Latest Advances in Automated Essay Scoring: A Survey of Machine Learning and Deep Learning Methods

Khawar Iqbal Malik[1*], Dr. Qaisar Abbas[2], Waqas Ahmad Khan[3] and Hira Arooj[4]

[1]Riphah School of Computing and Innovation, Riphah International University, Lahore, Pakistan
[2]Department of Computer Science and IT, Islamic University of Madinah, Saudi Arabia
[3]Department of Computer Science, University of Sargodha, Pakistan
[4]Department of Mathematics and Statistics, University of Lahore, Sargodha Campus, Pakistan

ABSTRACT The current study aims to assess the reliability of automated essay scoring (AES) through the comparison of the mean scores assigned by an AES tool in the context of a growing educational institution with a rising student population. A survey was conducted to test the reliability and validity of the E-Grading device, as well as to evaluate the use of holistic scores generated by both human and computer scoring, as a better solution for AES systems. While previous research found no significant mean score differences between human and AES scoring, this paper does not confirm these findings. In recent years, several algorithms have been proposed for AES and comparative studies have been conducted to evaluate the effectiveness of these algorithms. Instead, it reviews and examines earlier concepts and techniques applied in AES.

INDEX TERMS automated essay scoring (AES), BILSTM, grading and feedback, human raters, natural language processing (NLP), rubric

I. INTRODUCTION

An essay is a piece of text written in response to a prompt. It is a crucial testing tool used to measure academic achievement in educational institutes. Through essays, students can better explain and recall their knowledge. Many educational institutes are moving towards free-text responses to judge their students' abilities. Subjective type exams are conducted to evaluate students' ability, analytical clarity, and overall progress in the subject [1].

Natural Language Processing (NLP) is the branch of artificial intelligence (AI) which encompasses several theories and methods that enable us to analyze and understand the meaning of words, similar to how human brain comprehends them. With the aid of NLP, individuals can efficiently deal with language-based data, thanks to its understanding of grammar, vocabulary, syntax, enhanced algorithms, and powerful computing capabilities. Machine learning (ML) methods, as well as rule-based and statistical approaches, are utilized in NLP with sentence segmentation, normalization, and syntactic parsing being their distinctive features. Automated essay scoring (AES) analyzes students' responses, assigns grades, and provides feedback on how they can improve their thinking and writing abilities. AES applies to both long and short answers and scores are assigned based on linguistic criteria [2].

*Corresponding Author: khawar.iqbal@riphah.edu.pk

There are two primary reasons for conducting this study. Firstly, the corpus is manually annotated with holistic scores, which facilitates the development of learning-based holistic score engines. Secondly, holistic score engines are highly valued commercially as they enable the automation of grading millions of essays written for tests like SAT (Scholastic Aptitude Test), GRE (Graduate Record Examination), and TOEFL (Test of English as a Foreign Language). They have the advantage of being able to save a significant amount of time and effort in manual grading.

The objective of the current study is to build and implement a deep learning-based AES system that assesses and assigns grades to essays. AES is commonly used to assign grades to students in educational institutions. Our model would be beneficial for educational institutes to conduct exams and assign grades to students, given the recent advancements in machine learning and deep learning. AES is considered a useful tool to enhance human creativity in essay writing. This model is expected to address the statistical problems that arise when classifying large texts into small categories that match the corresponding score. This study predicts the writing style quality of Grade 7-10 students who describe postulates based on the Assessment Students Prize Datasets (ASAP) [3]. In most educational institutions worldwide, essays are an essential part of their curriculum to assess the overall progress of students. However, early ML-based AES systems lacked accuracy due to various reasons, including excessive consumption of resources, such as time, space, and human effort. Furthermore, the same essay may be assigned different grades for different students, as each teacher has a different perception and knowledge of the particular student. Manual feature extraction is also difficult as it is ambiguous to assign grades manually. Such problems can cause confusion in the minds of students and ambiguity in essay grading results. Hence, we propose employing a DL framework to enhance essay grading. Previous techniques, such as RNN, struggled with longer texts, frequently losing their initial context and meaning. Our strategy aims to address these challenges more effectively to deliver superior outcomes.

II. LITERATURE REVIEW

This paper explores the role of linguistic evaluation in interpreting natural language[4]. Additionally, it clarifies how the rich lexicon and intricate syntax structure of Arabic may render interpreting text challenging. The study suggests a novel method to enhance the accuracy of Arabic sentiment analysis in order to address these issues. The researchers tested it on a variety of data sets and obtained favorable results, proving that the strategy is effective [4]. The study shows that a linear regression framework can be used to investigate various aspects of the text, such as phrase and word size, paragraph length, word count, and even uncommon words. The goal is to discover the connections or patterns among these attributes and the larger dataset. Linear regression is a widely used scientific approach that establishes a direct relationship between variables, making it easier to understand how one variable affects another.

Scientists have been investigating ways to enhance the quality of NLP by combining statistical algorithmic systems based on rules and artificial neural networks (ANNs). For instance, rules like the ‘Rule of Apology’ assist computers to

comprehend the particular kind of content, such as the phrases of apology.

The study also examines two commonly employed methodologies to AES, such as the use of RNN-based techniques and the Enhanced AI Scoring Engine. RNN variants are considered stronger due to their statistical orientation. Even though it is not frequently apparent how the model determines what attributes are the most important, its findings show that it does so successfully.

Subsequently, the review examines multiple research projects on AES systems, including creation testing and comparison. It addresses the use of ML-based NLP and DL-based methodologies in these systems, while also highlighting their drawbacks, such as challenges in identifying originality or writing styles. Overall, the review paints a picture of the current state of AES studies, emphasizing the need for additional advancement in this area [5].

