (and counts) of errors of various types. These micro-features are aggregated into a set of features that fall into 8 broad categories: grammar, usage, mechanics (GUM), style, organization, development, lexical complexity, and topic-specific vocabulary usage (i.e. content) (Enright & Quinlan, 2010). E-rater uses multiple regression, weighted by quantitative evaluation of features (e.g. proportion of grammar errors, proportion of usage errors, etc.) found in a set of human-scored training essays, to predict a final holistic score representing the average score produced by two human judges. Scoring models may be prompt-specific or generic, and both model types strongly predict scores assigned by human judges (Attali & Burstein, 2016). Even with a generic scoring model, the correlation between human and e-rater scores ($r = 0.76$) is comparable or slightly higher than between paired human raters ($r = 0.70$) (Attali 2021). E-rater provides scoring of the Graduate Management Achievement Test (GMAT) and one of two scores on the writing portion of the Test of English as a Foreign Language (TOEFL) iBT; the second score is assigned by a human expert. Scores are produced from 0 to 6. If there is a discrepancy between the scores of more than 1.5 point, the passage is sent to a second human for arbitration (Enright & Quinlan, 2010).

N-grams

An n-gram is a multi-item written unit (syllables, letters, or words); in the context of automated writing scoring, it typically refers to words. N-gram generically includes the set unigram (e.g. "beautiful"), bigram (e.g. "beautiful day"), trigram ("a beautiful day"), etc. When analyzing a large corpus, a parser can break the text into a series of overlapping

n-grams, and the frequency of all n-grams in the corpus can be obtained. The presence or distribution of n-grams in the sample can be used in scoring writing. Often, lower frequency n-grams are associated with higher quality. For example, an essay about "a beautiful day", a frequent trigram, is fine albeit average, but an essay about "a gorgeous day", a less frequent trigram, is probably written by a more skilled student. However, probabilistic models looking only n-grams are insufficient: "a day beautiful" is also infrequent but should not be associated with a higher score. N-gram analyses are often combined with LSA or POS approaches to help determine the accuracy or appropriacy of word sequences in the student's writing.

Automated Scoring of Speaking

While automated systems for evaluation and remediation of pronunciation have been available for some time (e.g. Franco, Harry, Romain, Venkata, Rao, Elizabeth, Victor & Kristin, 2010), systems which evaluate the wider set of competencies required for spoken communication are now becoming more common. These systems must evaluate 'what' was said as well as 'how' it was said. Automated scoring of spoken language tests requires three important models to be developed: acoustic model; language model; and scoring model. The construction of these three models, together with test and task design, makes accurate automated scoring possible. Therefore, before describing test providers and the differences in their scoring models, this section elaborates on the three kinds of models employed in automated scoring of speech.

Acoustic Model

The acoustic model is the main component of the speech recognizer. Speech recognition software applies hidden Markov models (HMMs) to represent each phone, or sound (Young, 2021). Recognition can be thought of as a series of probabilities; the model estimates the likeliest phoneme or word from among a number of possibilities. Spectral features are extracted for each 10-millisecond frame of the speech, and a model associates probabilities for each possible phone. The output is the best statistical estimate of which words were spoken, in the form of a transcript. Acoustic models must be "trained" or optimized on a set of speech data. The training process involves pairing the audio speech with transcriptions of that speech, so that the model associates sounds with orthographic representations. Any speech data can be used for training, including radio broadcasts or audio from Youtube (Hinton, Li, Dong, George, Abdel-rahman, Navdeep, Andrew, Vincent, Patrick, Tara, & Bria, 2012). However, the acoustic model should be trained on speech that matches the speech to be recognized. Background noise, microphone type, and resolution of the speech signal will all impact the quality of the training data. Speakers' accents are also a crucial factor, as each potential test taker demographic may have a distinct pronunciation repertoire. Thus, it is best to gather speech data and train the acoustic model on a sample that matches the target population of the test.

