University of Ibadan Journal of Science and Logics in ICT Research (UIJSLICTR)

ISSN: 2714-3627

A Journal of the Department of Computer Science, University of Ibadan, Ibadan, Nigeria

Volume 14 No. 1, June, 2025

journals.ui.edu.ng/uijslictr
http://uijslictr.org.ng/
uijslictr@gmail.com

University of Ibadan
Journal of Science and Logics in ICT Research

Development of a Nigerian English classification model: for AI-Driven Grading Systems

[1] ✉ **Asoro, O. B and [2]Osunade, O.**

[1,2] Dept of Computer Science, University of Ibadan, Ibadan, Nigeria
[1]raymondblessing5@gmail.com and [2]o.osunade@ui.edu.ng

Abstract
The current Automated Essay Assessment Systems (AEAS) are predominantly trained on native Standard English, thereby introducing bias when grading essays written in other variations of English- Nigerian English. This bias leads to unfair grading, misclassification of valid linguistic features, and an increased failure rate among students. Nigerian English, the official language of Nigeria, incorporates linguistic features that differ from native English expressions. This study aims to enhance grading fairness by developing a Nigerian English classification model using K-Nearest Neighbors (KNN) and Term Frequency-Inverse Document Frequency (TF-IDF). The model successfully identifies and classifies Nigerian lexical features, by incorporating Nigerian English dictionaries and crowdsourced speech resources, aiding in unbiased assessments. Results suggest that this approach significantly improves recognition of Nigerian English expressions, contributing to fairer academic evaluations.

Keywords: *Natural Language Processing, ESL Writers, Educational technology, Context and computational approach*

1. Introduction

Language variation presents a unique challenge in computational linguistic assessments, particularly for non-native English varieties such as Nigerian English. The categorization of Nigerian English within the broader spectrum of World Englishes places it outside the Kachru (1983, 2017) Inner/Native Circle category, (countries with English as their native language), but in the Outer circle (countries with English as their second language).

The growing use of Automated Essay Assessment Systems (AEAS) in academic evaluations has brought efficiency but also fairness concerns, especially regarding linguistic diversity. Current AEAS are trained on native English, disregarding linguistic features of Kachru's Outer and Expanding circle variation of English users. Nigerian English, which serves as Nigeria's official language, are characterized with nativised lexicons like "longthroat" which translates to greed and "brideprice" which translates to dowry in native English. These lexicons are integral to Nigerian English yet are often flagged as incorrect by traditional assessment systems (Dada *et al.*, 2018) resulting in linguistic grading bias within existing AEAS. The objective of this research is to develop a model that accurately classifies Nigerian English features, enabling AI-driven essay grading systems to adapt to linguistic diversity.

2. Related Works

Yang *et al*,. (2022) in their study noted that the core AEAS design method can be grouped into three classes: the design centered on humans and computers, a method that pays attention to the association between essay grads and external writing measures and lastly, the design approach with a sole focus on grading process. This section of related works shares more light on the design evolution of AEAS.

The first AEAS system tagged Project Essay Grader (PEG) was developed by Ellis Page in 1966. This research was put in motion in response to rising limitations associated with traditional form of pen on paper grader done by human graders. The dataset for this work consists of four hundred pre-graded essays, written by students using the Kachru's inner circle of English language. Assessment of these essays was based solely on writing style of the students. Thus, linguistic features for which

Asoro, O. B and 2Osunade, O. (2025). Development of a Nigerian English classification model: for AI-Driven Grading Systems. *University of Ibadan Journal of Science and Logics in ICT Research (UIJSLICTR),* Vol. 14 No. 1, pp. 199 – 205

regression techniques were applied for grading includes average word length, parts of speech, essay length, number of commas, preposition, adjective and rarely used tenses. Unfortunately, PEG was found wanting by the educational stakeholders, because users were able to fool the systems into awarding higher assessment scores by merely giving relevant or irrelevant length of essays.