C-rater first employed the Goldmap answer identification algorithm, a rule-based technique that yielded straightforward yes or no answers (0 or 1). While easy to understand, this approach lacks freedom and can occasionally cause confusion. The probabilistic methods, such as the Naïve Bayes algorithm, tend to be more flexible and appropriate to deal with data variations.

To determine how well C-rater's outcomes matched those with actual graders, the quadratic weighted kappa technique was used. The experiments were conducted in Python, with packages such as NumPy, SciPy, NLTK, and mining text tools. Word strength, length, and tokenised text was obtained and analysed. Every attribute was compared to human assessments and kappa values were determined. Less beneficial attributes were slowly eliminated through the addition of a single function at a time, according to that of priority.

In general, the paper demonstrates how C-rater, a system intended to autonomously score short responses, can be enhanced by switching from rule-driven to probabilistic approaches. The study demonstrates that Naïve Bayes, when combined with appropriate feature mining and appraisal techniques, can improve the performance of automated grading systems [6].

A. AES USING MACHINE LEARNING

The methodology is explained in exhaustive detail, starting with the initial processing of the essay to eliminate unnecessary details and to draw a standardize layout. Subsequently, NLP methods were employed to extract feature characteristics from the documents, such as the number of specific words or phrase length. The discovered features were fed into a model based on machine learning. The model was trained on a set of edited essays to predict the grade of the newly created essays. The system was tested on a sample of essays produced by middle school students. The investigators noticed that it outperformed human judges in terms of accuracy. They also mentioned that the automated system showed the capacity to provide students with more comprehensive critiques than human instructors, which might enhance their writing skills.

In general, the particular paper illustrates the possible application of ML methods to automate essay grading. This might conserve both time and money for teachers, while offering more accurate and comprehensive feedback to students [7].

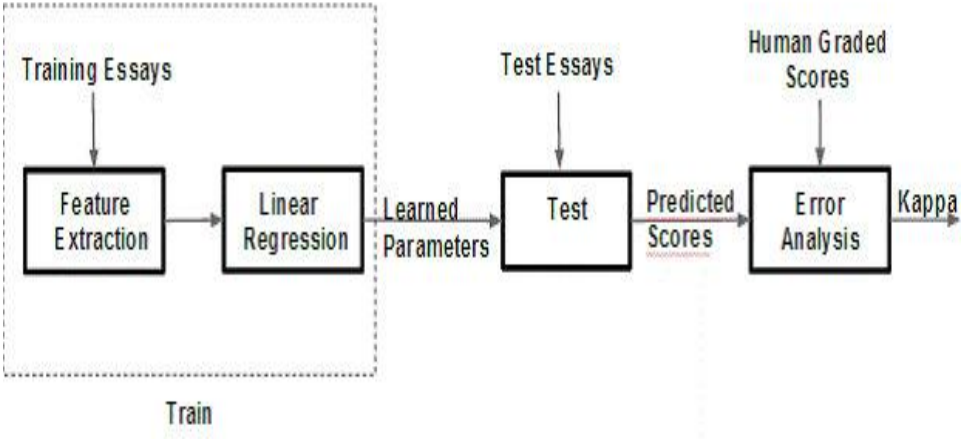

FIGURE 1. Linear regression methodology [7]

B.

TEXT CLASSIFICATION USING KNN

The research suggests a method for assigning personal texts to established groups and scoring the results based on similarities with the training data. It outlines the steps of preprocessing used on the raw text data, such as stemming, stop word removal, and selecting features via term frequency-inverse document frequency (TF-IDF). The researchers applied the K-nearest neighbor (KNN) technique to divide the provided text data into specific groups, based on the similarities to training data. The suggested approach was tested on a collection of subjective written information to determine its efficacy. The results showed that the proposed method achieved outstanding precision in categorizing and scoring textual information.

The study indicates that the proposed technique has numerous applications in a variety of fields, including sentiment evaluation, data mining, and system optimization. It also implies that the KNN algorithm can be a successful technique to identify texts and assessment tasks [8].

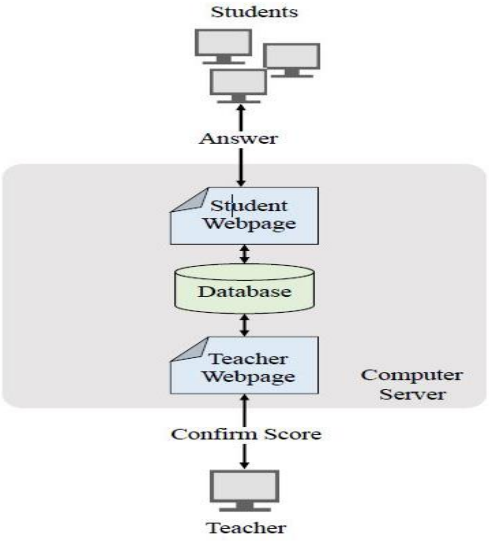

FIGURE 2. KNN scoring system [8]

C. GRADING SYSTEM USING CHATBOT

An innovative method was used to implement the grading process through chatbots, alongside ML algorithms. The

researchers suggested an assessment system that employed NLP techniques to analyse the answers to questions regarding essays and also assigned them a rating depending on their quality. The system was developed on a dataset of graded writings and imitated the method of assessment used by classroom teachers. It was anticipated to reduce the instructors' workload, while providing students with immediate advice on their grades. The recommended system's functionality was assessed using a limited number of scholar responses. The outcomes showed that the algorithm can grade essays with great precision. The study concluded by addressing the proposed approach's effects on education and recommending areas for further study.