Language Model

Next, the language model comprises words that are likely to be spoken in the response. The model is made of frequencies for n-grams. Thus, if the language task is to describe a picture of a girl eating an apple, the bigram "an apple" and the trigram "eating an apple" are likely to appear frequently in test-taker responses. To ensure a representative sample of responses, test items are typically trialed on the target population. For example, in the PTE Academic, each item was presented to over 300 learners in field-testing and their responses were recorded (Pearson, 2009). Using this prior information, the language model assists in the assigning of probabilities in the word recognition process. Speech is easier to recognize if the model can anticipate what the speaker may say. Thus, tasks such as reading aloud or repeating sentences can result in word recognition accuracy well above 90%, even in accented speech (Balogh, Jared, Jian., Alistair, & Masanori, 2012). On the other hand, recognizing spontaneous speech can be very difficult. Chen, Zechner & Xi (2019) report recognition accuracy at 34% (one in every three words recognized accurately) on spontaneous speech from learners of English in the TOEFL iBT. By comparison, the gold standard for recognition of spontaneous native-speaker speech under optimal conditions and with ample computing power is above 80%, depending on the speech to be recognized (Hinton et al. 2012).

Scoring Model

The scoring model refers to the method for selecting features from the speech recognition process and applying them to predict human ratings. In language assessment, the test developer should consider not only features of the output which are the best predictors of the human score, but also which predictors are *relevant* to the human score. For example, imagine that a duration measure such as speech rate were a strong statistical predictor of human judgments of grammar ability. If neither the rating scales nor the human raters paid attention to speech rate in evaluation of grammar ability, then it would not be responsible to use this measure in the scoring model. The Versant tests exemplify several kinds of scoring models (Bernstein, Van-Moere & Cheng, 2010). Pronunciation models are developed from spectral properties that are extracted during the recognition process. First, the rhythmic and segmental aspects of a performance are defined as likelihoods with reference to a set of native speaker utterances, and then these measures are used in non-linear models to predict human judgments of pronunciation. Fluency models are similarly developed from durations of events, such as response latency, words per time, segments per articulation time, and inter-word times. Sentence mastery is mainly derived from the number of word errors given in a response and scaled using a partial credit Rasch model. Vocabulary items which elicit single words or phrases simply use the recognized output to ascertain whether the response was correct or incorrect, and the result is entered into dichotomous Rasch models. In constructed responses, such as retelling a story, the recognized output is used in a Latent Semantic Analysis approach to predict human judgments of vocabulary

*Corresponding Author: AKINTUNDE, Abraham Femi, (Ph.D.)

coverage. Thus, the Versant tests attempt to preserve construct validity by using distinct, relevant features in models to predict each trait separately, and then combine the subscores to create an overall score.

A very different approach to scoring models using TOEFL iBT's SpeechRater 1.0 was described by Xi, Derrick, Klaus, and Davi (2018). Numerous features associated with pronunciation, fluency, and vocabularies were entered into a regression model to predict holistic scores on the speaking tasks, rather than individual traits. Then, a committee evaluated and selected the final features, taking into account the construct representation of each feature (i.e. its linkage to the rating scale) and the strength of its relationship with human scores. Based on the data, the committee advised on overweighting fluency variables which were the strongest predictors, and under-weighting grammar and vocabulary which were weakest for predicting human ratings. Interestingly, durational variables associated with fluency are the best-understood and measurable aspects of spoken proficiency (Cucchiari, Strik & Boves 2012). In Xi et al.'s model (2018), words-per-second is the best predictor of all, and correlated with human holistic scores at 0.49.

Machine Performance

Having developed these three models, how well can machine scores be expected to predict human scores for unseen (or new) spoken responses? Any evaluation of machine-to-human correspondence must be interpreted in light of the human rater reliability as this is the standard that the machine is measured against. The Versant tests report machine-to-human correlation of 0.97 for overall scores and 0.88-0.97 at the trait level (Pearson, 2009). Human split-half reliability is 0.99 for overall scores and 0.93-0.99 at the trait level. This is based on validation data consisting of a flat (rather than normal) score distribution that spans the entire score scale. Operationally, performance is somewhat lower; Bernstein, Van-Moere and Cheng, (2010) report overall machine-to-human correlation of 0.94, and split-half reliability 0.84-0.88 at the trait level, based on a class of English major undergraduates at a Hong Kong university. Xi et al.'s (2008) study using SpeechRater 1.0 reports machine-to-human correlation of 0.57 for the spoken section of the TOEFL iBT test, where human agreement was 0.74. This is with a dataset with restricted score distribution. Using test-taker responses from the TOEFL iBT field study which had a wider score distribution more representative of operational testing, the correlation was 0.68, where human agreement was 0.94. Trait level scores were not computed separately, as the scoring model combined them into one holistic model, as described above.

