PeerJ Computer Science

NepAES: exploring the promise of automated essay scoring for Nepali essays

Sweta Poudel[1,*], Kritesh Rauniyar[2,3,*], Ashish Acharya[4], Junaid Rashid[5], Surabhi Adhikari[6], Usman Naseem[7] and Surendrabikram Thapa[8]

[1] Kathmandu Engineering College, Tribhuvan University, Kathmandu, Nepal
[2] Delhi Technological University, New Delhi, India
[3] IIMS College, Kathmandu, Nepal
[4] Kathmandu University, Kathmandu, Nepal
[5] Sejong University, Seoul, Republic of South Korea
[6] Columbia University, New York, United States
[7] Macquarie University, Sydney, Australia
[8] Virginia Polytechnic Institute and State University (Virginia Tech), Blacksburg, United States
[*] These authors contributed equally to this work.

ABSTRACT

Natural language processing (NLP) has been extensively studied and developed for the purpose of automated essay scoring (AES). This field of research has attracted significant attention and has been explored across multiple languages and machine-learning models. Researchers, over the years, have dedicated significant resources to enhance the accuracy and dependability of AES systems. Multiple studies have shown that by utilizing advanced NLP approaches, AES models can attain performance levels that are equivalent to those of human evaluators. This has been accomplished by continuously improving algorithms and using extensive datasets for training, enabling these models to gain a deeper understanding of the contents of the essay and evaluate the subtleties of written language. However, these systems have primarily been developed for English and other high-resource languages. Nepali, which is a low-resource language based on the Devanagari script, remains unexplored in the context of AES due to its complex script formation and low research effort. In this article, we prepare a large translated dataset using machine translation algorithms and evaluate the efficiency of various machine learning and deep learning models in Nepali AES using scores like the Quadratic Weighted Kappa (QWK) score. For the classical machine learning (ML) approach, we used a feature-based method. Meanwhile, for state-of-the-art transformer-based models, we fine-tuned the models based on the transformer architecture. Our findings demonstrate that the effectiveness of AES systems is greatly influenced by the quality of translations, as the accuracy and precision of the translation process have a direct impact on the overall performance of the AES models. By comparing various models using the QWK score, we have demonstrated that fine-tuned transformer architectures perform quite similar to the traditional feature-based ML method. Our research efforts are a step further in enabling deep learning and artificial intelligence (AI) access to the Nepali-speaking community. The dataset is available at https://github.com/rkritesh210/NepAES.

Submitted 5 May 2025
Accepted 8 September 2025
Published 14 October 2025

Corresponding author
Surendrabikram Thapa,
Surendrabikram@vt.edu

Academic editor
Varun Gupta

Additional Information and Declarations can be found on page 22

DOI **10.7717/peerj-cs.3253**

How to cite this article Poudel S, Rauniyar K, Acharya A, Rashid J, Adhikari S, Naseem U, Thapa S. 2025. NepAES: exploring the promise of automated essay scoring for Nepali essays. **PeerJ Comput. Sci. 11:e3253 DOI 10.7717/peerj-cs.3253**

Subjects Computational Linguistics, Data Science, Emerging Technologies, Natural Language and Speech, Text Mining
Keywords Natural language processing, Automated essay scoring, Low-resource language

INTRODUCTION

Essays are primarily important in educational systems worldwide as they provide a comprehensive means for assessing students' understanding, critical thinking, and writing skills (*Norton, 1990*). Unlike multiple-choice tests, essays require students to articulate their thoughts, construct arguments, and demonstrate a deep understanding of the subject matter. This makes essays a valuable tool for educators to evaluate higher-order cognitive skills and the ability to communicate effectively. However, scoring essays is inherently challenging due to their subjective nature (*Gierl et al., 2014*; *Hussein, Hassan & Nassef, 2019*). Human evaluators may have different interpretations of the same piece of writing, leading to inconsistencies in scoring. Furthermore, human evaluators can be influenced by their own biases, consciously or unconsciously, which can affect the fairness of the scoring. Additionally, human evaluators may experience fatigue or lose concentration over time, impacting the accuracy and consistency of their scoring. Varying standards and expectations among evaluators can also contribute to discrepancies in scores.

