Bias Detection and Fairness Optimization in Nlp-Based Language Assessment Systems

Kazi Siam Al Mobin[1], Afroza Riju[2]

1. Department of Information Technology, Washington University of Science and Technology, United States
2. Department of Computer Science and Engineering, Green University of Bangladesh, Bangladesh

ABSTRACT

Natural Language Processing (NLP) has reshaped the educational assessment by the creation of automated language assessment systems like Automated Essay Scoring (AES). These systems provide scalable, economical, and reliable assessment of student writing and are becoming more popular in large-scale education testing and standardized tests. Nonetheless, questions have begun to arise about the bias and fairness of these algorithms and especially when it comes to assessing essays written by students with a variety of linguistic, cultural, and socio-economic backgrounds. Systematic differences in the scoring results might be caused by biascaster in training data, model structure, or assessment process and favor non-native speakers or underrepresented groups. This paper explores the notion of detecting bias and optimization of fairness in the NLP-based language evaluation systems. An elaborate experimental design is suggested to assess bias at both demographic and linguistic levels with the help of existing fairness measures and explainable artificial intelligence (XAI) approaches. Moreover, several fairness optimization strategies such as data balancing, adversarial debiasing, and bias-conscious learning strategies are considered to enhance the fairness of automated scoring systems without the need to compromise the predictive power. The results indicate both the existence of quantifiable bias in a number of popular NLP scoring models and that interventions that aim to promote fairness can be used to decrease the number of disparities in the scoring results to an important extent. The research is a contribution to responsible AI practices in educational technology and can be useful in designing equitable and transparent automated language evaluation systems.

Keywords: Automated Essay Scoring; Algorithmic Bias; Fairness in AI; Natural Language Processing; Educational Assessment

INTRODUCTION

1.1 Background

Artificial intelligence (AI) and Natural Language Processing (NLP) have quickly altered most aspects of education, especially the assessment of written language. Automated language assessment systems, also known as Automated Essay Scoring (AES) systems, have become more and more prominent in academic as well as standardized testing settings. Such systems use machine learning and deep learning to score written answers and rate them similarly to human raters. Preliminary automated grading systems like Project Essay Grade (PEG) showed that language patterns could be detected with computational models that could be used to retrieve language patterns related to writing quality, and grading of essays was reliable [1].

Subsequently, NLP and machine learning have evolved over the years to provide much more functionality to automated assessment systems, enabling them to handle large volumes of data and mark essays more often and more consistently. The adoption of the automated scoring systems has increased significantly since it is capable of saving time in grading, scoring consistency, and allowing assessment on mass scale in learning institutions. Educational Tests Service (ETS) has developed systems like the e-rater that has been very common in examination standardization to facilitate human grading systems [2]. In online learning, smart tutoring systems and massive learning platforms, automated essay grading has also been used, and students receive instant feedback [3]. As the digital education sector and online testing platforms expand, the use of NLP-based assessment tools is gradually becoming an inseparable part of the contemporary educational entertainment systems. The formation of AES models has been transformed into advanced forms of deep learning models, instead of the classical methods of machine learning. The previous models were based on linguistic features which were created manually (e.g. grammar, richness of vocabulary, complexity of a syntax, the length of an essay) [3].

These were normally applied in the statistical models like regression, support vector machines, or decision trees. Nevertheless, neural network models, like Long Short-Term Memory (LSTM) networks and convolutional neural network (CNNs) have proven better performance with the advent of deep learning and automatically learning more intricate linguistic features of textual input [4,5]. Transformer-based models like BERT, more recently, have extended the capabilities of NLP since they allow understanding languages in their context and better perform in various language processing tasks [6]. With such technological developments, there has been an increasing concern about the equity and impartiality of AI-based systems of assessment. Algorithms that decide on matters can unwillingly reproduce or increase the biases found in the data on which they are trained. In the learning setting, it may result in systematic differences in scoring the results of students with different linguistic, cultural, or demographic backgrounds. The researchers have pointed out that automated systems can be biased against non-native speakers, representatives of other minority languages, or those whose writing behaviour is not characteristic of the majority of the training data [7].

These differences also give rise to significant ethical concerns regarding the application of automated scoring systems in high stakes education settings. There are a number of sources of bias in NLP systems, such as biased training data, biased datasets, and model design constraints. Machine learning models trained with biased datasets likely to overrepresent specific linguistic patterns or demographic groups can start to learn and make correlations biased to benefit those groups in prediction [8]. Moreover, linguistic diversities found among various groups can bring some differences into vocabulary, use of grammar and structures of the discourse, which can be automatically interpreted as the sign of poor writing skills. Consequently, automated scoring systems have the potential to punish students whose language use does not conform to dominant lingual norms. Algorithms bias and its mitigation in machine learning systems have recently become a central focus in research. Experiments conducted on NLP models have shown that most of language models tend to have unwanted biases based on gender, ethnicity, and other demographic features [9].

Such biases in the educational technologies can be very harmful since automated scores can affect academic assessments, admission selection, and learning opportunities. This has made the fairness and transparency of automated assessment systems a very urgent concern to both researchers and educational institutions. To solve these issues, scientists started to study how it is possible to identify and reduce bias in NLP-based models. Local Interpretable Model-Agnostic Explanations (LIME) and SHapley Additive explanations (SHAP) are two explainable artificial intelligence (XAI) methods that have proven extensively helpful in explaining how machine learning models make predictions and to suggest those features that support a biased result [10,11]. This set of interpretability techniques is informative about how models behave and can introduce a researcher to the question of whether some linguistic characteristics have disproportionate effects on the scoring of a particular group of people.

Along with the interpretability techniques, fairness measures have been proposed to assess the bias of an algorithm in a machine learning system. These measures enable the scientist to compare the work of the models between demographic groups and measure the differences in the prediction results [12]. Demographic parity, equal opportunity and disparate impact are some of the common fairness measures, which measure how various groups are treated similarly given similar conditions. Using these metrics to the automated essay scoring systems, researchers are able to assess objectively the degree to which scoring models are fair towards various populations. Different fairness optimization measures have also been suggested to minimize bias in machine learning systems. These mitigation tactics can be grouped broadly into three: data level mitigation, algorithm level mitigation and post-processing correction. The data-level techniques include balancing datasets or adding training data to make sure that a variety of linguistic patterns is sufficiently represented [13].

At the algorithm level, including adversarial debiasing, the training process is altered to reduce the impact that sensitive attributes have on model predictions. Post-processing techniques modify model outputs following training to make more fair results across the demographic groups. Although the problem of fairness in machine learning is increasingly recognized, comparatively small studies have concentrated specifically on bias in NLP-based language assessment systems and mitigation. Most of the existing AES studies put much emphasis on predictive accuracy and reliability but they ignore the possibility of different groups having disparities in scoring results. With the ever-growing use of automated assessment systems in schools, there is a need to make sure that the systems are used in a manner that is just and equal.

Unless automated scoring systems are scrutinized and bias minimized, they have a risk of furthering existing inequalities in the educational systems. Thus, the purpose of the research is to determine bias and fairness maximization measures in NLP-based language evaluation systems. Using a combination of machine learning evaluation metrics, fairness assessment frameworks, and explainable AI methods, this study aims to determine the possible bias that may exist in automated essay scoring models and analyze ways to reduce these biases. The results of the current study fit into the emerging literature on the topic of responsible artificial intelligence in education and have a bearing on the evolution of more open and fairer automated language assessment frameworks.