Differences in reliability and machine-to-human correlation among tests can be partly explained by the different modeling employed, and partly explained by the test and task design. The Versant test and the PTE Academic were both designed with automated scoring, as well as construct representativeness in mind. They present approximately 62 and 36 items respectively, where the item-types elicit a mix of

constrained speech (read aloud and sentence repeats) and constructed speech which is controlled in terms of output (describing a picture, retelling a lecture or story). The mix of item-types plays to strengths of the automated system: pronunciation can be measured more accurately from constrained speech than from unconstrained speech (Chen, Zechner & Xi, 2019); vocabulary, language use and communication effectiveness can be measured more appropriately from constructed or communicative speech tasks. In contrast, the TOEFL iBT presents six items that elicit long turns. The speech exhibits degrees of spontaneity - some items elicit opinions on a given topic, some items are in response to stimulus. Because the speech is unpredictable and has high perplexity the recognition performance is low, and so the scoring relies heavily on fluency measures irrespective of the content (vocabulary and language use) in the speech.

When enquiring about automated scoring of spoken language, the question is often asked, "What kind of speech recognition software do you use?" However, it should be clear after reading the preceding sections that the actual speech recognition software is of little importance. Rather, the important variables are: the type of speech data used in the optimization of the acoustic models; the predictability and perplexity of the speech elicited and the resulting usefulness of language models; the quality and reliability of the human ratings which the models are developed to predict; and the features which are extracted from the recognition process, and how the features are combined and formulated to predict human ratings.

Automated Scoring and Evaluation

Automated scoring and evaluation are crucial aspect of language assessment that utilizes Artificial Intelligence (AI) technologies to analyze and assess various language skills. It encompasses multiple domains: automated essay scoring, spoken language evaluation, and grammar and vocabulary assessment. These applications of AI in language assessment have revolutionized the evaluation process, offering efficient, objective, and consistent results (Mizumoto & Eguchi, 2023).

Spoken Language Evaluation

The evaluation of spoken language skills is another area where AI has made significant contributions. AI technologies, such as speech recognition and natural language processing, are vital in assessing spoken language proficiency. Speech recognition algorithms convert spoken language into text, enabling the analysis of various linguistic aspects, such as pronunciation, fluency, and intonation. Natural language processing techniques then process the transcribed text to evaluate language accuracy, vocabulary usage, and coherence in spoken responses (Zhai & Wibowo, 2023). Spoken language evaluation using AI offers numerous benefits, including scalability, objectivity, and consistency. By automating the assessment process, AI efficiently evaluates a large number of spoken language samples. AI systems apply predefined scoring criteria consistently, eliminating human biases and ensuring fairness in evaluation. Additionally, AI-based spoken language evaluation provides immediate

feedback to learners, enabling them to identify and address areas of improvement in their spoken language skills (Jones, Laxton, & Galaczi, 2021).

Grammar and Vocabulary Assessment

AI technologies have also been applied to assess grammar and vocabulary proficiency in language assessment. Automated error detection and correction systems utilize AI algorithms to analyze written or spoken language samples and identify grammatical and vocabulary errors. These systems employ rule-based approaches, statistical methods and machine-learning techniques to detect errors and suggest corrections. Grammar and vocabulary assessment using AI offers several advantages. It allows for efficient and objective evaluation of language skills, particularly in identifying common errors and providing targeted feedback to learners. Automated error detection and correction systems can help learners improve their language accuracy and expand their vocabulary by offering suggestions for error correction and alternative word choices. These AI-driven tools complement traditional grammar and vocabulary instruction, providing learners personalized support in their language learning journey (Yu & Deng, 2015). Automated scoring and evaluation encompass various applications of AI in language assessment including automated essay scoring, spoken language evaluation, and grammar and vocabulary assessment. These applications have transformed the evaluation process by offering efficiency, objectivity and consistency. AI-based systems provide timely feedback, handle large volumes of assessments and assist learners in improving their language skills. Ongoing research and development in automated scoring and evaluation continue to enhance the reliability and effectiveness of these AI-driven tools, empowering educators and learners in language assessment.