The increasing use of artificial intelligence (AI) in educational technologies has transformed many aspects of teaching and learning (*Brusilovsky, 2024*; *Onesi-Ozigagun et al., 2024*; *Guilherme, 2019*; *Holmes & Tuomi, 2022*). AI tools are increasingly being integrated into educational use cases to enhance learning experiences and outcomes. For instance, AI-driven platforms can provide personalized learning paths, adapt to the unique needs of each student, and offer real-time feedback (*Kamalov, Santandreu Calonge & Gurrib, 2023*; *Huang, Lu & Yang, 2023*). In language learning, AI tools can help students practice pronunciation, grammar, and vocabulary with instant corrections and suggestions (*Gayed et al., 2022*; *Liang et al., 2023*). Furthermore, AI-based tutoring systems can assist students in subjects like mathematics and science by offering interactive problem-solving sessions and tailored explanations (*Hwang & Tu, 2021*; *Shin & Shim, 2021*). Similarly, in response to the challenges of manual evaluation of essays, automated essay scoring (AES) is being explored as an alternative tool to existing human evaluators to provide a more consistent and objective method for essay evaluations (*Lagakis & Demetriadis, 2021*; *Ramesh & Sanampudi, 2022*). AES utilizes advanced natural language processing (NLP) techniques and machine learning (ML) algorithms to assess the quality of writing based on various linguistic and structural features (*Hussein, Hassan & Nassef, 2019*). The primary goal of AES is to achieve performance levels that are equivalent to those of human evaluators, ensuring reliable and fair assessments. Notably, standardized tests such as the Test of English as a Foreign Language (TOEFL) and the Graduate Record Examinations (GRE) have begun incorporating AES to provide quicker and more consistent scoring of written responses (*Weigle, 2013*; *Ramesh & Sanampudi, 2022*). This shows the growing trust and reliance on these systems which calls for the need for ongoing research to enhance their accuracy and effectiveness across different languages and contexts.

Following the COVID-19 pandemic, there has been a significant shift towards digitizing educational content, moving away from traditional article-based assessments (*Maity, Sahu & Sen, 2021*). The digitization of educational content offers several advantages, including easier storage, retrieval, and analysis of student performance data (*Rodríguez & Pulido-Montes, 2022*). This trend was not only seen in developed countries but the developing countries as well (*Shrestha et al., 2022*; *Acharya & Rana, 2024*). This transition has also created new opportunities for the implementation of AES, as digital platforms can be easily integrated with AI tools for AES. Additionally, the ability to provide instant feedback through AES can significantly enhance the learning experience by allowing students to quickly understand their strengths and areas for improvement.

While these tools are widely used in English language contexts, there are numerous other regional languages that also need to be addressed. It is necessary for researchers to adhere to the United Nation's principle of "leave no one behind" (LNOB) (*Cordery, Arora & Manochin, 2023*) in the digital transformation of education. The Nepali language, spoken by approximately 30 million people, is one such example (*Rauniyar et al., 2023*). Nepali is part of the national curriculum in Nepal and is taught compulsorily up to high school. Several other countries like India also have the Nepali language as a part of their curriculum. Despite its significant number of speakers and its importance in the national education system, there are very few resources available for NLP in Nepali.

While some efforts are being made to create NLP datasets and resources in areas such as healthcare (*Adhikari et al., 2022*; *Thapa et al., 2020*), sentiment analysis (*Gupta & Bal, 2015*; *Piryani et al., 2020*; *Singh et al., 2020*; *Shahi, Sitaula & Paudel, 2022*; *Sitaula & Shahi, 2024*), POS tagging (*Pradhan & Yajnik, 2021*; *Paul, Purkayastha & Sarkar, 2015*), named entity recognition (*Maharjan, Bal & Regmi, 2019*), image captioning (*Adhikari & Ghimire, 2019*), *etc*., there is a notable gap in the field involving educational applications of AI. To fill this gap, we have generated a large translated dataset using machine translation algorithms and evaluated the efficiency of various ML and deep learning (DL) models in Nepali AES using scores like the Quadratic Weighted Kappa (QWK) score. Our main contributions are:

- We create a substantial dataset for Nepali AES by utilizing machine translation algorithms addressing the lack of resources in this field.
- We evaluate the efficiency of various ML and DL models in the context of Nepali AES, using metrics such as the QWK score.
- We analyze the impact of translation quality on the performance of AES models, emphasizing the importance of accurate translations for effective AES in low-resource languages.
- We demonstrate that fine-tuned transformer-based models can perform comparably to traditional feature-based ML methods, even in a low-resource setting like Nepali.