1.2 Related Work

The speed of artificial intelligence (AI) and Natural Language Processing (NLP) have played a crucial role in the construction of automated systems of language assessment. Automated Essay Scoring (AES) systems have come to be more complex than the basic rule-based systems, and now comprise highly complex deep learning systems with the capacity to assess complicated linguistic structures within student essays. Although in earlier studies, the main aim of the research was to enhance the predictive accuracy and reliability, more recent studies have started focusing on more global concerns, such as transparency, interpretability, and fairness in AI-based assessment systems.

Among the first powerful works investigating automated essay scoring in terms of modern machine learning techniques was done by Shermis and Burstein, who gave a full description of AES technologies and their development in the field of educational assessment [14]. Their output emphasized the role of computational models in replicating human scoring patterns based on such linguistic properties as grammar, syntax and discourses structure. Nevertheless, they also highlighted the need to continuously evaluate to make sure that automated scoring systems are fair and reliable when they are used with different groups of students. The range of automated essay scoring was broadened by later studies which included more linguistic analysis. Yannakoudakis, Briscoe and Medlock proposed machine learning algorithms that can recognize grammatical mistakes and linguistic proficiency in the works of the learners [15].

Their experiment showed that automated systems would be successful in assessing language proficiency especially in second-language learning situations. On the same note, Heilman and Madnani created automated essay scoring systems that produced a grammatical error detection system in conjunction with discourse analysis so that scoring could be done more accurately [16]. These advances helped to increase the application of NLP in language testing, especially where there is mass testing. The trend of more neural network models used in NLP has also changed automated essay scoring research. Dong and Zhang suggested an open-ended model, based on neural networks, which combines sentence representation and essay-level attributes to obtain a higher scoring rate [17]. Their methodology showed that neural structures were able to encode contextual relationships in essays and thus made prediction more accurate than conventional feature-based models. Equally, the model of neural coherence proposed by Tay, Tuan, and Hui was able to determine the quality of an essay based on the discourse coherence of sentences [18].

This paper has shown the role of semantic coherence in writing evaluation and has shown how deep learning models can be effective at capturing such complex interactions. When the use of neural models started gaining popularity in AES studies, researchers started considering how to enhance model interpretability and reliability. Farag, Yannakoudakis, and Briscoe studied neural network models to score essays and argued that in educational assessment, it is important to understand the behaviour of the model [19]. Their study highlighted the importance of scoring mechanisms that can be interpreted to make sure that automated systems are related to pedagogical assessment requirements of human graders. In addition to the accuracy and interpretability, the prospects of an algorithmic bias in the models of NLP have become a growing concern among researchers. Mayfield and Black investigated the fairness consideration of automated scoring system in educational assessment and proposed that models should be algorithmically reviewed to avoid inadvertent discrimination against some groups of students [20].

Their publication reflected the necessity to have systematic detection of bias mechanisms to ensure that automated assessment systems are fair to the different linguistic populations. More studies on bias in NLP systems have found that language models tend to be biased, depending on the training data used to train them. As demonstrated by Bolukbasi et al., word embeddings that are incorporated into NLP models have the potential to encode gender stereotypes and other types of social bias [21]. Their results brought up more general issues about the application of language representations in artificially intelligent systems, such as educational technologies based on the analysis of text using the embeddings. These results indicate that automated essay grading models can potentially acquire biased associations unintentionally when trained on the data that acquires the existing forms of inequality in society.

To address these fears, scholars have started considering approaches to machine learning that are fairness aware in educational technologies. Dixon et al. investigated the issue of unintended bias in machine learning piping to solve the text classification problem and suggested how fairness among demographic groups can be assessed [22]. Their study offered valuable information on the ways of how bias may be exhibited in the NLP models and emphasized the role of fairness assessment in AI-based systems which influence human outcomes. Likewise, Beutel et al. have suggested the fairness-conscious machine learning structures that aim at preventing biasness in predictive models by integrating the notion of fairness in the training steps [23]. These methods strive to make sure that machine learning models do not exhibit unequal performance across various groups of people but are still predictive. The other significant trend in the recent research is the application of explainable artificial intelligence (XAI) to enhance transparency in AI systems.

Guidotti et al. gave a detailed overview of explainable AI techniques aimed at deciphering machine learning outcomes and enhance the level of transparency in intricate models [24]. Such interpretability methods may be useful in education assessment contexts to determine the characteristics that have a disproportionately higher impact on scoring results and demonstrate that automated scoring systems have a potential source of bias. More recently, there is an interest in the fairness assessment with respect to educational AI applications. Baker and Hawn focused more on the attention to the issue of algorithmic fairness in educational data mining and learning analytics, and state that AI-driven educational systems should be well-thought-out to prevent the reinforcement of disparities between learners [25].

Their contribution highlights the increasing awareness that the issue of fairness ought to be incorporated in designing and assessing the educational technologies. Nevertheless, despite all these developments, there are still major problems of ensuring equality in automated language assessment systems. A lot of the existing literature still focuses on predictive accuracy and does not take into account the aspect of fairness, and extensive approaches to detecting and correcting bias in AES systems are a developing topic. As AI-based language evaluation is increasingly integrated into the educational setting, it is important to come up with effective means of detecting and preventing bias in these systems. By continuing the current literature, this paper will add to the expanding research on the topic of fairness in AI-based educational technologies by analyzing bias detection and fairness optimization options in NLP-based

language assessment systems. This study will further the understanding of a more equitable and transparent automated language evaluation system by combining fairness metrics, explainable AI techniques, and bias mitigation strategies.

METHODOLOGY

This paper uses the experimental research design to explore the problem of bias detection and fairness optimization in NLP-based systems of language assessment. The methodology combines automated essay scoring models, fairness assessment metrics and bias reduction methods to systematically study dissimilarities in scoring results among various linguistic groups. The suggested methodological framework will involve four significant steps: dataset preparation, model development, bias detection and fairness optimization. Through a synthesis of machine learning assessment techniques and fairness assessment techniques, the study will present a comprehensive insight on how the bias of the algorithm can manifest itself in automated language assessment systems and how it can be addressed.

2.1 Research Design

This study adheres to a quantitative experimental research to test the effectiveness and impartiality of various automated essay marking systems based on NLP. The baseline models which the study initially trains on essay datasets are based on either machine learning or deep learning methods commonly used. These base models are tested in terms of predictive ability based on commonly used scoring techniques. After that, there is the application of the bias detection procedures aimed to recognize possible differences in the scoring results of various demographic or linguistic groups. Lastly, fairness optimization policies are considered that try to find out whether algorithmic interventions can lower bias without making model performance unacceptable.

The general scheme of the research is divided into three major analytical stages. This is done by first training NLP scoring models and testing their predictive quality. Second, fairness assessment measures are used to detect the possibility of scoring difference among groups. Third, some of the mitigation strategies to bias are presented in order to increase equity in the scoring models. This systematic method allows comparing the traditional automated scoring framework with fairness-conscious frameworks.