Natural Language Processing in Language Assessment

Natural Language Processing (NLP) is a field of AI that focuses on the interaction between computers and human language (Jones et al., 2021). NLP is crucial in analyzing and processing written texts in language assessment to provide valuable insights and assessments. There are two critical applications of NLP in language assessment: sentiment analysis and textual complexity and readability assessment.

1. Sentiment Analysis

Sentiment analysis uses AI and NLP techniques to analyze and interpret the sentiment and emotions expressed in written texts. In language assessment, sentiment analysis can provide valuable insights into the emotional tone, attitudes, and opinions in essays, reviews, or other written responses (Kennedy, 2012). AI algorithms for sentiment analysis are trained on large datasets of annotated texts, allowing them to identify and classify sentiments as positive, negative, or neutral. These algorithms employ various NLP techniques, including text classification, machine learning, and deep learning approaches, to accurately identify sentiment-bearing words, phrases, and contextual cues (Lund, 2020).

2. Textual Complexity and Readability Assessment

Textual complexity and readability assessment involve evaluating written texts' difficulty level, comprehension requirements, and linguistic features. AI techniques, particularly NLP, enable automated assessment of the complexity and readability of texts, providing valuable information for language assessment purposes (Kotani, Yoshimi, & Sahara, 2021). AI algorithms can analyze various linguistic features, including sentence structure, vocabulary richness, syntactic complexity, and discourse organization, to assess the complexity of written texts. These algorithms utilize computational models and statistical approaches to determine the appropriate language levels and readability of texts, considering factors such as word frequency, sentence length, and cohesion. Textual complexity and readability assessment using AI provide valuable insights for educators, learners and assessment designers. It helps identify texts that align with specific language proficiency levels, ensuring that learners are exposed to appropriate materials that challenge and support their language development. Additionally, it assists in creating reading materials that cater to the diverse needs of learners, considering factors such as age, language background and reading ability (Kotani et al. 2021).

The application of NLP in textual complexity and readability assessment contributes to more informed and data-driven language assessment practices. By automating the assessment process, AI-powered systems can efficiently handle large volumes of text, saving educators and assessment administrators' time and effort. It is important to note that while AI-based sentiment analysis and textual complexity assessment offer valuable insights, they are not infallible. Contextual & cultural understanding can sometimes pose challenges for automated systems. Therefore, human oversight and interpretation remain essential in language assessment to ensure accurate evaluations (Burstein et al., 2014). Natural language processing (NLP) plays a significant role in language assessment, particularly in sentiment analysis, textual complexity, and readability assessment. NLP techniques enable the analysis of sentiments and emotions expressed in written texts, providing deeper insights into the writer's language proficiency and communication skills. Additionally, NLP facilitates the automated assessment of textual complexity and readability, ensuring appropriate material selection and tailored instruction for learners. Continued advancements in AI and NLP technologies will further enhance language assessment's accuracy, efficiency, and effectiveness, supporting educators, learners, and assessment designers in their language development endeavors.

Advantages of Using AI in Language Assessment

1. Enhanced Efficiency

One of the key advantages of using AI in language assessment is the significant improvement in efficiency. AI-powered systems can process and evaluate a large volume of assessments relatively quickly (Kanase-Patil, Kaldate, Lokhande, Panchal, Suresh, & Priya, 2020). Automated

scoring and evaluation eliminate the need for manual grading, saving time and resources for educators and administrators. Additionally, AI algorithms can generate prompt feedback, providing learners with timely insights to improve their language skills. This enhanced evaluation and feedback generation efficiency streamlines the language assessment process and supports faster learning progress.

2. Increased Objectivity

AI brings increased objectivity to language assessment. Human grading can be influenced by personal biases or subjectivity, leading to inconsistencies in evaluation. AI-based systems use predefined criteria and algorithms to assess written or spoken language proficiency, eliminating subjective biases. By employing a standardized approach, AI ensures more objective and consistent evaluations. This objectivity enhances the credibility and fairness of language assessment outcomes, providing a level playing field for all test-takers (Chapelle & Chung, 2010).