In improving this, Page and Peterson (1995), collaborated and made use of natural language processing method to develop AEAS that can handle essay assessment focusing on spellings and grammatical construct. The dataset used for training this model was again obtained from native English writers, with about five hundred linguistic feature extractions carried out on them.

E-rater1 is an AI automated essay grading system developed by Educational Testing Service (ETS) (1999). It made use of regression techniques for feature extraction ranging from 10-12, for automatic grading, for grading. E-rater1 is the core engine for some standardized testing platforms such as Test of English as a Foreign Language (TOEFL), GMAT, Graduate Record Examination (GRE), and College-Level Examination Program (CLEP) Haberman (2011).

Unfortunately, these developed AEAS still come with their own form of bias, in form of architectural designs model in choice of algorithms and methodology used in building them. These AEAS models rely solely on statistical models and feature extraction techniques for essay assessment lacking semantic knowledge of words (Li et al., 2023).

Moving away from standardized bodies, individuals like Alikaniotis et al (2016), Taghipour and Ng (2016), Dong and Zang (2016), Dasgupta et al (2018) Kumar and Boulanger (2020) made use of various Deep Neural Network (DNN) models for development of contextual AEAS, achieving various Quadartic Weighted Kappa score of approximately 0.91, 0.76, 0.73, 0.97 and 0.80 respectively. DNN design based AEAS are said to achieve better accuracy for assessment purposes.

Poonpon *et. al..*, (2023) explored the effectiveness of techniques used for automated essay grading of non-native English users, and noted that the majority of AEAS grader is designed for the Inner Circle of English users, while neglecting the other circle of English users. The research gap identified includes cultural differences, and linguistic patterns. The study proposed a joint learning technique of various text representations in essays, and applied Long Short Term Memory algorithm for feature extraction and QWK result the model.

Doi *et al* (2024) worked on the development of AEAS in a bid to understand the implication of grammatical variety and error type in essay assessment scoring. The study made use of two linguistic features, focusing on the correctness of grammar used in essay and the errors resulting from grammar usage. This result shows that essay grammar plays a vital role in averaging high scores in holistic essay assessment score.

Faseeh *et al* (2024) developed a hybrid AEAS scoring model using Lightweight Extreme Gradient Boost (LwXGBoost) and RoBERTa for integration of deep learning embedding with handcrafted linguistic features for better accuracy in scoring. The study achieved a QWK score of 0.941, justifying the need for hybrid model design. Essays used for this study is from Kaggle ASAP dataset.

Ndukwe *et. al.*, (2020) utilizes Sentence BERT (SBERT) language model for assessment of short answers question for 228 essays on Computer Network courses. The Quadratic-Weighted Kappa (QWK) was used to test the agreement level between human assessor and the language model, on three variations of questions, including description, comparison and listing. Result shows that the model performed very well on the comparison and description questions compared to the listing question. The dataset used for this work is written in native Standard English.

Li *et al* (2023) designed AEAS system using SBERT. The study outlined limitations of deep learning AEAS models: inability to extract shallow linguistic features, and limited extraction of linguistic features based on sentences. As a solution, the work proposed a multi-layer scale features AEAS design, using SBERT for sentence vectorisation. The result from this study shows QWK score of 79.3% when tested on Kaggle and ASAP dataset.

Poonpon and Chansanam (2025) used an ensemble technique for the design of an AEAS system. The study used a combination of models: Bidirectional Encoding Representation from Transformer (BERT), Extreme Gradient Boosting (XGBoost), and Ridge Regression technique and achieved a high Quadratic Weighted Kappa (QWK) score a reduced Root Mean Squared Error (RMSE). It further used figures obtained from SHAP AI for analysis of feature importance. The study considered features from three groups: linguistic features (word count, the sentence and vocabulary difficulty), semantic features (TF-IDF and word vectorisation for identification of content importance) and essay specific feature (focusing on domain specific items that are necessary to score allocation in essay). Dataset for this work is totals of to 17,793 English essays: 12,976 from the Automated Student Assessment Prize (ASAP) dataset and 4,817 from the Khon Kaen University Academic English Language Test.