Through this work, we contribute to the ongoing efforts to make AI accessible to all. Our research aims to enhance the resources available for NLP in Nepali, ensuring that advanced educational technologies like AES are not limited to high-resource languages.

This effort aligns with the broader goal of making AI tools inclusive and beneficial for a diverse range of linguistic communities. The remainder of this article is organized as follows: In the 'Related Works', we review the related work in AES, along with some background on the state of NLP in the Nepali language and the need for AES in Nepali. In the 'Dataset/Corpora', we detail our dataset and the curation process. In the 'Methodology', we describe our methodology, including feature extraction and algorithms. In the 'Experimental Setup', we outline our experimental settings and evaluation metrics. The 'Results and Analysis' presents the results and analysis. Finally, the last section provides the conclusion and suggestions for future research directions.

RELATED WORKS

Using cutting-edge algorithms and ML approaches, the area of NLP has developed AES with notable progress. These systems are rapidly being implemented in educational settings because they can give faster responses and objective assessments of student work. Recent breakthroughs in AI have led to the creation of Essay Grading Systems that employ technical and semantic elements to enhance essay grading. These properties include referential coherence, lexical variety, syntactic complexity, and topic overlap estimate (*Ikram & Castle, 2020*). Additionally, a rank-based strategy that combines the agreement between human and machine raters has been developed, leading to enhanced AES performance (*Chen & He, 2013*). Below, we further discuss some of the recent advancements in AES.

Recent advancements in automated essay scoring

AES has enhanced classroom education and student's writing abilities by offering an effective and impartial essay grading system. Researchers have created different systems, such as e-raterTM, Intellimetric, and the Intelligent Essay Assessor, and examined psychometric challenges and improvements in the area (*Shermis & Burstein, 2003*). Recently, substantial research efforts have been focused on building automated essay assessment systems. These techniques have been effectively utilized in tests like TOEFL, and GRE, where essays are scored by both humans and automated algorithms (*Beseiso, 2021*).

Significant advancements in the field of AES have been realized through the creative application of ML techniques. *Ke & Ng (2019)* provided an overview of the ranking approach and analyzed several ways for evaluating coherence in learner texts using the Automated Assessment (AA) framework and the Incremental Semantic Analysis (ISA) model adapted for semantic coherence, highlighting feature sets that include length-based, category-based, syntactic, semantic, and discourse. They used a dataset of various student responses from the Cambridge student *Corpus*'s First Certificate in English (FCE) exam. *Vajjala (2018)* presented the role of several linguistic variables in AES using two publicly accessible datasets, the TOEFLSUBSET and the FCE, of non-native English essays produced in test-taking circumstances. Among the six extensive feature sets they offered for evaluating student language were word level, part of speech, grammatical features, discourse attributes, errors, and others (prompt and L1).

Similarly, there has been some notable research carried out on AES in the Arabic language. *Lotfy et al. (2023)* proposed an AES system in the Arabic language that utilizes a dataset from a sociology course with 270 essays and comprises two major components: a grading engine and an adaptive fusion engine. The grading engine looks at both string-based and *corpus*-based criteria to see how similar student answers are to model answers in different situations. The adaptive fusion engine then combines these scores using six ML algorithms and feature selection techniques to make the system more accurate and less likely to make mistakes. Similarly, *Reafat et al. (2012)* suggested an approach employing latent semantic analysis (LSA) and cosine similarity to grade Arabic essays, with an experiment with 29 student submissions. This research focuses on reducing the use of stopwords to achieve an acceptable score level. *Ramalingam et al. (2018)* presented an AES using various ML techniques. With 8,900 essays divided into 8 sets extracted from Kaggle, they used linear regression as their main technique for training their model. In addition to linear regression, they used other classification and clustering algorithms to improve the system's performance.