3. Scalability

AI enables scalable language assessment, accommodating a more significant number of test-takers efficiently. Traditional language assessments often face challenges handling a high volume of assessments, especially during peak periods. AI-powered systems can process assessments simultaneously, allowing for quick and accurate evaluation of multiple test-takers. This scalability is particularly beneficial in educational institutions or language testing centers where many students or individuals require language assessment services. AI's ability to handle scalability ensures that language assessment remains accessible and manageable. AI-generated analytics provide valuable diagnostic insights into individual strengths and weaknesses in language proficiency. AI algorithms can identify specific areas where learners excel or struggle by analyzing a wide range of linguistic features and patterns. These insights go beyond a simple score or evaluation and offer a detailed breakdown of the learner's performance, highlighting specific language skills that require improvement. This diagnostic information enables educators and learners to tailor instruction and learning resources to address individual needs effectively. By focusing on targeted areas of improvement, learners can enhance their language skills more efficiently and effectively.

Constraint of Using AI in Language Assessment

1. Data Privacy and Security

The use of AI in language assessment involves collecting and analyzing large amounts of data. It is crucial to address the ethical implications surrounding data privacy and security. Institutions and organizations must ensure that appropriate measures are in place to protect the personal information of test-takers. Transparency and consent should be prioritized when collecting and utilizing data, adhering to relevant data protection regulations. Data storage and transmission should also be secured to prevent unauthorized access or breaches.

2. Fairness and Bias

While AI systems can reduce subjective biases, there is a risk of introducing or perpetuating new ones. Language assessment algorithms should be carefully designed and validated to minimize unfair advantages or disadvantages based on gender, ethnicity, or cultural background. Regular monitoring and audits of AI systems are necessary to detect and address potential biases. Inclusivity and fairness should be fundamental considerations throughout developing and deploying AI-driven language assessment tools (Edwards, Edwards, Spence, & Lin, 2018).

3. Human-AI Collaboration

It is essential to recognize the role of human experts alongside AI systems in language assessment. While AI can provide automated scoring and evaluation, human expertise is still crucial in interpreting complex language details. Addressing context-specific challenges and making subjective judgments in some instances. Collaboration between human experts and AI systems can lead to more accurate and reliable assessments. Clear guidelines and protocols should be established to facilitate the interaction and collaboration between human assessors and AI algorithms, ensuring a seamless integration of their strengths and expertise (Cope & Kalantzis, 2015). By addressing the challenges and considerations mentioned above, the application of AI in language assessment can be optimized to ensure data privacy and security, promote fairness and inclusivity, and harness the combined strengths of human expertise and AI technology. These considerations will contribute to AI's responsible and effective use in language assessment, benefiting learners and providing reliable and valuable insights into their language proficiency.

Conclusion

The power of AI in language assessment is vast and has the potential to revolutionize the field. Continued research and development in this area are essential to refine AI algorithms, address ethical considerations and maximize the benefits of AI in accurately measuring language proficiency. By embracing AI we can create more efficient, effective and fair language assessment systems that benefit learners, educators and institutions alike.

Recommendations

As AI becomes more prevalent in language assessment, it is essential to establish ethical guidelines and standards. These guidelines should address data privacy, fairness, transparency, and accountability concerns. Institutions and organizations involved in language assessment should develop clear policies and procedures for the responsible and ethical use of AI. Collaborative efforts among researchers, educators, policymakers, and relevant stakeholders can contribute to developing comprehensive ethical guidelines and standards that ensure the responsible deployment of AI in language assessment.