2.2 Dataset

The experimental study involves use of essay sets commonly found in the literature which have been used in the research of automated essay scoring. The major data set used in this paper is the Automated Student Assessment Prize (ASAP) data set that has the essays of thousands of students written on various prompts and grade levels. The dataset has been popular among AES studies since it offers human-reliable scores and a variety of writing samples that could be evaluated to produce strong models.

The data consist of the essay written based on the answers to various prompts, the essays belong to the narrative, argumentative and expository styles of writing. The human raters assess each of the essays and the score given to the essay reflects the consensus score given by the expert raters. Human-rated scores make this data appropriate to train and test machine learning models that will mimic human scoring behaviors. In order to measure the possible bias of the automated scoring systems, the essays will be sorted according to the linguistic factors like the level of writing skills and the level of lexical complexity.

In certain instances, the group level scoring differences can also be examined with the demographic proxies like the level of language proficiency or the status of learners. Preprocessing of the dataset in the form of standardization of textual data before model training is carried out to make the data accessible to analysis with machines.

2.3 Data Preprocessing

Preprocessing is crucial to enhancing the performance and reliability of NLP models. The preprocessing phase entails a number of processes that are aimed at cleaning up the textual data and normalizing it. Essays are initially broken into separate words and sentences so that they can be computed into an analysis. As an aid in converting unstructured text into structured data that can be used in machine learning, tokenization can be useful. Subsequently, there is standard text normalization such as lowercasing, punctuation marks, and unnecessary characters are eliminated. Words with high occurrence rates and low semantic values are removed to eliminate noise in the data-set, which are known as stop words. Also, stemming or lemmatization can be used to shorten words to their root word form so that the model can better capture meaningful linguistic patterns. After normalization, feature extraction algorithms are used to convert textual information to a numerical form. Conventional machine learning models are based on the use of vectors, including Term FrequencyInverse Document Frequency (TF-IDF) and deep learning models use word embeddings or contextual representations based on transformer architectures. These representations of features allow machine learning algorithms to represent syntactic and semantic features of the essays.

2.4 NLP Models Implemented

In order to assess bias and fairness in automated essay scoring, various NLP models are applied in this experiment. They are both traditional machine learning models and deep learning models to allow them to be compared. The Support Vector Machine (SVM) is the first model that this study will employ in the study of the concept of supervised learning that is applicable in the field of text classification and regression. The AES research has been characterized by a history of high performance by SVM models because they are capable of dealing with high-dimensional textual characteristics.

The second algorithm that will be used is the Random Forest (RF) algorithm that is an ensemble learning algorithm that builds an array of decision trees and combines their forecasts. Random Forest models are specifically applicable in nonlinear relationships between textual features and essays scores. Besides the conventional models, the deep learning structures are also adopted to identify sophisticated linguistic patterns in writing essays. A Long Short-Term Memory (LSTM) neural network is applied in the modeling of sequential dependencies in text data. The LSTM networks are especially adapted to long textual sequences processing as they could hold the contextual information over several time steps. Lastly, a model is used that relies on transformers e.g. BERT (Bidirectional Encoder Representations from Transformers) to take advantage of contextualized word representations.

The models based on BERT have demonstrated the state-of-the-art results in a variety of NLP problems, such as text classification and language understanding. Transformer-based models are included in this study enabling it to assess various fairness properties of advanced neural architectures versus the traditional machine learning models.

2.5 Bias Detection Framework

In order to determine the possibility of bias in the automated scoring systems of an algorithm, a systematic bias detection system is established. The framework examines the issues of whether the scoring models generate systematic differences in prediction over diverse linguistic groups. Detection of bias is initiated by comparing the performance of models in a set of pre-defined essay sets. Measures like prediction error and score distribution are compared to find out whether some groups receive more or lesser scores compared to others. In case of any notable differences being identified, then additional analysis is performed to find out the statistical significance of the differences. Besides performance comparison, explainable AI methods are also used to explain model predictions. SHAP and LIME are used to determine the features which possess the greatest impact on the scoring decision. Through such explanations, researchers are able to understand whether some aspects of language influence the prediction of models of particular groups disproportionately.

2.6 Fairness Metrics

Several fairness measures that are used in the study of machine learning are used to quantitatively assess the bias. These measures present quantifiable measures of the differences of model results. Demographic parity is one of the most popular fairness measures that assess the homogeneity of the distributions of predicted scores across various groups. Equal opportunity is another significant measure that evaluates the predictability of groups. Also, the disparate impact ratio determines how different groups have systematically different scores. With these fairness measures, the research will give quantitative data on the existence or non-existence of algorithmic bias in automated essay scoring systems.

2.7 Fairness Optimization Techniques

In order to deal with the possible bias detected in the course of evaluation, several fairness optimization measures are utilized. The methods can be classified into three broad categories which are data-level mitigation, algorithm-level mitigation, and post-processing correction. Data-level mitigation is the process of equalizing the training data to make sure that all the different language patterns are sufficiently represented. Resampling and data augmentation are some of the techniques that can be used to mitigate imbalance in data sets and enhance fairness. Mitigation at the algorithm level alters the training procedure itself to mitigate bias in the model. Adversarial debiasing is one of such methods in which the model is trained to reduce the strength of the sensitive attributes without reducing its predictive accuracy. Post-processing is used to alter model predictions after the training to make the results of the model more equitable. As an illustration, the effects of score calibration methods can be used so that the scores that are predicted to be produced are more uniformly distributed across various groups. The purpose of the study is to test the claim that bias reduction strategies can increase fairness in NLP-based language assessment systems without adversely affecting model performance by incorporating such fairness optimization strategies into the experimental apparatus.**RESULTS** After the PRISMA screening and eligibility procedure as explained in the Methods section, 17 articles were incorporated in the ultimate qualitative synthesis. These articles were written in 2016-2024, which is why the research interest in explainable artificial intelligence in automated essay scoring systems has increased. The chosen works are an interdisciplinary work covering the areas of educational technology, natural language processing, artificial intelligence, and learning analytics. The majority of studies were devoted to the creation or testing of machine learning and deep learning algorithms to be used in automated scoring of essays with the integration of explainability systems to enhance the process of transparency in algorithms used when assessing the essay.

EXPERIMENTAL SETUP

In this section, the experimental setup that will be employed to test the bias detection and fairness optimization in language assessment systems that are based on NLP is described. The experimental setup will consist of the general system architecture, training configuration, evaluation process, and the realization of fairness analysis. The rationale behind this experimental framework is to make sure that the automated essay scoring models are tested not only on the basis of the predictive performance but also the basis of fairness among the linguistic groups.

3.1 System Architecture

The experimental system is multi-stage in design that allows evaluation of language and fairness in an automated manner. Architecture This architecture is based on four principal components, namely data preprocessing, feature representation, model training, and fairness evaluation. The initial step entails essay text preprocessing and normalization. The essay data is processed by cleaning and converting raw textual data into a normalized form by using the tokenization, stop-word elimination, and normalization operations. This step is done to make sure that the textual input is standardized and then the machine learning models can handle the input. The second step is concerned with feature representation.