Advancing beyond traditional ML techniques, the field of AES has seen significant contributions from DL models. *Lu & Cutumisu (2021)* evaluated and compared three algorithms, namely convolutional neural networks (CNNs), CNN+long short-term memory (LSTM), and CNN+bidirectional long short-term memory (Bi-LSTM), to evaluate their performances on AES tasks using QWK as the evaluation metric within the same context. 12,979 essays were released from a Kaggle challenge named Automated Student Assessment Prize (ASAP), which has eight prompts and four genres namely narrative, persuasive, experiential, and source-dependent responses. Similarly, *Rodriguez, Jafari & Ormerod (2019)* provided a detailed exploration of the network architectures of BERT and XLNet, using the Kaggle AES dataset to benchmark their models, comparing the results with traditional methods such as bag of words and LSTM networks. Furthermore, *Wang et al. (2022)* presented a novel approach to AES using BERT, focusing on multi-scale essay representation using the ASAP dataset. *Li & Liu (2024)* worked on the feasibility of using LLMs for AES and the role of prompt engineering in developing the LLM-based AES. Similarly, *Atkinson & Palma (2025)* introduced a hybrid approach that combines linguistic features (lexical, readability, and grammatical diversity) and context embedding, which uses LLM models. *Song et al. (2024)* tested the ability of open-source LLMs to score the essay and assist in improvement, which employed methods like few-shot learning. *Xiao et al. (2025)* suggested a dual-process framework that utilizes cognitive theory, with a Slow Module for in-depth explanations and a Fast Module for rapid scoring that is triggered by confidence thresholds. The system demonstrates that human-AI cooperation works best in low-confidence scenarios, with the united team surpassing its separate parts through favorable performance. *Stahl et al. (2024)* examined LLM prompting techniques for automatic essay evaluation and feedback production with Mistral-7B (*Jiang et al., 2023*). By testing several personas, task sequences, and instructional methods on the ASAP dataset, it was discovered that although LLMs attain competitive scores (QWK 0.53), autonomous feedback generation provides the most beneficial outcomes. The combination of joint scoring and feedback production

demonstrates minimal advantages, indicating that these educational tasks are more effectively managed independently. *Seßler et al. (2025)* compared five LLMs against 37 German teachers who graded student essays using ten different criteria. It was found that all the models overrated essays and struggled with content evaluation.

AES has not been substantially studied for low-resource languages, owing to a lack of large and varied datasets. Low-resource languages often have less data accessible, making it difficult to design trustworthy algorithms for these languages (*Wang et al., 2023*). AES systems depend significantly on big datasets of annotated essays to train algorithms capable of evaluating and scoring student work. For widely spoken languages like English or Chinese, there is an abundance of educational materials, research, and data, making it possible to construct solid AES systems. In contrast, low-resource languages often lack annotated instructional resources. Furthermore, the complexity and variety of grammar, syntax, and idioms in less common languages make it difficult to create accurate and reliable models (*Singh et al., 2023*).

Natural language processing in Nepali language

Despite significant progress in NLP, which has expanded the capabilities of data analysis and information extraction, allowing computers to accurately comprehend human language, the field is still relatively new in the context of the Nepali language. Nepali is a complicated and monographically rich language, making the NLP tasks tough (*Niraula, Dulal & Koirala, 2021*). Regardless of the obstacles, initiatives are being carried out to promote incorporating the Nepali language in research and development. Below, we highlight some of the research done in the Nepali language in the realm of NLP.

Koirala & Niraula (2021) explored 25 state-of-the-art Word Embeddings for the Nepali language utilizing Global Vectors for Word Representation (GloVe), Word2Vec, fastText, and Bidirectional Encoder Representations from Transformers (BERT). With 700,000 unique news articles crawled from Nepali online news media, 39,000 Nepali articles from Wikipedia, and around 1.2GB dataset from Open Super-large Crawled ALMAnaCH *corpus* (OSCAR), they presented well-established metrics to evaluate these embeddings. Similarly, *Timilsina, Gautam & Bhattarai (2022)* presented NepBERTa, a BERT-based model trained on a *corpus* of 0.8B words from 36 distinct Nepali news sites and evaluated the performance with named-entity recognition, content classification, POS tagging, and Categorical Pair Similarity. *Shahi & Pant (2018)* presented a Nepali news dataset that consisted of 4,964 documents crawled from different Nepali news portals with 20 different categories. Baselines for the classification tasks were presented through the utilization of various ML and DL models.