The study determined the value of Cohen's Kappa using the following formula:

$$K = (na-ne) / (n-ne) = \quad 0.6 \qquad (1)$$

In the above scenario, 'k' indicates the value of kappa, 'n' signifies the entire number of students, 'ne' reflects the number of choices by chance, and 'na' stands for the total number of agreements [9].

D. SUBJECTIVE ANSWER GRADER SYSTEM (SAGS)

This particular review provides a summary of recent advances in the field of automatic short answer assessment using methods based on deep learning. Different strategies used for automated grading of short answer queries are discussed, which include feature-based methodologies, neural network-inspired techniques, and combination methods that incorporate multiple technologies [10]. An in-depth review of the latest developments in automated short answer assessment using DL methods is offered. It contains pre-graded answers, as well as actual grader scores ranging from 0 to 3 marks [11].

The method for automating long answer evaluation with ML methods and terminologies is described. The authors proposed a system that requires specific ontological expertise and evaluates subjective answers provided by students using an ML-based algorithm. The technique undergoes training on a dataset of predetermined answers. Afterwards, it uses the ontology to identify specific characteristics, while the ML algorithm determines the subjective answer's score. The authors used arbitrary answers to assess the efficiency of their system and compared their findings to other techniques. Various studies showed that the suggested system can accurately evaluate subjective responses and is superior to alternatives. The study concluded by addressing the possibilities of their approach towards improving educational evaluation, while highlighting areas for further study. Overall, the paper presents a novel technique for automated human answer evaluation utilising machine learning along with semantic techniques [12].

Finally, the results were calculated using the KNN method with a value of 2. The selected two highest-levels of already graded answers were used to calculate the grade for the student's given answer when reacting to a prompt.

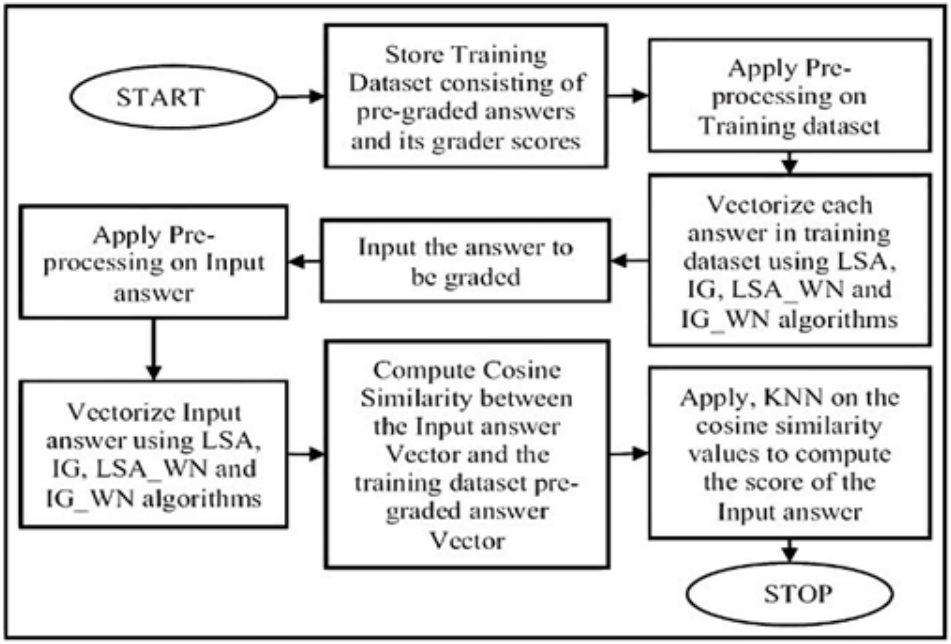

FIGURE 3. Model diagram of SAGS [11]

E. SUBJECTIVE ANSWER EVALUATION SYSTEM

The review describes a novel technique for automated subjective answer assessments based on NLP and ML methodologies. It provides an approach to train the ML model on the dataset of previously evaluated answers before using it to evaluate new answers provided by the students. The system starts with preprocessing steps including tokenization, stemming, and stop word removal, followed by determining key characteristics from the answer. The outcomes of various machine learning techniques, such as Support Vector Machine (SVM), Random Forest (RF), and Naive Bayes, are presented. The challenges faced in developing an accurate and efficient system are addressed, especially when dealing with similar words, homophones, and grammatical mistakes. The researchers used subjective answers to assess the success of their system and then contrasted their findings to those of human evaluator. The findings showed that the suggested system can accurately evaluate subjective responses and perform similarly to human evaluator. The researchers finished by addressing the possibility of their strategy to improve educational assessment and highlighted areas for further study. Overall, the paper presents an innovative method to automate subjective answer assessment utilizing NLP and ML approach [13].

Basically, this model considers four types of similarities which are as under.

1) COSINE SIMILARITY

In space with n dimension, cosine similarity determines the correlation among the vector representations of both words. It is irrelevant if the size of the two words are distinct; the cosine similarity provides a precise similarity metric.

2) JACCARD SIMILARITY

Following the initial processing, the algorithm generates a pair of words: the answer within the study and the reference answer. The connection is used to indicate common words, while the combination is utilised to integrate a list of comparable words [14].

3) BIGRAM SIMILARITY

The term similarity refers to different forms of language that function independently of one another. It is used to determine the analogy among two consecutive collections of words of any length.

Synonym analogy: If the test subject's answer is different from the model answer, it is assigned an aggregate score. If the substitute answer is equivalent to the prototype answer or an alternative of the substitute answer is discovered, it is deemed to be the right answer [15].