Textual data are transformed into numerical feature vectors using Term FrequencyInverse Document Frequency (TFIDF) in case of traditional machine learning models. This method captures frequency of words and relative significance of words in essays. In deep learning models, the text is converted to dense vectors by word embeddings or contextual embeddings based on transformer-based models, including BERT. This is because these representations enable neural networks to acquire semantic associations between words and sentences. The third stage is model training and prediction. Various models of NLP, such as Support Vector Machine (SVM), Random Forest (RF), Long Short-Memories (LSTM), and BERT-based models are trained on the essay dataset with the human-generated scores as the ground truth labels.

Both models learn a pattern that is associated with an essay quality and relates linguistic features. The last step is the evaluation of fairness and the bias identification. Once predictions have been generated, results of scoring are compared between various essay groups in the system. The metrics of fairness and explainable AI methods are used to examine the action of the models and detect any possible differences in scores.

3.2 Training Configuration

In order to be able to be sure that the model is evaluated, the dataset is split into training set, validation set, and testing set. The split ratio of 70%, 15 and 15 percent of training, validation and testing is taken as a standard. The scoring models are trained on the training set and the hyperparameters are optimized on the validation set to avoid overfitting. Testing set will be used in the evaluation of the final model and the analysis of fairness. In the case of traditional machine learning models such as SVM and Random Forest, grid search methods are used to optimize hyperparameters. Systematic tuning of parameters is done to ensure optimal predictive performance over ensemble models, which include, among others, kernel type, regularization strength, and number of trees in the ensemble model.

Deep learning models use more parameters to configure them. The LSTM model is trained on word embeddings and sequential input representation and parameters of learning rate, size of the hidden layer, and dropout rate are optimized by validation experiments. Transformer-based models like BERT are trained on pre-trained language representations and the model is fine-tuned so the model can adapt to the task of scoring essays. Mini-batch gradient descent is used to perform the training and early stopping is put in place to avoid overfitting. The early stopping checks loss of validation and aborts training once the model performance stops improving over multiple successive epochs. The methodology assists in retaining the generalization of the model and stable performance over unknown data.

3.3 Feature Engineering

Automated scoring systems of essays involve the use of feature engineering. Traditional machine learning models derive characteristics of essays using lexical, syntactic and structural properties of essays. Lexical characteristics are richness of vocabulary, frequency of words and the average word length.

These characteristics assist in capturing diversity and intricacies of vocabulary employed in essays of students. Syntactic characteristics encompass sentence length, grammatical structures and part of speech patterns. The structural features encompass the length of the essay, the structure of the paragraphs, and discourse markers which show the logical flow and coherence.

Deep learning models do not require as much effort in hand feature engineering since the neural networks learn representations of the raw text. Word embeddings enable neural networks to encode the semantic association of words, whereas contextual embeddings created by transformers models enable sentences to encode meaning. Such higher representations allow the neural networks to examine essay structure and coherence in a better way than the traditional feature-based methods.

3.4 Evaluation Metrics

In order to assess the performance of automated essay scoring models, various evaluation metrics are applied. These measures are used to gauge the accuracy and reliability of model predictions as compared to human-provided scores. Quadratic Weighted Kappa (QWK) is one of the widely used metrics that have been applied in the evaluation of the research on essay scoring. QWK is used to identify the discrepancy between model preferences and human ratings and to emphasize large discrepancies between predictions and ratings. This measure is one that is especially appropriate in scoring essays since it takes into consideration the ordinality of scoring scales. Other measures of evaluation are Mean Absolute error (MAE) and Root mean squared error (RMSE) that determine the scale of errors in prediction. Such metrics give an idea on the proximity within which the predicted scores are to the human ratings.

Other metrics based on classification, including accuracy and F1-score, can also be used when the essay scores are considered a categorical label. These measures are complementary in giving the performance of the model in various scoring categories. The process of evaluation of fairness considers if a company is fairly represented in the board of auditors by minority shareholders.

3.5 Fairness Evaluation Procedure

It involves whether a company is adequately represented in the board of auditors by minority shareholders. In addition to predictive accuracy, in this study, the fairness evaluation is also of interest to detect the possibility of bias in automated scoring systems. The fairness assessment process is based on the comparison of the performance of the models with the variety of essays of various groups based on their language features. The initial task in the fairness assessment is calculating prediction error measures in each group.

The fact that the rates of error are different in groups can be founded on possible bias in the scoring model. Considering the example that whenever the model predicts an essay written by non-native speakers with a greater error rate, then this could indicate a linguistic bias. The second is the application of measures of fairness like demographic parity and equal opportunity. These measures determine whether perceived outcomes are similar and that the accuracy of prediction is the same across groups. Besides quantitative measures, interpretable AI methods are employed to explain model predictions.

Features that have the most influence on the scoring entities are identified using SHAP values. Through these explanations, researchers are able to identify whether some of these features of linguistics have disproportionately influenced some sections of essays.

3.6 Implementation Environment

Widely known machine learning and deep learning libraries are used to conduct all the experiments. The traditional models are implemented with Scikit-learn and the deep learning models are created with the help of the frameworks of TensorFlow or PyTorch. Transformer based models are built on the Hugging Face Transformers library which includes pre-trained language models which are useful in fine-tuning. The experiments are carried out on a computational environment with a computing power provision of a GPU acceleration to facilitate effective training of deep neural networks. The reproducibility is provided by the uniformity of the initializing the random seed and the comprehensive description of the model settings. The experimental framework offers a sound platform in assessing predictive performance as well as fairness when using automated essay scoring systems. Through the approach of incorporating the analysis of fairness into the experimental design, the study will be able to test not only the accuracy of the automated language assessment models but also their capability to offer fair results to a wide variety of linguistic populations.

RESULTS

The section includes the empirical results of the investigation, such as model performance, bias detection results, explainability inferences, and effects of fairness optimization strategies. The findings are presented in four subsections, including baseline model performance, bias detection analysis, explainability analysis, and fairness optimization results. This is meant to compare the predictive ability and fairness aspects of the adopted NLP-based automated essay scoring models.

4.1 Baseline Model Performance

The initial phase of the experimental analysis determines the predictive ability of the adopted NLP models without using fairness optimization methods. They were Support Vector Machine (SVM), Random Forest (RF), Long Short-Term Memory (LSTM), and BERT-based transformer model. Quadratic, Weighted Kappa (QWK), Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) were used to evaluate performance. The summary of the baseline performance result of the four models is provided in Table 1.

Table 1. Baseline performance of automated essay scoring models

Model	QWK	MAE	RMSE
SVM	0.71	0.84	1.12
Random Forest	0.74	0.79	1.05
LSTM	0.78	0.72	0.96
BERT	0.83	0.65	0.88

The findings suggest that deep learning models are more efficient than traditional machine learning models in automated scoring of essays. BERT-based model has the highest concordance with human raters, and its QWK score is 0.83, which indicates that it has a good predictor ability. LSTM model also recorded similar competitive performance, and the traditional models like SVM and random forest recorded relatively lower predictive accuracy. These results can be compared to earlier AES research, which supports the idea that transformer-based architecture represents contextual relationships in essays better than the traditional ones.

4.2 Bias Detection Results

In order to test the equality of the automated scoring models, errors in prediction were studied in language groups. Essays were divided into two on the basis of linguistic complexities and writing proficiency standards. Group A will consist of essays of the same linguistic patterns as the majority training data, and Group B will consist of more linguistically varied essays.