Moreover, notable studies in sentiment analysis for the Nepali language have been carried out, suggesting a growing interest in the analysis of digital communication within this unique linguistic context. *Gupta & Bal (2015)* outlined two essential methodologies for Nepali text sentiment analysis. The initial strategy was identifying emotional phrases in Nepali texts to establish the document's tone. The second technique employed annotated Nepali text data to create an ML-based text classifier for categorizing content. *Shrestha & Bal (2020)* presented named-entity recognition to identify political

personalities in texts, followed by anaphora resolution, and finally, these figures are matched to related viewpoints, offering insights into the public attitude toward political entities. With a sentiment *corpus* of 3,490 phrases from Nepali news media articles, they proposed an ML-based sentiment classifier. Similarly, *Singh et al. (2020)* presented a dataset for a multilingual BERT model for aspect term extraction and a BiLSTM model for sentiment classification using social media data in Nepali. With 3,068 comments retrieved from 37 YouTube videos from nine distinct channels, they categorized it into six aspects: general, profanity, violence, feedback, sarcasm, and out-of-scope, and four target entities: person, organization, location, and miscellaneous.

Research in the Nepali language also includes work on hate speech identification, a vital topic in today's digital era for maintaining a courteous and secure online community. *Rauniyar et al. (2023)* presented 4,445 manually annotated tweets with a multi-aspect annotation consisting of seven basic classes: relevance, sentiment analysis, satire, hate speech, direction, targets, and hope speech for Nepali election discourse. The benchmarks show possibilities for improved automatic speech identification in Nepali. Apart from this, there have been few studies done in the field of medical NLP. *Adhikari et al. (2022)* accomplished manual translations to produce a Nepali Alzheimer's disease (AD) dataset comprising transcripts from 168 Alzheimer's disease patients and 98 control normal (CN) individuals from Dementia Bank. A total of 499 transcripts made up the dataset; 255 of the transcripts belonged to AD patients and 244 to CN individuals. Using Nepali transcripts, they proposed an NLP-based framework for the early identification of AD patients.

While research in the Nepali language in the context of NLP has covered a wide range of topics, including healthcare, AES is still largely unexplored. Improving AES for the Nepali language makes it feasible to dramatically improve the educational experience, giving timely and fair assessments while lowering the effort on educators. This innovation is vital for keeping pace with global educational technology trends and satisfying the increasing demands of the learning community in Nepal. Table 1 provides a comparison of datasets across various languages, offering insights into the current state of the data.

Need of AES techniques for Nepali

With the rapid advancement of technology, everything is moving online, including education. Assessment in the educational system is critical in determining student performance. As teacher-to-student ratios rise, the manual assessment procedure becomes increasingly difficult. Furthermore, in a nation like Nepal, where urban, rural, and regional growth is imbalanced, the education system confronts even greater obstacles. There is also a significant imbalance between supply and demand for educational resources in Nepal. As a result, Nepal has become a major student exporter, with many youths choosing to further their studies overseas. This causes major quality and talent migration issues in the country. Introducing a computer based assessment system that automatically scores or marks student responses aids in improving the quality of education and addresses some of these challenges. AES also relieves instructors of a heavy task by automating the essay grading process. The use of AES is not confined to the classroom; it also aids in educational reform by encouraging the revision of educational processes and regulations to align with global

Table 1 Overview of other datasets in existing literature.

Authors	Method	Dataset	No. of essays	Language
Song et al. (2020)	Multi-stage pre-training	School students	121,515	Chinese
Xie et al. (2022)	Neural pairwise contrastive regression	ASAP	12,978	English
Mizumoto & Eguchi (2023)	GPT (ChatGPT)	TOEFL11	12,100	English
Hu, Yang & Yang (2022)	Transformer	VisEssay	13,000	English
Ouahrani & Bennouar (2020)	Correlated Occurrence Analogue to Lexical Semantic	Arabic Dataset for automatic short answer grading	2,133	Arabic
Gaheen, ElEraky & Ewees (2021)	eJaya-NN	Personal dataset	240	Arabic
Hirao et al. (2020)	Feature-based and neural-network	GoodWriting	More than 800	Japanese
Marinho, Anchiêta & Moura (2022)	Feature-based	Essay-BR	4,570	Portuguese

technology. The pace of global knowledge and technological innovation has intensified in the digital age. Science, technology, AI, and digitalization are all intended to play an essential part in societal and economic progress. A complete overhaul of education policy is required to address numerous issues in the educational system, which in turn helps boost the country's prosperity. Thus, AES essentially represents a major advancement in the country's educational path by being at the forefront of fusing technical innovation with educational fairness and reform.