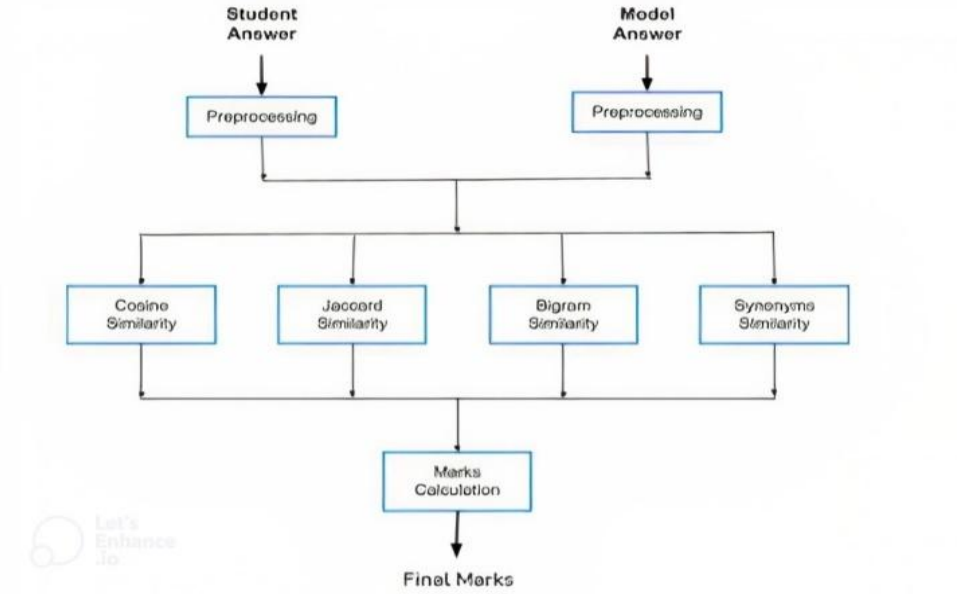

FIGURE 4. Answer evaluation using similarity measure [13]

F. HYBRID FRAMEWORK OF AES

It presents a method for automated essay assessment in Bahasa Indonesia using latent semantic analysis. The method is designed to assess essays based on their significance rather than fundamental features, such as grammar and spelling. The authors tested their method on a collection of essays to evaluate the results and to compare them with other methods already in use. The findings revealed that their system was more accurate for assessing essays in Bahasa Indonesia [16]. The paper introduces a relevancy-driven automated essay evaluation method built upon an organizational recurrent model. The method attempts to assess essays based on their material relevance to the chosen topic, rather than examining surface level characteristics only, such as punctuation and spelling. The investigator evaluated the system on a collection of essays and compared its effectiveness to other standard methods. The results showed that the proposed strategy surpasses other

techniques in term of its precision and efficacy [17].

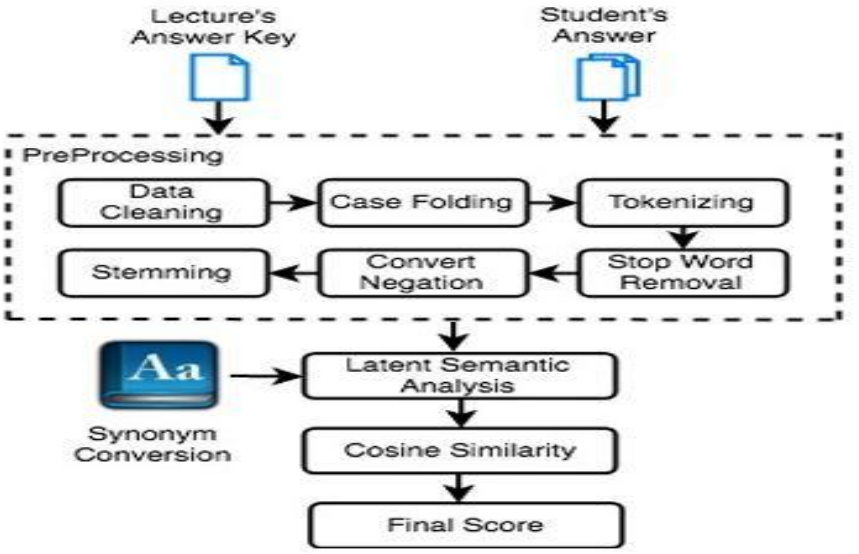

FIGURE 5. Methodology[16]

G. AES-BASED STEMMING TECHNIQUE

This paper presents an automated system for grading essays written in Arabic. The proposed system uses stemming technique and Levenshtein edit distance to process the input text and calculate the score. The system was tested on a dataset of 100 essays (written by students) and achieved a high level of accuracy in grading them. The paper also discusses the limitations of the system and suggests areas for future research [18].