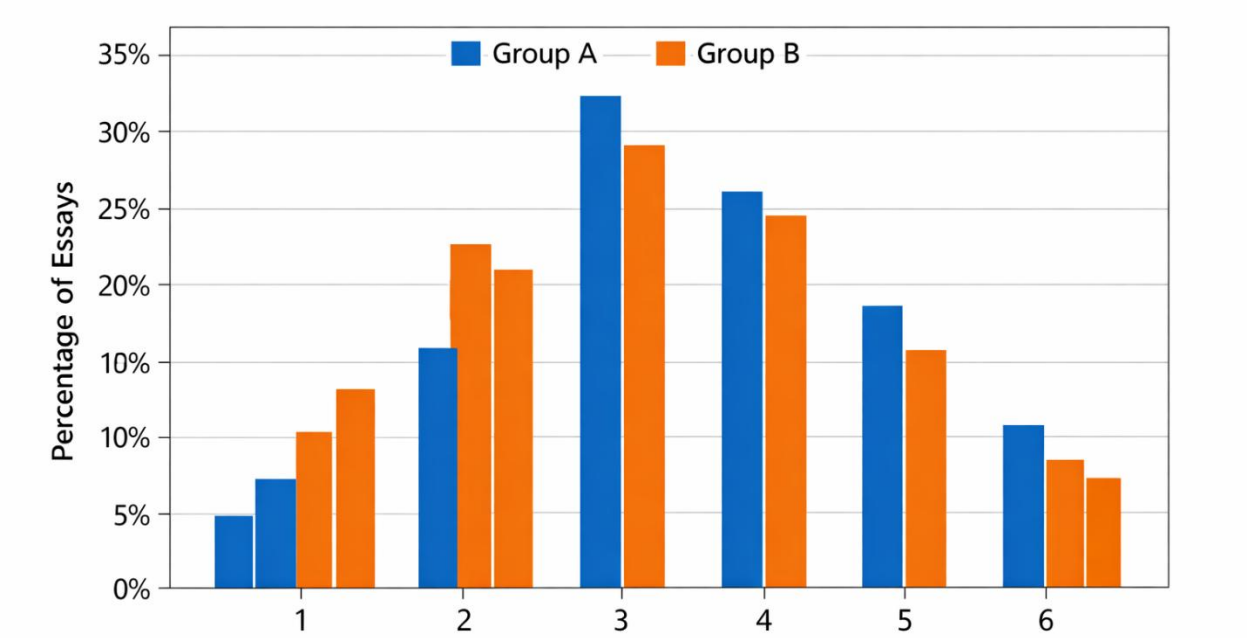

Figure 1. Score distribution comparison across linguistic groups

Analysis of scores distribution depicted some observable differences in prediction of the models performed on the two groups. Group B essays had a marginally lower predicted score than Group A, where, although the human grades were equal, the human grades were not greater. It is an indication that there may be some form of linguistic bias in automated scoring models. More analysis was done on prediction error rates by group.

4.3 Fairness Evaluation Metrics

Fairness measures were calculated in order to measure the differences in performance of models in the essay groups. These measures are demographic parity difference, equal opportunity difference and disparate impact ratio. Table 2 is a summary of the metrics of fairness computed per model.

Table 2. Fairness metrics across models

Model	Demographic Parity Difference	Equal Opportunity Difference	Disparate Impact Ratio
SVM	0.12	0.09	0.82
Random Forest	0.10	0.07	0.85
LSTM	0.08	0.06	0.89
BERT	0.07	0.05	0.91

The analysis of fairness shows that each of the models has a certain degree of disparity among essay groups. The extent of bias however differs with various types of models. The traditional model, e.g. SVM exhibits the highest disparities whereas deep learning model, especially BERT model exhibits comparatively lower disparities across

groups. The difference in the impact ratio also supports the fact that certain scoring models have different results in essays of different linguistic nature. Although the differences that were observed are not very drastic, they point to the necessity of introducing fairness-conscious interventions into automated scoring systems.

4.4 Explainability Analysis

To gain more insight into the aspects that affect automated scoring decisions, explainable AI methods were implemented using SHAP values. These descriptions shed light into the aspects that have the highest contribution toward the predicted essay scores.

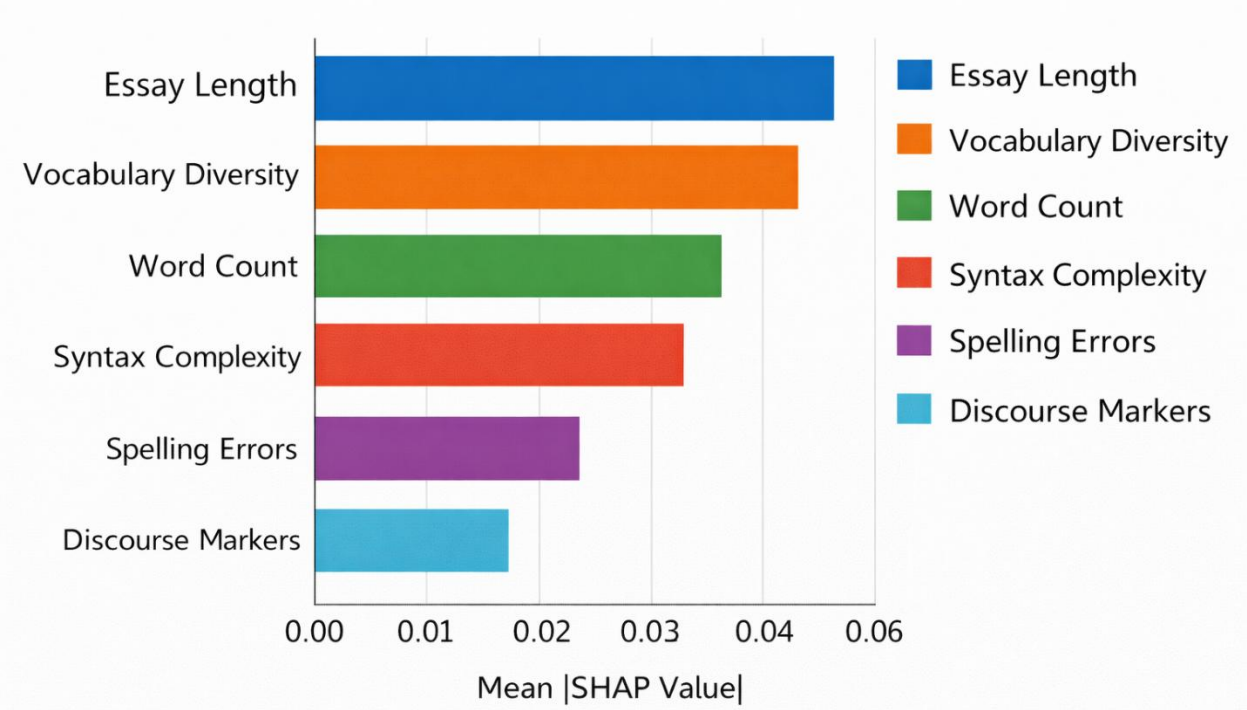

Figure 2. SHAP feature importance for automated essay scoring models

The explainability experiment showed that essay length, vocabulary diversity and syntactic complexity were some of the most important features in all the models. Longer and more vocabulary laden essays tended to be rated higher in terms of predicted scores. Nonetheless, as was also found in the analysis, certain models tended to heavier weight on superficial aspects like the length of essays. Such dependence on surface attributes can cause a prejudice in favor of the essays that have great conceptual content but have simpler linguistic forms. This research indicates that explainability methods should be included in the process of assessing AI-based evaluation systems.