In all of these developed AEAS, one take home is that they were able to mitigate the challenges arising from traditional human grader, such as result processing time, raters fatigue, raters errors (consciously or un consciously), raters bias etc.

Xu *et al.,* (2024) did a systematic review of current AEAS systems, from 104 publications and concludes that despite best efforts of these AEAS models, teachers and writers still find them lacking. These findings were also supported by Ramesh and Sanampudi 2022. Existing AEAS have one thing in common: native circle English training dataset. As a solution, educational body like the World Variation of Englishes and various scholars have called for the development of AEAS that takes writers context into consideration (Zang 2021), (Poonpon et al 2021), (Vaijala 2018) and (Xu et al., 2024).

3. Methodology

To address grading bias, this work developed a Nigerian English classification model was using a combination of (TF-IDF) and K Nearest Neighbor (KNN) for feature extraction and text classification. It also made use of a secondary dictionary training feature that helps augment TFIDF vectorised feature, improving classification accuracy. The dataset is made up of 762,166 high-qualities crowd-sourced from open speech learning resources and essays. Preprocessing techniques, including lemmatization, part-of-speech tagging (POS), and word vectorization, were applied to refine linguistic characteristics. The classification model was evaluated using the F1-score, which measures precision and recall ensuring robust performance. Figure 1 illustrates the modified classification model for this work.

3.1 Data collection Process

Essay data collection was drawn from selected schools in Southern, Western and Eastern Nigeria regions. The schools falls into the public and private school categories, with students ranging from class 1 to 3 of Senior Secondary School. The essays were anomalyzed for removal of personal identifiable information. The essays were scanned using Microsoft lens and entered into the model for training.

A. Input

This is where the raw essay text enters the system. It could be pasted or uploaded as a text file.

B. Preprocessing

Here, the essay is prepared for analysis. This stage involves:

- Text cleaning: removing punctuation, special characters, and converting text to lowercase.
- Tokenization: breaking the essay into individual words or tokens.
- Lemmatization: reducing words to their base forms (e.g., “running” → “run”).
- Part-of-Speech (POS) tagging: labeling tokens by grammatical role (noun, verb, etc.).

This step ensures that the model focuses on the linguistic structure rather than irrelevant text noise.

C. Feature Extraction

This stage identifies useful patterns in the text that can distinguish Nigerian English from Standard English. It includes identification of specific vocabulary or expressions native to Nigerian English. These features act like “clues” the classifier later uses.