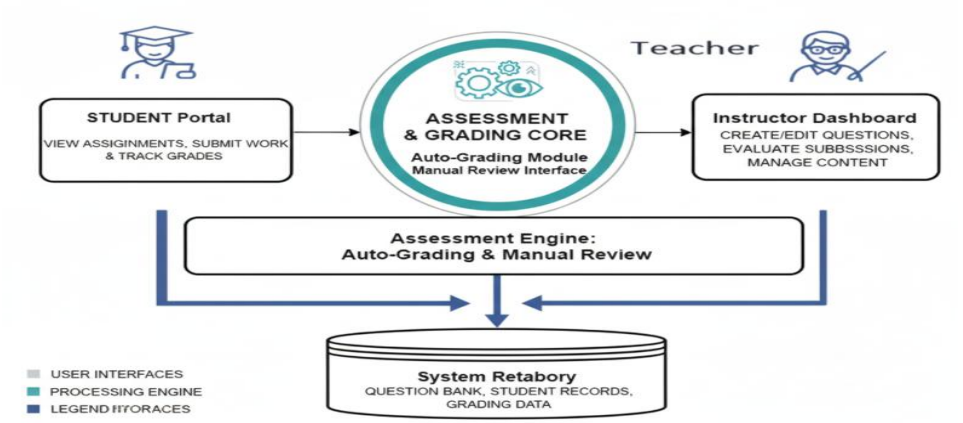

FIGURE 6. Model diagram of stemming based AES [18]

H. ESSAY SCORING BASED ON FEATURE OPTIMIZATION

Figure 7 shows the overall system architecture of the AES system based on machine learning. This paper investigates the impact of feature optimization on the accuracy of AES systems. The authors used a dataset consisting of essays written by Korean students and explored various feature selection and dimensionality reduction techniques to identify the most relevant features for essay scoring. The results showed that optimizing the feature selection process significantly improved the accuracy of the AES system [19]. After that, the labeled essay is entered into the training module. The results of the training model depend on the learner models. At the end, the training model was used to compute the final grades used in the predictor model [20].

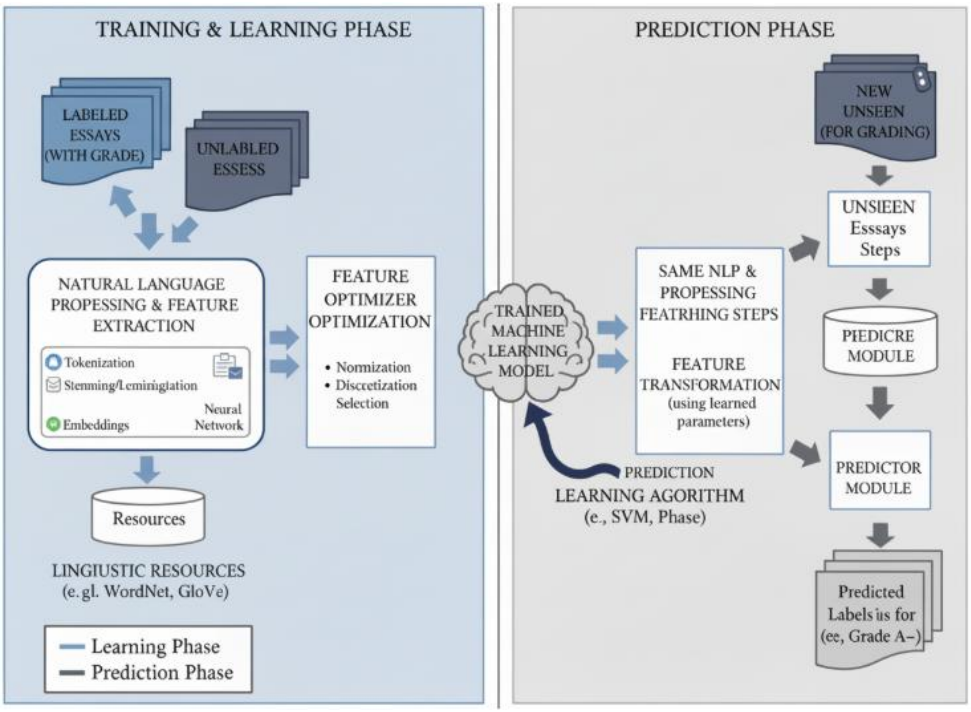

FIGURE 7. ML-based AES system [19]

I. MULTI MODEL MACHINE LEARNING

Sun introduced an innovative AES system which employs numerous ML models to enhance assessment accuracy. The method uses a total of three models: bag-of-words, convolutional neural network (CNN), and long short-term memory (LSTM). The bag-of-words model captures the essay's fundamental traits, whereas the CNN and LSTM models identify deeper traits, including sentence construction and background. The experimental findings showed that the suggested model exceeds several conventional AES systems with regard to reliability and precision [21].

The article discusses AES along with its various uses. It begins with an outline of

AES, followed by a description of how the technology works, the different approaches used, and the difficulties and obstacles associated with this technology. The researchers proceed to address the benefits and drawbacks of AES in comparison with conventional human grading methods, as well as its possible effects on the educational system. They find that, while AES technology continues to evolve, it is still not an alternative for human grading and should be utilized only additionally in the process of learning [22].