References

- Akintunde, A. F., & Abdallah, F. S. (2024). Blended Learning in a Language Classroom: Implication for Pedagogy. *European Journal of Contemporary Education and E-Learning*, 2(6), 200-211. [https://doi.org/10.59324/ejceel.2024.2\(6\).12](https://doi.org/10.59324/ejceel.2024.2(6).12)
- Akintunde, A. F. (2024). Exploring some artificial intelligence technologies that can be applied to learning in English as a Second Language Classroom. *Multidisciplinary Journal of Arts and Language Education* 5(1), 166-176.
- Attali, Y. & Burstein, J. 2016. Automated essay scoring with e-rater® V.2. *Journal of Technology, Learning, And Assessment*, 4(3). <http://www.jtla.org> (accessed 22 October 2013).
- Attali, Y. (2021). Automated subscores for TOEFL iBT® independent essays. *ETS research Report*. RR-11- 39. Princeton, NJ: Educational Testing Service.
- Balogh, J., Jared, B., Jian, C., Alistair, V, & Masanori, S. (2012). Validation of automated scoring of oral reading. *Educational and Psychological Measurement* 72(3). 435-452.
- Bello, V. I., Chukwuemeka, E. J., & Dr Ohiare-Udebu, M. F. (2024). Integrating Phonological Features and Technology in Designing a Comprehensive English Diction Curriculum for Effective Learning. *Global Scientific and Academic Research Journal of Education and literature*, 2(9), 16-22.
- Bernstein, J., Alistair, V. M. & Jian, C.. (2010). Validating automated speaking tests. *Language Testing* 27(3). 355-377.
- Burstein, J., Shore, J., Sabatini, J., Moulder, B., Lentini, J., Biggers, K., Holtzman, S. (2014). From Teacher Professional Development to the classroom: How NLP technology can enhance teachers' linguistic awareness to support curriculum development for English language learners. *Journal Educational Computing Research*. 5 (1). 119 -144. <https://doi.org/10.1080/03634523.2018.1502459>.
- Brown, H. D. (2023). *Language assessment principles and classroom practices*. San Francisco, California: Longman.
- Chapelle, C. A. & Chung, Y. R. (2010). The promise of NLP and speech processing technologies in language assessment. *Language Testing*. 27 (3), 301-315. <https://doi.org/10.1177/0265532210364405>.
- Chapelle, C. A. & Douglas, D. (2017). *Assessing language through computer technology*. United Kingdom: University Press. Cambridge.
- Chen, L., Klaus, Z. & Xiaoming, X.. (2019). Improved pronunciation features for construct-driven assessment of non-native spontaneous speech. *Proceedings of human language technologies: The 2009 annual conference of the North American chapter of the association for Computational Linguistics*, 442-449. Stroudsburg: Association for Computational Linguistics.
- Cope, B. & Kalantzis, M. (2015). Sources of evidence-of-learning: Learning and assessment in the era of big data. *Open Review Of Educational Research*. 2 (1), 194-217. <https://doi.org/10.1080/23265507.2015.1074869>.
- Dodigovic, M. (2020). Natural Language Processing (NLP) as an instrument of raising the language awareness of learners of English as a Second Language. *Language Awareness*, 12 (3-4), 187-203. <https://doi.org/10.1080/09658410308667076>.
- Edwards, C., Edwards, A., Spence, P. R., & Lin, X. (2018). I Teacher: Using artificial intelligence (AI) and social robots in communication and instruction. *Communication Education*. 67 (4), 473-480. <https://doi.org/10.1080/03634523.2018.1502459>.
- Enright, Mary K. & Thomas Quinlan. (2010). Complementing human judgment of essays written by English language learners with e-rater scoring. *Language Testing* 27(3), 317-334.
- Farhady, H. (2019). Language assessment: A linguametric perspective. *Language Assessment Quarterly* https://doi.org/10.1207/s15434311laq0202_3.
- Franco, H., Harry, B., Romain, R., Venkata, R., Rao, G., Shriberg, E., Abrash, V., & Kristin, P., (2010). EduSpeak®: A speech recognition and pronunciation scoring toolkit for computer-aided language learning applications. *Language Testing* 27(3). 401-418.
- Foltz, Peter W., Lynn A. Streeter, Karen E. Lochbaum & Thomas K Landauer. (2013). Implementation and applications of the Intelligent Essay Assessor. In Mark D. Shermis & Jill Burstein (eds.), *Handbook of automated essay evaluation: Current applications and new directions*. New York, NY: Routledge.
- Hinton, G., Li, D., Dong, Y., George, D., Abdelrahman, M., Navdeep, J., Andrew, S., Vincent, V., Patrick, N., Tara, & Bria, K. (2012). Deep neural networks for acoustic modeling in speech recognition. *IEEE Signal Processing Magazine* 29(6). 8297.
- Jones, E., Laxton, V., & Galaczi, E. (2021). Assessing L2 English speaking using automated scoring technology: Examining automarker reliability. *Assessment in Education: Principles, Policy & Practice*. 28 (21), 411-436. <https://doi.org/10.1080/096594X.2021.1979467>.
- Kaldaras, L., Haudek, K. C. (2022). Validation of automated scoring for learning progression-Aligned next generation science standards performance assessments. *Frontiers in Education*. <https://doi.org/10.3389/educ.2022.968289>.
- Kanase-Patil, A. B., Kaldate, A. P., Lokhande, S. D., Panchal, H., Suresh, M., & Priya, V.(2020). A review of artificial intelligence-based optimization