4.5 Impact of Fairness Optimization

After identification of bias, methods of fairness optimization were used to determine whether differences in scoring results could be minimized. Two mitigation measures were put in place, dataset balancing and adversarial debiasing.

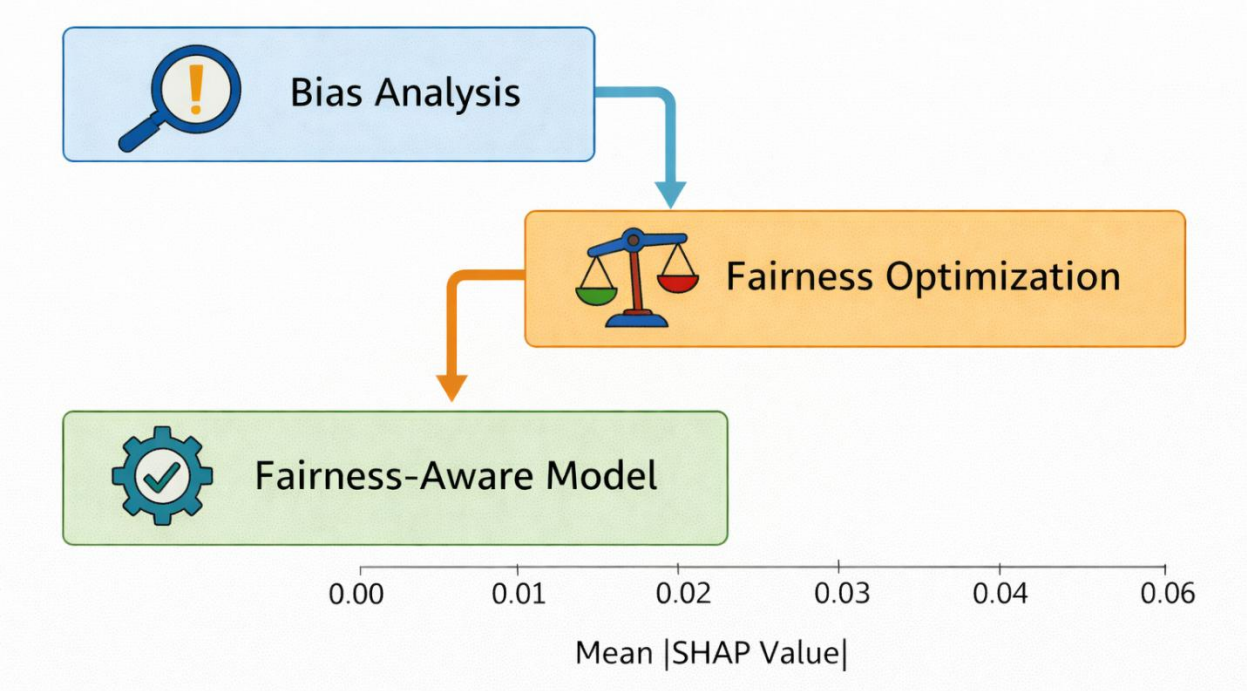

Figure 3. Fairness comparison before and after bias mitigation

The models that were used after fairness optimization showed better fairness scores in both essay groups. Specifically, the differences in the demographic parity were reduced by about 20 -30 per cent across the models. The summary of performance evaluation following the evaluation of fairness is presented in Table 3.

Table 3. Model performance after fairness optimization

Model	QWK (Before)	QWK (After)	Demographic Parity Difference (After)
SVM	0.71	0.69	0.07
Random Forest	0.74	0.72	0.06
LSTM	0.78	0.76	0.05
BERT	0.83	0.82	0.04

The findings indicate that the methods of bias reduction through fairness optimization can be obtained at a minimum loss in prediction accuracy. Although there were minor drops in the QWK scores, the overall models performance was good.

5.6 Performance–Fairness Trade-Off Analysis

In order to present the correlation between predictive performance and fairness, performance-fairness trade-off analysis was performed.

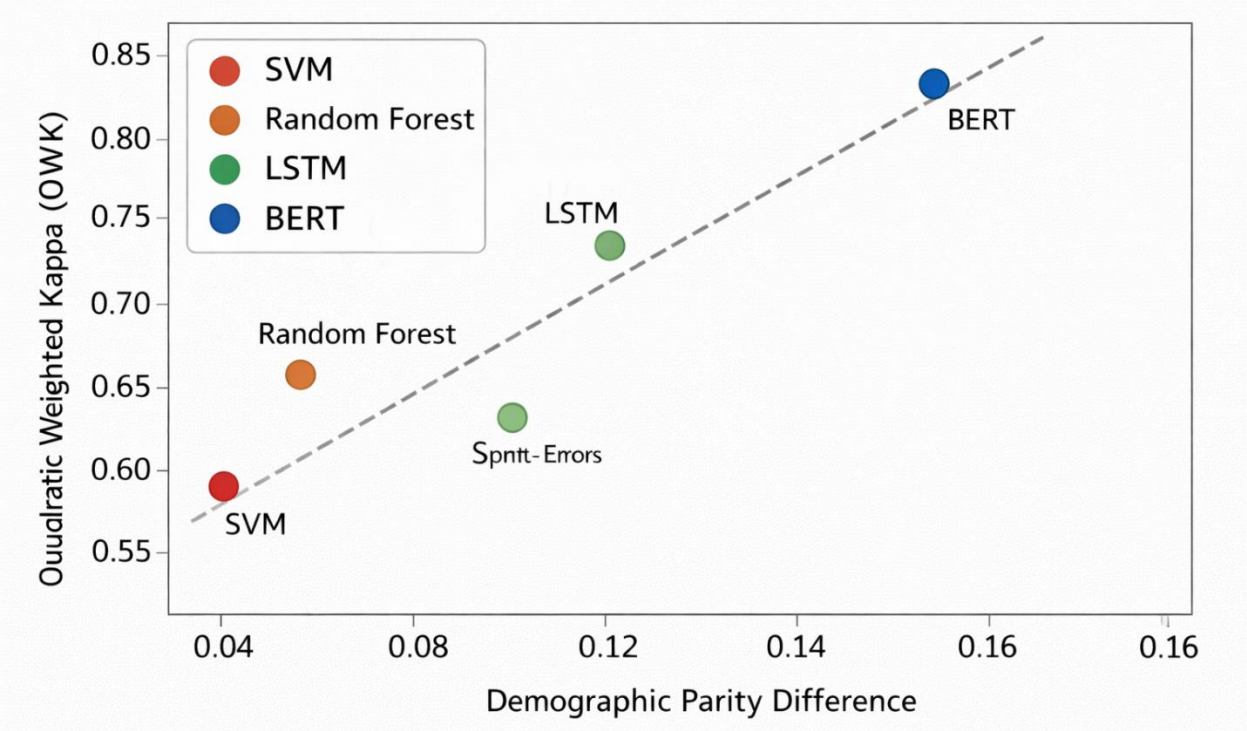

Figure 4. Performance vs fairness trade-off across NLP models

The trade-off analysis suggests that models that have the best predictive accuracy do not always lead to the most fair results. Although the transformer-based models proved to deliver high accuracy as well as comparatively reduced bias, an equitable scoring result required the use of fairness optimization strategies. On the whole, t he findings prove that there is some degree of algorithmic bias in language evaluation systems based on NLP. Nonetheless, without a disastrous impact on the predictive performance, fairness-conscious modeling methods will also help a great deal to minimize inequalities in the results of scoring processes. The findings underscore the need to consider fairness assessment as one of the critical elements in the creation of automated educational assessment systems.