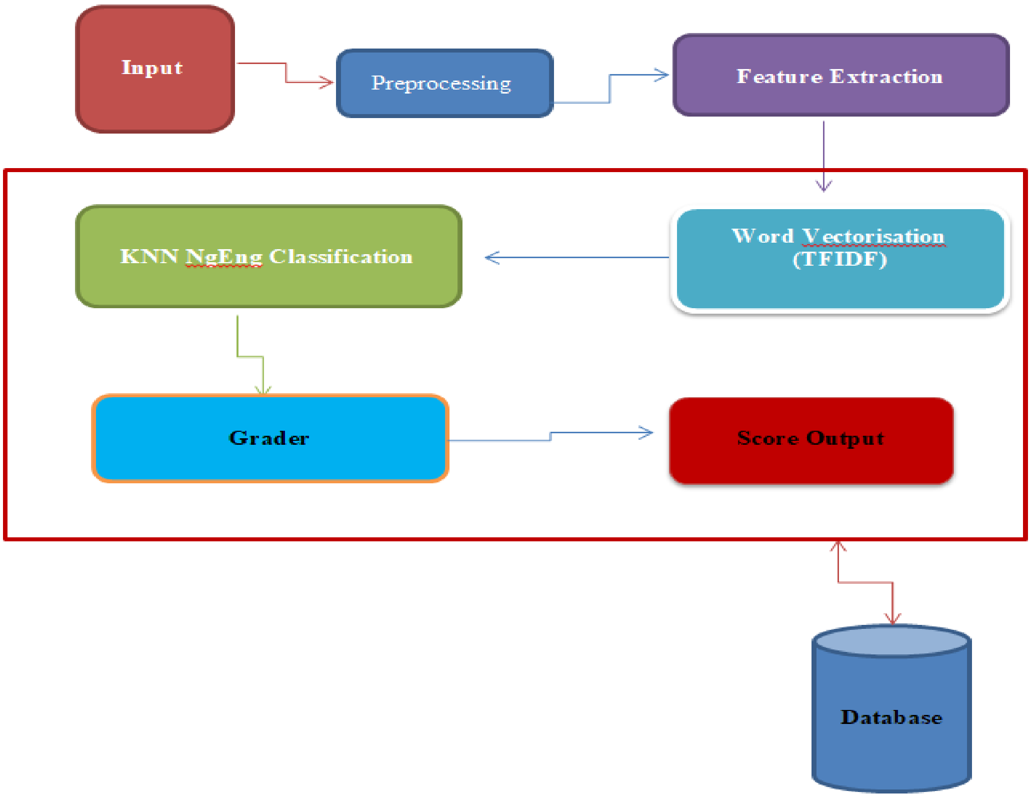

Figure 1: Modified classification model for Nigerian English

D. Word Vectorization

Words are converted into numerical representations using techniques like TF-IDF (Term Frequency–Inverse Document Frequency). These vectors reflect how important a word is within the essay and across the entire dataset

.

E. KNN Nigerian English Classifier

The K-Nearest Neighbors algorithm classifies words, phrases, or entire sentences as Nigerian English or not, based on similarity to examples in the training dataset. It works by:

- Measuring distance between the essay's feature vectors and known samples.
- Predicting the label (Nigerian English or not) based on the "votes" of nearby neighbors.

This stage helps filter out misclassified expressions and correct unfair bias before grading.

F. Grader

Now the system evaluates the essay's quality. This grader could apply rubrics like coherence, grammar, and relevance.

G. Score Output

The system generates a numeric grade based on the essay's alignment with assessment criteria and its effective use of Nigerian English.

H. Database

All inputs and final scores are stored in a central database.

3.5 Building the Nigerian English Classifier, Using KNN Algorithm

For this work, K-Nearest Neighbors (KNN) was used for text classification, with K values of 5, based on the extracted linguistic features captured. The dataset was split into training, testing, using the 5-fold cross-validation technique. This technique gave room for proper model performance evaluation, since the dataset was shared into five parts: four parts for training, and one part for testing iteratively.

4. Results and Discussion

The model demonstrated improved accuracy in recognizing and classifying Nigerian English lexical items, with accuracy score of 0.84, precision score of 0.83, and recall score of 0.77. It also gave an F1-score of 0.68, indicating moderate performance in detecting Nigerian English lexical features across diverse writing samples. Given the nuanced nature of Nigerian

English and its overlap with Native Standard English syntax and semantics, this result demonstrates the model's foundational ability to capture culturally relevant language use. However, the score also highlights areas for refinement, especially in handling context-dependent expressions and low-frequency lexical items." Table 3 gives a breakdown of the various metrics.

Table 1 gives an illustration of the linguistic features the NgEng classifier was modeled after. This features act as an anchor to the classification capacity of KNN.

Sample Essay Input:
"I eat puffpuff and Fanta for lunch. Dinner will be swallow, probably fufu or eba with egusi or okro soup. The boy died of poison because of his longthroat, always eating in random places—even in naming ceremonies or wake keepings. The girl's brideprice had to be reduced because she was disvirgined by her ex-boyfriend."

From the results shown in table 2, it can be seen that the classification model successfully cross-matched culturally rooted terms via lexicon lookup and learned from usage patterns in training data. The classifier was also able to combine both statistical and linguistic signals to detect uncommon and code-switched expressions.