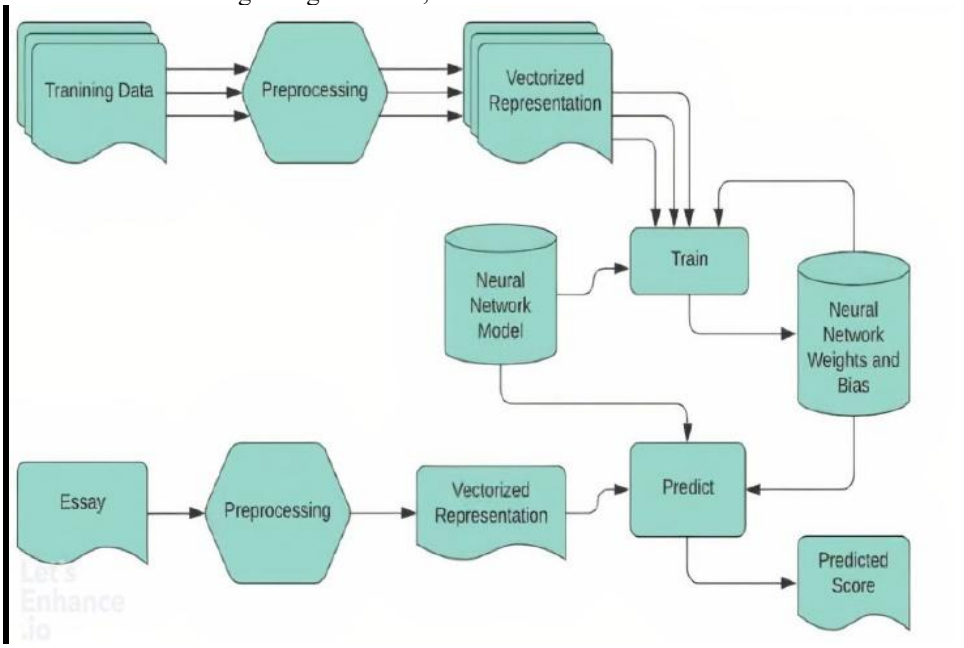

FIGURE 8. Model diagram of multi model [21]

J. AES USING HIERARCHICAL RECURRENT MODEL

K. Taghipour and H. T. Ng presented an AES method based on neural networks in their research. They used an LSTM network, developed with a ranking loss algorithm, to calculate an essay's score using the question being asked and essay text. In two separate datasets, the suggested approach outperformed a variety of baseline models. In addition, the researchers conducted a series of experiments to assess the neural network's behaviour and the impact of multiple characteristics on the model's efficacy [23]. Yang proposed an automated assessment system based on a neural network model that combines convolutional and recurrent neural networks with a method for applying attention. The suggested approach was trained using a corpus of student essays and their scores. It performed well in comparison with other innovative methods. The investigators further examined the attention method to understand the parts of the essay the model focuses on as it formulates its predictions [24].

Figure 9 represents the essay as E and the question as P to learn the current model. Then, to get the appropriate subject matter for essay and prompt, component-by-

component multiplication is done. Figure 9 shows the current model for essay scoring with prompt awareness. The essay (E) and the prompt (P) are first converted into word embeddings and then passed through sentence and document level encoders. This step helps the system capture both the local meaning and the overall structure. In the interaction zone, two operations take place. Concatenation joins the features of essay and prompt, while element-wise multiplication highlights where they align closely. This design allows the model to check not only the quality of the essay but also its relevance to the prompt. The combined features then move to a scoring layer with a sigmoid function, which produces the final predicted score in a smooth and reliable way.

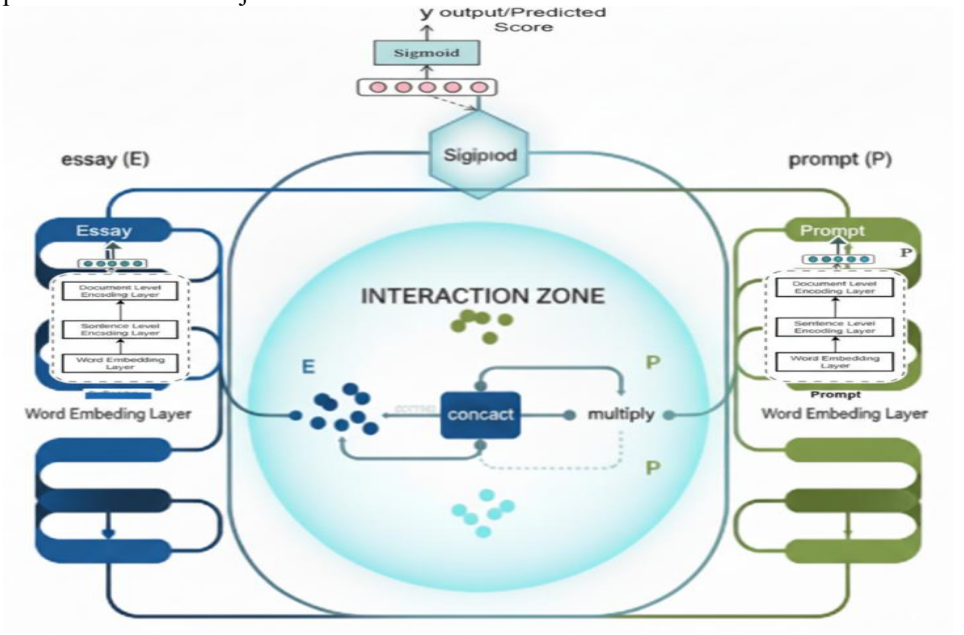

FIGURE 9. Hierarchical recurrent model [17]

K. AES USING TREE-LSTM

Engineering and Systems in 2020 proposed an AES system for short essays written in the Indonesian language. The proposed system uses a combination of transfer learning and Dependency Tree LSTM (DT-LSTM) to capture the semantic and syntactic features of the essays. The authors also conducted experiments on a dataset of short essays written by high school students. They demonstrated that the proposed system achieved better performance, as compared to several baseline methods. A value of 1 indicates retaining the information, while a value of 0 indicates ignoring it. The weight of h_t-1 is w_{fh} and the weight of the input x is $w_{fx,}$ while the bias is calculated as

$$f_t = \sigma(W_f [h_{t-1}, X_t] + b_f) \quad 2$$

In the source corpus, every word's vector in each essay sentence constitutes the input data. Next, DT-LSTM was employed on the source data, followed by modeling. The initial weights were assigned randomly but updated subsequently after each batch. Various factors impact or update the

weights. However, in this investigation, the loss function and multiple cross-entropy function governed the weight modifications [25].

The hidden state h_t is obtained after the processing of DT-LSTM and then applied to the softmax function for classification. The whole process of DT-LSTM is shown below in Figure 10.