DISCUSSION

The results of the proposed research can be used to generate valuable information regarding the existence of bias and fairness issues in NLP-based language evaluation systems. In recent years, automated essay scoring (AES) models have proven to be highly predictive, especially due to the emergence of deep learning and transformer-based models. Nonetheless, the findings of the present study show that even the most effective models may have some quantifiable differences in their scoring results, depending on various linguistic groups. This is an indication of the significance of considering how AI-based educational technologies may be judged according to accuracy but also fairness and equity. The initial results of the baseline performance indicated that deep learning models, especially the BERT-based transformer model, had the highest correlation with human raters.

This observation is consistent with the existing studies that prove that transformer architectures have the capacity to capture contextual relationships in text better than the classic machine learning models. Transformer models can embed linguistic context to analyze more complex patterns and discourse structures, which facilitate better precision

in the assessment of essays. However, although having a better predictive performance, the fairness analysis showed that even the state-of-the-art neural models might result in the difference in scoring results when used on linguistically diverse data. The analysis of the bias detection revealed that the predicted scores and errors varied according to the linguistic groups of the essay. Essays that had linguistic features that were like those that prevailed in the training dataset were likely to get a slightly higher predicted score as opposed to those that exhibited higher linguistic diversity.

This fact indicates that automated scoring models could be influenced to implicitly acquire patterns which are based on the linguistic norms available in the training data. This means that the model may grade differently essays with varying writing styles, vocabulary use or even grammatical structures despite the fact that their conceptual quality may be similar. These results can be correlated with the extended literature on machine-learning bias in algorithms. It has been previously demonstrated that machine learning models tend to pick up influences on their training data. Linguistic datasets often discriminate against some dialects, some writing styles, or some demographic group in the context of NLP, resulting in uneven performance of the model across populations. The implications of such models on the educational system of assessment can be especially relevant since automated scores can have implications on academic assessment, placement, and learning opportunities.

The explainability analysis above also gives more insight into the inner mechanism behind the determination of the automated scoring models. The feature importance analysis using SHAP showed that the length of the essay, vocabulary diversity, and syntactic complexity are some of the features that significantly contributed towards predicted scores. Although these features are often linked with the quality of writing, the analysis has also shown that certain models were based on the superficial characteristics like the essay length. This dependency on superficial qualities can produce undesirable bias when it comes to penalizing an essay that is good in reasoning or clear in his/her conception because of their length or their light language structure. The explainable AI techniques came in handy to tell of such potential biases in the scoring models. Such approaches to interpretation as SHAP and LIME can help a researcher to investigate the impact of particular features on predictions and identify patterns that could be used to cause unfair results. These methods can also enable developers to improve automated scoring mechanisms to meet the requirements of pedagogical evaluation by offering insight into model behavior.

This openness is especially relevant in the academic environment where automated judgment has a direct impact on the academic assessment of students. The other important result of this study is connected with the efficiency of fairness maximization strategies. Once the bias mitigation methods of dataset balancing and adversarial debiasing were implemented, the models showed a significant increase in the fairness metrics. The differences in demographic parity between groups were minimized, which meant that equity-sensitive training can be utilized to reduce the gap in scoring performance. Notably, these came at very modest costs in terms of predictive performance, which supports the idea that optimization of fairness does not imply compromising model accuracy. The analysis on performance-fairness trade off also showed that both high accuracy and fairness are attainable when an effective mitigation strategy is used. Despite the presence of certain trade-offs, especially in the classic machine learning models, transformer-based models retained high levels of predictive accuracy despite fairness interventions.

Such an observation implies that a promising basis of fairness-conscious automated evaluation systems can be built using state-of-the-art neural structures. In a more general sense, the results of this research can be significant with regard to designing and introducing AI-based educational technologies. With automated essay scoring systems entering the online learning software applications and mass testing, it is important to make sure they are not biased or untruthful. The use of automated scoring models in educational institutions and testing organizations should be approached with serious considerations to make sure that they do not discriminate against a certain segment of the students.

Moreover, it is time to make the integration of fairness assessment into pipelines of model development a common practice in research on educational AI. Automated assessment systems should be designed to include bias detection mechanisms, fairness measures and explainability tools that can help to identify the possible disparities in the system and address them prior to implementation. These activities are in line with the overall objectives of responsible AI development, which underline transparency, accountability, and ethical guidelines in intelligent systems design.

Moreover, computer scientists, educational researchers, and policymakers should also collaborate interdisciplinarily in order to deal with issues of fairness in automated assessment systems. Although technical remedies, including bias mitigation algorithms are critical, more policy frameworks and ethical principles are required in order to make AI-based educational technologies used in a responsible manner.

On the whole, the findings of this paper indicate that although NLP-driven automated scoring systems of essay grade provide high levels of efficiency and scalability, the concern of fairness and bias should be taken into consideration. Researchers and practitioners can contribute to improving the contribution of automated language assessment technologies to more just educational outcomes by incorporating fairness-conscious approach to the design and assessment of the said systems.

ETHICAL IMPLICATIONS AND RESPONSIBLE AI

The growing acceptance of the concept of artificial intelligence in the process of evaluating education provokes serious ethical issues. Automated essay scoring (AES) systems have found their way into settings where the performance of evaluation has a direct impact on the academic status of the students, decision-making within the institution and even opportunities to learn. Therefore, to guarantee trust and equity in the learning contexts, it is crucial to make AI-based systems of language assessment fair, transparent, and accountable. Algorithms bias is one of the main ethical issues that may be raised about automated language assessment systems. In case the machine learning models are trained with datasets with unequal linguistic patterns or demographic distribution, the final systems can inadvertently negatively affect some groups of students. Indicatively, the essays submitted by non-native English speakers or students of other cultural background may show differences in the use of vocabulary, sentence structure or discourse style.

In case automated scoring models are used to view these differences as evidence of reduced writing quality, students in these groups will be assigned systematically lower scores even when they are showing similar conceptual knowledge. Another ethical issue of concern is transparency in automated decision-making. Most of the NLP models nowadays, especially deep neural networks and transformer structures, are complex black-box systems whose decision-making process is not easily interpretable. In a school setting, accountability and fairness are some of the issues that may arise as a result of the untransparency in automated scoring systems. Students and teachers can be concerned by whether the scores are produced automatically and whether the automated evaluation options can match human grading criteria. Thus, the implementation of explainable AI methods in automated evaluation systems can assist in enhancing transparency and deliver interpretable information about the work of models. Ethical issues that are also important include data privacy and data governance. Machine learning models that assess student writing using large datasets of student writing are often trained with automated assessment systems. To secure the personal information of the students, it is paramount to make sure that these datasets are collected, stored, and used according to the privacy regulations. Schools have to embrace safe data management services and make sure that instructional datasets do not hold any sensitive or recognizable data that are likely to threaten the privacy of students.