- techniques for the sizing of integrated renewable energy systems in smart cities. *Environmental Technology Reviews*. 9 (1), 111-136. <https://doi.org/10.1080/21622515.2020.1836035>
24. Kennedy, H. (2012). Perspectives on Sentiment Analysis. *Journal of Broadcasting and Electronic Media*. 56 (4), 435-450. <https://doi.org/10.1080/08838151.2012.732141>
 25. Kotani, K., Yoshimi, T., & Isahara, H. (2021). A machine learning approach to measurement of text readability for EFL learners using various linguistic features. *David publishing*. <https://files.eric.ed.gov/fulltext/ED529383.pdf>.
 26. Landauer, Thomas, Darrell Laham & Peter Foltz. (2013). Automated scoring and annotation of essays with the Intelligent Essay Assessor. In Mark D. Shermis & Jill Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective*. Mahwah, NJ: Lawrence Erlbaum Associates.
 27. Lund, B. D. (2020). Assessing library topics using sentiment analysis in R& A discussion and code sample. *Public Services Quarterly*. 16 (2), 112-123. <https://doi.org/10.1080/15228959.2020.1731402>.
 28. Mizumoto, A., Eguchi, M., (2023). Exploring the potential of using an AI language model for automated essay scoring. *Research Method in Applied Linguistic*. 2 (2). <https://doi.org/10.1016/j.rmal.2023.100050>.
 29. Page, E. B. (2003). Project essay grade: PEG. In Mark D. Shermis & Jill Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective*, Mahwah, NJ: Lawrence Erlbaum Associates.
 30. Pearson. (2009). *Versant English test: Test description and validation summary*. Menlo Park, CA: Pearson. <http://www.versanttest.com/technology/VersantEnglishTestValidation.pdf>
 31. Pearson. (2013). *English for professionals exam*. Menlo Park, CA: Pearson. http://www.euproexam.com/pdfs/euro_validation_report.pdf
 32. Phongsirikul, M. (2018). Traditional and alternative assessments in ELT: Students and teacher's perceptions. *rEFLECTIONS*. 25 (1).
 33. Ramesh, D., & Sanampudi, S. K., (2021). An automated essay scoring systems: A systematic literature review. *Artificial intelligence Review An International Science and Engineering Journal*. 55, 2495-2527. <https://doi.org/10.1007/s10462-021-10068-2>.
 34. Shi, Genghu., Lippert, A. M., Shubeck, K., Fang, Y., Chen, S., Jr. P. P., Greenberg, D., Graesser, A. C., (2018). Exploring an intelligent tutoring system as a conversation-based assessment tool for reading comprehension. *Behavioral Metrics*. <https://doi.org/10.1007/s41237-018-0065-9>.
 35. Wooldridge, M. (2021). *A brief history of artificial intelligence: What it is, where we are, where we are going*. Great Britain, UK: Flatiron Books.
 36. Xi, U., Derrick, M., Klaus, S. K., & Davi, I. P. (2018). Automated scoring and feedback systems: Where are we and where are we heading? *Language Testing* 27(3). 291-300.
 37. Yu, D. & Deng, L. (2015). *Automatic speech recognition A deep learning approach*. London: Springer-Verlag London.
 38. Zhai, C., & Wibowo, S. (2023). *A systematic review on artificial intelligence dialogue systems for enhancing English as Foreign Language students*. Switzerland: Springer