Table 1 showing the design of the database structure for the NgEng Classification model

S/N	lexical	Synonyms	Antonyms	Sentence Example	POS	Lexical Definitions	Morphology	Alternate Spells	Related Words
1	Longthroat	Greed, gluttony	Generosity, selflessness	It is a fact that some people, including adults have longthroat.	Noun	To show unreasonable amount of interest in other people's belongings	Long+throat	Long throat	Food, money, attitude
2	Disvirgin	Deflower	Preserve, protect	She was disvirgined even before becoming an adult	Verb	Have sexual intercourse with someone who has never experienced sex	Dis+virgin	disvirgine	Lover, relationship, sex, female, innocence
3	Barbing salon	Barbershop		I need to visit the barbing salon, my hair is looking bushy		A shop where people get their haircut	Barbing+salon	Barbing saloon	Haircut, barber, rough hair, bush hair, shave, trim,

Table 2 illustrates the classifier output for Nigerian English lexical

Word / Phrase	TF-IDF+KNN Score	Lexicon Match	Flagged As Nigerian English
Puffpuff	✓	✓	✓
Fanta	✓	✗	✓
Swallow	✓	✓	✓
Fufu	✗	✓	✓
Eba	✗	✓	✓
Egusi	✓	✓	✓
Okro	✓	✓	✓
Longthroat	✓	✓	✓
Naming	✓	✗	✓(contextual)
wake keeping	✗	✓	✓
Brideprice	✓	✓	✓
Disvirgined	✓	✓	✓

Table 3: F1 score for the NgEng Classification Model built with KNN

KNN Classification Model
F1- score of 0.68
Precision 0.83
Recall 0.77.
Cross validation 0.80

Table 3 gives a breakdown of the evaluation metrics used for this study. The values show that the developed Nigerian English classifier is able to handle the classification task given to it. NgEng was able to recognize Nigerian English lexicals and classify them appropriately.

5. Conclusion

This study presented the design and implementation of a culturally-aware Nigerian English classification model, built to support automated essay scoring in a multilingual context. The classifier integrates a hybrid architecture that combines a TF-IDF-based KNN model trained on essay data with a linguistically enriched lexicon of Nigerian English terms. While the KNN classifier provides statistical insight into word usage patterns, the lexicon operates in parallel as a semantic knowledge base—allowing for the identification of low-frequency, culturally grounded expressions often ignored by frequency-based models.

Moreover, implementing attention-based models or transformer fine-tuning on domain-specific Nigerian English corpora would further enhance contextual sensitivity. A transparent visual interface for flagged terms, coupled with real-time feedback, could also increase pedagogical value in assessment environments. In sum, the classifier is not only a computational solution but also a sociolinguistic tool—offering fairer, more representative scoring for multilingual writers. It lays the groundwork for a new generation of Automated Essay Scoring systems that can be localized, explainable, and inclusive.

Future enhancements could integrate word embeddings (such as BERT trained on Nigerian English dataset) and syntactic dependency parsing to improve contextual accuracy. Additionally, incorporating human-in-the-loop validation will refine the model's interpretability, ensuring alignment with human grading standards.

References

Alikaniotis, D., Yannakoudakis, H., and Rei, M. 2016. *Automatic Text Scoring Using Neural Networks*. https://doi.org/10.18653/v1/P16-1068

Doi. K. K., Sudoh., K. and Nakamura, S. 2024. Automated Essay Scoring Using Grammatical Variety and Errors with Multi-Task Learning and Item Response Theory. *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 316–329. *Association for Computational Linguistics.*

Dong, F., and Zhang, Y. 2016. Automatic Features for Essay Scoring – An Empirical Study. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing.* (pp. 1072–1077). https://doi.org/10.18653/v1/d16-1115

Dada, S. A., Owolabi, D. and Bamigboye .O. 2018. Nigerian English as a Compulsive Variety in Braj Kachru's Concentric Circles. *Ebonyi Journal of Language and Literary Studies 1 (2) Volume 1, No. 2: 88–98 (2018) ISSN 9091 4582 7142.*