AES is also an important tool for boosting NLP research in the Nepali language. Due to the morphological rich characteristics and complicated sentence structure (*Niraula, Dulal & Koirala, 2021*), NLP in the Nepali language is exceptionally demanding and requires much study (*Timilsina, Gautam & Bhattarai, 2022*). The progress in NLP in low-resource languages such as Nepali is often hindered by the scarcity of pre-training data, lack of resource consistency, and limited computational resources. The development of AES demands the establishment of advanced linguistic models and huge databases adapted to the complexities of the Nepali language. This, in turn, drives NLP innovation, which contributes to the larger aims of language preservation and technological advancement in computational linguistics withing the language. Thus, AES for Nepali is not merely an educational tool but also a bridge between real educational needs and the rising horizon of language technological research.

DATASET/CORPORA

The ASAP dataset comprises eight sets of essays, each generated from a single prompt (P). These essays vary in length from 150 to 550 words and are written by students across grade levels 7 to 10. The dataset includes both source-dependent and independent essays, all of which were hand-graded and double-scored to ensure reliability. This variety in the dataset is intended to challenge and evaluate the capabilities of AES engines. The essays were collected as part of a competition aimed at advancing the technology for automatically

scoring written responses. The following is a summary of the eight prompts included in the dataset.

- **P1:** People hold different views on the impact of computers on society. Some believe computers offer numerous benefits, while others worry they limit physical activity and personal contact. Write a letter to your local newspaper explaining your opinion on this subject.
- **P2:** Censorship in libraries raises concerns regarding access to content considered offensive. Write a persuasive essay, arguing whether such materials should be removed, and back your opinion with personal experiences and observations.
- **P3:** Writers read a passage from *ROUGH ROAD AHEAD: Do Not Exceed Posted Speed Limit* by Joe Kurmaskie and discussed how the setting affected the cyclist's experience.
- **P4:** The writer had to read a part of *Winter Hibiscus* by Minfong Ho. Then, they needed to clarify why the writer ends the story like that.
- **P5:** The writers were tasked with reading a paragraph from *Narciso Rodriguez* by Narciso Rodriguez. They were then asked to explain the feeling/mood expressed by the author in the text, using relevant information from the paragraph to back up their description.
- **P6:** The task involves describing the challenges experienced by the Empire State Building constructors in accommodating dirigible docking. Writers must use details from the article *The Mooring Mast* by Marcia Amidon Lüsted to support their explanation of those obstacles.
- **P7:** Write a story about a time when you, or someone you know, showed patience. Describe how they remained calm and tolerant, handling difficulties without complaining.
- **P8:** The task was to tell a true story where laughter played a role. Writers should show how laughter was an important part of the relationship in the story.

The ASAP dataset consists of three types of essays: Argumentative, Source Dependent, and Narrative essay.

- **Argumentative essays:** These essays require the writer to convince the reader of their viewpoint on a specific subject by presenting evidence and logical arguments to support their view, whether they are in favor or against the topic. P1 and P2 fall under this category.
- **Source Dependent essays:** These essays require the writer to respond to a query regarding a source text, typically by expressing their viewpoint or analysis of an event or situation presented in the text. P3-P6 were included in this category.
- **Narrative essays:** These essays consist of themes that require the writer to narrate a story or provide a comprehensive depiction of an event, experience, or subject. Finally, P7 and P8 were categorized under narrative essay.

Table 2 Summary of ASAP dataset.

Prompt (P)	No. of essay	Score range
P1	1,783	2–12
P2	1,800	1–6
P3	1,726	0–3
P4	1,772	0–3
P5	1,805	0–4
P6	1,800	0–4
P7	1,569	0–30
P8	723	0–60

The ASAP dataset, hosted on Kaggle (https://www.kaggle.com/competitions/asap-aes/data), provides an opportunity for researchers to test and demonstrate their scoring algorithms' effectiveness against a benchmark of human-graded essays. Table 2 represents a brief description of the ASAP dataset.

NepAES dataset

The ASAP *corpus* has been used as a benchmark for evaluating different English language models (*Ke & Ng, 2019*). We chose to analyze a version of the ASAP dataset translated into Nepali since it is widely utilized in AES research. This dataset provides a significant amount of diverse data, including different types of prompts like narrative, argumentative, and source-dependent responses (*Mathias & Bhattacharyya, 2018*). Translation has played a crucial role in the field of NLP, particularly in the creation of multilingual datasets and architectures. Translation plays an integral part in enhancing the possibilities of language-related research and technology. We utilized translation models such as mBART-50 (*Tang et al., 2020*) and Google Translate to produce the NepAES datasets.