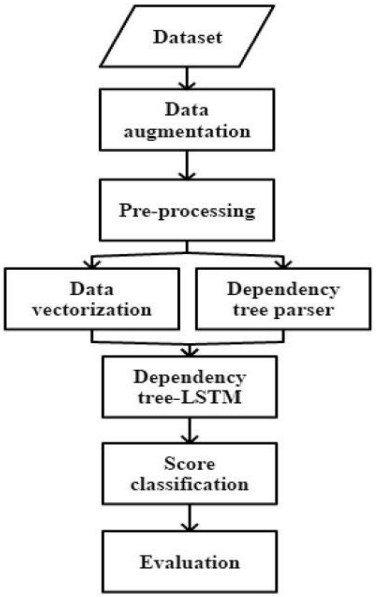

FIGURE 10. Model diagram of DT-LSTM model [25]

III. CONCLUSION AND FUTURE WORK

Over the past few years, the deployment of machine learning (ML) and deep learning (DL) methods has piqued the researchers' interest to develop automated essay scoring (AES) systems. The objective is to assess students' free-text replies and contextual writings with precision. Through the discussion of various methods and techniques employed in this field, the latest research on AES and its potential applications in education have been highlighted. The examination of ML-based AES systems revealed that they often rely on the contextual meaning of the text, with six classes (A, B, C, D, E, F) typically used for their classification. However, while such systems show promise, many existing models have limitations in terms of feedback and coherence, which results in lower accuracy scores.

Despite of these drawbacks, AES systems possess a substantial potential to assist education through all its stages and especially during difficult times, such as the recent COVID-19 pandemic, since when online classes have become more common. As a result, it is suggested that new techniques are needed to improve the precision and efficiency of automated essay scoring.

Incorporating domain-specific expertise and NLP techniques into DL models is a promising approach. By utilizing such resources, the effectiveness of AES systems can be improved, especially when evaluating the fundamental ideas of the essays. For example, DL models, such as LSTM and BILSTM, demonstrate excellent results in automating the assessment of plain-text answers and corresponding scripts.

To conclude, while ML-based AES systems have drawbacks, DL-based techniques can significantly enhance the accuracy of automated essay evaluation. Utilizing domain-specific knowledge and NLP techniques can improve the efficacy of such systems for assessing the quality of essays, promoting education at all levels, and addressing the difficulties created by distance learning.

TABLE 1

MODEL COMPARISON FOR AES

Domain Name	Algorithms	Feature Extraction	Datasets	Scoring Task	QWK	Evaluation Measures		Accuracy
						Z-Score	RMSE	
Machine Learning	Logistic Regression	TDM	The Hewlett Foundation: Automated Essay Scoring Data set[1]	Holistic	0.72	-1.45	0.41	0.68
Machine Learning	KNN	TF-IDF		Holistic	0.64	-1.27	0.37	0.74
Machine Learning	Chatbots	TF-IDF		Holistic	0.60	-1.63	0.45	70
Machine Learning	LSA, IG, LSA_WN and IG_WN algorithms	TDF matrix		Organization	0.68	-0.98	0.33	44,16,81 and 83
Machine Learning	SVM and LSA	Statistic (Word embedding)		Prompt Adherence	0.68	-0.97	0.37	0.89
Machine Learning	SVM and SVR	GloVe		Organization	0.69	-0.94	0.32	0.80
Deep Learning	Stemming-Based Mechanism	W2C		Organization	0.66	-1.15	0.35	0.77
Deep Learning	LSTM and GRU	GloVe		Prompt Adherence	0.71	-0.80	0.30	0.83
Deep Learning	LSTM-CNN-attention	GloVe		Persuasiveness	0.74	-0.60	0.25	0.90
Deep Learning	Dependency-tree LSTM	GloVe		Persuasiveness	0.72	-0.75	0.28	0.85

[1] https://www.kaggle.com/competitions/asap-aes

In the future, DL-based methods should be focused because most of the work in this field has been done using ML-based methods, with little emphasis on deep learning. The current approach would involve using LSTM and BILSTM models to automate the evaluation of free-text responses and students' contextualized scripts. Additionally, other DL variants should be explored to improve the performance of the AES systems.

CONFLICT OF INTEREST

The author of the manuscript has no financial or non-financial conflict of interest in the subject matter or materials discussed in this manuscript.

DATA AVAILABILITY STATEMENT

Data will be provided by corresponding author upon reasonable request.

FUNDING DETAILS

This research received no external funding.

REFERENCES

[1] B. Bohnet, R. McDonald, G. Simoes, D. Andor, E. Pitler, and J. Maynez, "Morphosyntactic tagging with a meta-BiLSTM model over context sensitive token encodings," 2018. [Online]. Available: https://arxiv.org/abs/1805.08237.

[2] D. Boulanger and V. Kumar, "Deep learning in automated essay scoring," in *Proc. Intell. Tutor. Syst. 14th Int. Conf., ITS 2018,* Montreal, QC, Canada, June 2018, pp. 294–299.

[3] C. T. Lim, C. H. Bong, W. S. Wong, and N. K. Lee, "A comprehensive review of automated essay scoring (AES) research and development," *Pertanika J. Sci. Technol*., vol. 29, no. 3, pp. 1875–1899, 2021, doi: https://doi.org/10.47836/pjst.29.3.27.

[4] H. Elfaik and E. H. NfaouI, "Deep contextualized embeddings for sentiment analysis of Arabic Book's reviews," *Proc. Comput. Sci.*, vol. 215, pp. 973–982, 2022, doi: https://doi.org/10.1016/j.procs.2022.12.100.