Dasgupta, T., Naskar, A., Saha, R., and Dey, L. 2018. Augmenting Textual Qualitative Features in Deep Convolution Recurrent Neural Network for Automatic Essay Scoring. *Aclweb.Org*, 93–102. Retrieved from http://aclweb.org/anthology/W18-3713

Faseeh, M., Jaleel, A., Iqbal,N., Ghani, A., Abdusalomov, A.,Mehmood, A. and Cho, Y.-I. 2024. Hybrid Approach to Automated Essay Scoring: Integrating Deep Learning Embeddings with Handcrafted Linguistic Features for Improved Accuracy. *Mathematics*,12, 3416.https://doi.org/10.3390/math12213416Academic

Haberman, S.J. 2011. Use of e-rater R in scoring of the TOEFL iBT R writing test. *ETS Research Report Series, 2011(1):i–13.*

Kachru, B. B. 2017. World Englishes and Culture Wars. *Cambridge University Press*. [Google Scholar]

Kachru, B. B. 1985. Standards, Codification and Sociolinguistic Realism: The English Language in the Outer Circle. In R.Quirk and H.G Widdowson (Eds.), English in the World: Teaching and Learning them Language and Literatures (pp.11-30). Cambridge University Press.

Kumar, V. S., Boulanger, D. 2020. Automated Essay Scoring and the Deep Learning Black Box: How

Are Rubric Scores Determined?. *Int J Artif Intell Educ*.
https://doi.org/10.1007/s40593-020-00211-5

Li, F., Xi, X., Cui, Z.: Li, D. and Zeng, W. 2023. Automatic Essay Scoring Method Based on Multi-Scale Features. Appl. Sci. 13, 6775. *Appl. Sci.* , *13*(11), 6775: https://doi.org/10.3390/app13116775

Ndukwe, G.I., Chwukudi, E.A., Nkomo, L.M. and Daniel, K.B. 2020. Automatic Grading System Using Sentence-BERT Network. *Artificial Intelligence in Education (pp.224-227). DOI:*10.1007/978-3-030-52240-7_41

Page, E. B. 1966. The imminence of grading essays by computer. *The Phi Delta Kappan, 47(5):238–243.*

Petersen, N.S. 1997. Automated scoring of written essays: Can such scores be valid? *Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago*

Poonpon, K., Manorom, P., and Chansanam, W. 2023. Exploring effective methods for automated essay scoring of non-native speakers. *Contemporary Educational Technology, 15(4), ep475. https://doi.org/10.30935/cedtech/13740*

Qiusi Zhang 2021. Impacts of World Englishes on local standardized language proficiency testing in the Expanding circle. A study on the College English Test (CET), in China. *English Today, Vol 38, Issue 4, pp 254-270 Doi:* https://doi.org/10.1017/S0266078421000158

Ramesh, D. and Sanampudi, S. K. 2022. An Automated Essay Scoring Systems: A Systematic Literature Review. *Artificial Intelligence Review*, 55(3), 2495-2527. doi: 10.1007/s10462-021-10068-2

Taghipour, K., and Ng, H. T. 2016. A Neural Approach to Automated Essay Scoring. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing* (pp. 1882–1891). https://doi.org/10.18653/v1/d16-1193

Xu, W., Mahmud, R. B., and Lam, H.W. 2024. A Systematic Literature Review: Are Automated Essay Scoring Systems Competent in Real-Life Education Scenarios? IEEE. DOI 10.1109/ACCESS.2024.3399163 https://www.researchgate.net/publication/380477783_A_Systematic_Literature_Review_Are_Automated_Essay_Scoring_Systems_Competent_in_Real-life_Education_Scenarios

Yang, X.; Yang, K.; Cui, T.; Chen, M.; He, L. 2022. A Study of Text Vectorization Method Combining Topic Model and Transfer Learning. Processes 10, 350. https://doi.org/10.3390/pr10020350