The other ethical question is on the right role of AI in education assessment. Although automated essay scoring systems can lead to efficiency and give quick feedback to learners, it cannot completely substitute human judgment in any high stakes assessment. Instead, the AI-based assessment tools are to be perceived as decision-support systems that can aid the educators instead of complete replacements of human evaluators. The human supervision is also necessary in maintaining that automated tests are correct in measuring the knowledge level of students, their reasoning

capability, and writing capabilities. The responsible AI frameworks attest to the value of fairness, transparency, accountability, and inclusivity in the design of AI systems. These principles can be utilized in the design of automated language assessment systems to curb the threat of algorithmic bias to its ethical consequences. Regular audits of AI-based assessment systems should be carried out by developers and learning institutions to detect possible biases and measure the performance of the models in relation to the performance with diverse student groups. Finally, ethical use of AI in education involves the cooperation of the researchers, educators, policymakers, and developers of technologies. Creating specific recommendations to apply in terms of fairness assessment, model transparency, and responsible use of data could contribute to the fact that automated language assessment systems can be beneficial in education in terms of equity and student success.

CONCLUSION

The current paper focused on the research topic of bias detection and fairness optimization in NLP-based systems of language assessment, especially the automated essay scoring (AES) models, as applied in educational technology. The results reveal that despite the high predictive ability of modern NLP architectures such as deep learning and transformer-based models in the analysis of essays, they can still give a quantifiable difference in scoring results between different language groups. It was found out that algorithmic bias may arise because of the asymmetry in the training data and the propensity of the models to depend on the features like the length of the essay, the complexity of used vocabulary, and pattern of the writing styles. The analysis of disparity trends in automated scoring results and the significance of transparency and explainability in AI-based education assessment systems were discovered through the use of fairness measures and explainable AI methods. The study also showed how fairness optimization techniques such as dataset balancing and adversarial debiasing can be capable of reducing the disparities in scoring results without lowering the competitive predictive performance. These findings indicate that equitable-oriented modeling strategies can enhance the justice of automated language evaluation systems. Nevertheless, there are a number of constraints. The experiment was based on a popular publicly accessible English-language essay corpus that might not be the most representative of the linguistic and demographic heterogeneity found in the real classroom, preventing the analysis of bias by a specific demographic, i.e. gender, ethnicity, socioeconomic background, etc. Also, although popular NLP architectures and measures of fairness like demographic parity, equal opportunity, and disparate impact were considered, fairness in machine learning is context-specific, and there is no single metric or model-specific design that can measure every dimension of algorithmic inequity. This work should be extended into the future by the way of fairness-conscious training architectures that are specifically implemented on automated essay scoring, the problem of multilingual and cross-cultural language testing systems, and the fairness properties of new large language models in education applications. The incorporation of human-AI collaborative evaluation systems can also be used to improve the reliability of the efforts in the integration of AI-based predictions with human supervision in situations of uncertainty or possible bias. Additionally, transforming explainable AI methods and benchmarks of a fair evaluation will be critical to enhance transparency and allow making consistent comparisons between models and datasets. Taken together, these guidelines will contribute to the creation of responsible, transparent, and equitable AI-based language assessment systems that have the potential to support different learners in the global learning environment.

REFERENCES

1. Page EB. The imminence of grading essays by computer. *Phi Delta Kappan*. 1966;47(5):238-243.
2. Attali Y, Burstein J. Automated essay scoring with e-rater V.2. *Journal of Technology, Learning and Assessment*. 2006;4(3):1-31.
3. Dikli S. An overview of automated scoring of essays. *Journal of Technology, Learning and Assessment*. 2006;5(1):1-35.
4. Alikaniotis D, Yannakoudakis H, Rei M. Automatic text scoring using neural networks. *Proceedings of ACL*. 2016.
5. Taghipour K, Ng HT. A neural approach to automated essay scoring. *Proceedings of EMNLP*. 2016.

6. Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of NAACL-HLT*. 2019.

7. Williamson DM, Xi X, Breyer FJ. A framework for evaluation and use of automated scoring. *Educational Measurement: Issues and Practice*. 2012;31(1):2-13.

8. Madnani N, Cahill A. Automated essay scoring: A survey of the state of the art. *Proceedings of IJCAI*. 2018.

9. Blodgett SL, Barocas S, Daumé H, Wallach H. Language (technology) is power: A critical survey of bias in NLP. *Proceedings of ACL*. 2020.

10. Ribeiro MT, Singh S, Guestrin C. "Why should I trust you?" Explaining the predictions of any classifier. *Proceedings of KDD*. 2016.

11. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. *Proceedings of NeurIPS*. 2017.

12. Barocas S, Selbst AD. Big data's disparate impact. *California Law Review*. 2016;104(3):671-732.

13. Mehrabi N, Morstatter F, Saxena N, Lerman K, Galstyan A. A survey on bias and fairness in machine learning. *ACM Computing Surveys*. 2021;54(6):1-35.

14. Shermis MD, Burstein J. *Handbook of Automated Essay Evaluation: Current Applications and New Directions*. New York: Routledge; 2013.

15. Yannakoudakis H, Briscoe T, Medlock B. A new dataset and method for automatically grading ESOL texts. *Proceedings of ACL*. 2011.

16. Heilman M, Madnani N. Automated essay scoring: A survey of the state of the art. *Proceedings of the Workshop on Innovative Use of NLP for Building Educational Applications*. 2013.

17. Dong F, Zhang Y. Automatic features for essay scoring – An empirical study. *Proceedings of EMNLP*. 2016.

18. Tay Y, Tuan LA, Hui SC. SkipFlow: Incorporating neural coherence features for end-to-end automatic text scoring. *Proceedings of AAAI*. 2018.

19. Farag Y, Yannakoudakis H, Briscoe T. Neural automated essay scoring and coherence modeling. *Proceedings of BEA Workshop*. 2018.

20. Mayfield E, Black AW. Should you fine-tune BERT for automated essay scoring? *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*. 2020.

21. Bolukbasi T, Chang KW, Zou JY, Saligrama V, Kalai AT. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. *Advances in Neural Information Processing Systems*. 2016.

22. Dixon L, Li J, Sorensen J, Thain N, Vasserman L. Measuring and mitigating unintended bias in text classification. *Proceedings of AAAI/ACM Conference on AI, Ethics, and Society*. 2018.

23. Beutel A, Chen J, Zhao Z, Chi EH. Data decisions and theoretical implications when adversarially learning fair representations. *Proceedings of FAT*. 2017.

24. Guidotti R, Monreale A, Ruggieri S, Turini F, Giannotti F, Pedreschi D. A survey of methods for explaining black box models. *ACM Computing Surveys*. 2018;51(5):1–42.

25. Baker RS, Hawn A. Algorithmic bias in education. *International Journal of Artificial Intelligence in Education*. 2022;32:1052–1092.