[5] M. A. Hussein, H. Hassan, and M. J. P. C. S. Nassef, "Automated language essay scoring systems: A literature review," *Peer. J. Comput. Sci.*, vol. 5, 2019, Art. no. 208, doi: http://doi.org/10.7717/peerj-cs.208.

[6] J. Z. Sukkarieh and J. Blackmore, "c-rater: Automatic content scoring for short constructed responses," in *Proc. 22nd Flairs Conf.*, 2009, pp. 290–295.

[7] M. Mahana, M. Johns, and A. *Apte. Automated essay grading using machine learning*. Stanford University, 2012. [Online]. Available: https://cs229.stanford.edu/proj2012/MahanaJohnsApte-AutomatedEssay GradingUsingMachineLearning.pdf

[8] K. Sriwanna, "Text classification for subjective scoring using K-nearest neighbors," in *Int. Conf. Digit. Arts Media Technol.*, 2018, pp. 139–142, doi: https://doi.org/10.1109/ICDAMT.2018.8 376511.

[9] I. G. Ndukwe, B. K. Daniel, and C. E. Amadi, "A machine learning grading system using chatbots," in *Artif. Intell. Edu. 20th Int. Conf.,* AIED 2019, Chicago, IL, USA, June, 2019, pp. 365–368.

[10] S. Bonthu, S. Rama Sree, and M. K. Prasad, "Automated short answer grading using deep learning: A survey," in *Mach. Learn. Knowledge Extract. 5th IFIP TC 5, TC 12, WG 8.4, WG 8.9, WG 12.9 Int. Cross-Domain Conf.*, Virtual Event, August, 2021, pp. 61–78.

[11] A. Sakhapara, D. Pawade, B. Chaudhari, R. Gada, A. Mishra, and S. Bhanushali, "Subjective answer grader

system based on machine learning," in *Soft Comput. Signal Proc. Proc. ICSCSP 2018*, 2019, pp. 347–355.

[12] M. S. Devi and H. Mittal, "Machine learning techniques with ontology for subjective answer evaluation," 2016. [Online]. Available: https://doi.org/10.48550/arXiv.1605.02442.

[13] A. Kumar, A. Kharadi, D. Singh, and M. Kumari, "Subjective answer evaluation system." Available: https://www.academia.edu/download/73054812/IJCST_V9I5P10.pdf.

[14] H. Henderi, Henderi Henderi, and W. Winarno, "Text mining an automatic short Answer Grading (ASAG), comparison of three methods of cosine similarity, Jaccard similarity and Dice's coefficient," *J. Appl. Data Sci.*, vol. 2, no. 2, pp. 45–54, 2021, doi: https://doi.org/10.47738/jads.v2i2.31.

[15] A. M. B. Omran and M. J. Ab Aziz, "Automatic essay grading system for short answers in English language," *J. Comput. Sci.*, vol. 9, no. 10, pp. 1369–1382, 2013, doi: https://doi.org/10.3844/jcssp.2013.1369.1382.

[16] A. Amalia, D. Gunawan, Y. Fithri, and I. Aulia, "Automated Bahasa Indonesia essay evaluation with latent semantic analysis," *J. Phy. Conf. Ser.*, vol. 1235, no. 1, 2019, Art. no. 012100, doi: https://doi.org/10.1088/1742-6596/1235/1/012100.

[17] M. Chen and X. Li, "Relevance-based automated essay scoring via hierarchical recurrent model," in *Int. Conf. Asian Lang. Process.*, 2018, pp. 378–383

[18] E. F. Al-Shalabi, "An automated system for essay scoring of online exams in Arabic based on stemming techniques and Levenshtein edit operations," 2016. [Online]. Available: https://doi.org/10.48550/arXiv.1611.02815.

[19] B.-J. Yi, D.-G. Lee, and H.-C. Rim, "The effects of feature optimization on high-dimensional essay data," *Math. Prob. Eng.*, vol. 2015, 2015, Art. no. 21642, doi: https://doi.org/10.1155/2015/421642.

[20] T. B. Adji, Z. Abidin, and H. A. Nugroho, "System of negative Indonesian website detection using TF-IDF and Vector Space Model," presented at the 2014 Int. Conf. Elect. Eng. Comput. Sci., Kuta, Bali, Indonesia, Nov. 24–25, 2014, pp. 174–178.

[21] W. Zhu and Y. Sun, "Automated essay scoring system using multi-model machine learning," *Comput. Sci. Info.*, vol. 10, no. 12, pp. 109–117, doi: https://doi.org/10.5121/csit.2020.101211.

[22] A. Lukic and V. J. R. U. Acuna, "Automated essay scoring," 2012. [Online]. Available: http://www.alenlukic.com/assets/docs/aes_report.pdf.

[23] K. Taghipour and H. T. Ng, "A neural approach to automated essay scoring," in *Proc. 2016 Conf. Empir. Meth. Nat. Lang. Proc.*, Austin, Texas, USA, Austin, Texas, Nov. 1–5, 2016, pp. 1882–1891.

[24] F. Dong, Y. Zhang, and J. Yang, "Attention-based recurrent convolutional neural network for automatic essay scoring," in *Proc. 21st Conf. Comput. Nat. lang. Lear.*, Vancouver, Canada, 2017, pp. 153–162.

[25] A. Wiratmo and C. Fatichah, "Indonesian short essay scoring using transfer learning dependency tree LSTM," *Int. J. Intell. Eng. Syst.*, vol. 13, no. 2, Jan. 2020, doi: https://doi.org/10.22266/ijies2020.0430.27.