We evaluated different subsets of the translated collection using a stratified approach to ensure coverage across the full range of essay types and quality levels. We selected essays from all eight prompts included in the dataset and sampled three distinct sets for each prompt representing high-scoring (best), mid-range (average), and low-scoring (worst) essays based on the original ASAP scores. These 24 stratified samples (3 per prompt $\times$ 8 prompts) were then evaluated for translation accuracy and contextual fidelity by a group of skilled bilingual individuals and professional academics. The evaluators included two university-level educators and one graduate-level language expert, all fluent in both English and Nepali. The reviewers assessed the quality of the translated content using a rubric focused on semantic preservation, syntactic correctness, and overall coherence. Their feedback indicated that the translations generally preserved the original context and conveyed meaning appropriately, although minor issues were noted in some low-scoring essays due to structural ambiguities in the source text. The quality of the content was hence found to be appropriate and to articulate the context well.

Neural machine translation approaches

Neural machine translation (NMT) is a noteworthy breakthrough in machine translation that uses deep learning neural networks to enhance the fluency and accuracy of language translation, surpassing prior methods (*Bahdanau, Cho & Bengio, 2014*). During the early 2010s, there was a significant change in translation methods that focused on end-to-end learning. In contrast to previous methods that divide the translation process into distinct stages, this modern approach of NMT considers the entire translation work as a single task. Consequently, the system no longer divides the process into separate components but rather treats translation as one cohesive task (*Vaswani et al., 2017*). NMT generally employs sophisticated neural network structures such as recurrent neural networks (RNNs), long short-term memory (LSTM) networks, and more recently, the Transformer model. These models have shown remarkable skill in analyzing sequences and grasping context (*Wu et al., 2016*; *Vaswani et al., 2017*). An outstanding characteristic of NMT is its context-awareness, offering translations that takes into account the entire sentence or phrase. This allows for translations that process the full sentence or phrase, resulting in an improved natural flow and consistency of the translated text. Algorithm 1 illustrates the fundamental process of the NMT model. In our research, we employed two translation models: Google Translate and mBART Translate.

Google translation

Google Translate employs a sophisticated algorithm known as the Google Neural Machine Translation system (GNMT). The GNMT works great in machine translation tasks, tackling issues such as translation quality, speed, and robustness. The system utilizes a complex LSTM network with eight encoder and eight decoder layers, incorporating residual connections and an attention mechanism. The primary feature is the bidirectional encoder in the initial layer, which captures contextual information from both directions of the input sentence, that enhances the comprehension of the text. The model utilizes parallelism to enhance computing efficiency by dividing the network across multiple GPUs. GNMT utilizes a Wordpiece model for segmentation, which divides words into sub-words, and also explores a Mixed Word/Character Model. Both methods efficiently address vocabulary size and the issue of rare words. In the GNMT system's decoder, beam search is utilized to identify the sequence Y that maximizes a score function $s(Y, X)$, given a trained model. This process includes two key refinements: a coverage penalty and length normalization. Length normalization addresses the challenge of comparing hypotheses of different lengths, as regular beam search tends to favor shorter results due to the accumulation of negative log probabilities at each step. This bias is mitigated by dividing the score by $(5 + |Y|)^{\alpha}$, where $|Y|$ is the length of the hypothesis and α is a parameter optimized on a development set, typically found to be between 0.6 and 0.7 (*Wu et al., 2016*). The coverage penalty $cp(X; Y)$ is included to encourage translations that fully cover the source sentence, as indicated by the attention module. The scoring function is defined in Eqs. (1), (2), and (3):

Algorithm 1 **Generalized neural machine translation.**

Input: Source text X
Output: Translated text Y
function TRANSLATE (X)
    $X_{\text{tokenized}} \leftarrow$TOKENIZE(X) ▷ Tokenization of the source texts
    $H \leftarrow$ENCODE$X_{\text{tokenized}}$ ▷ Encoding the tokenized text into hidden states
    $Y_{\text{init}} \leftarrow$ [start_token] ▷ Initializing the target sequence with a start token
    while end_token not in Y_{init} **do**
        $Y_{\text{init}} \leftarrow$DECODE($Y_{\text{init}}$, H) ▷ Decoding the sequence iteratively
    end while
    $Y \leftarrow$DETOKENIZE(Y_{init}) ▷ Detokenizing the output sequence
        return Y
end function
function TRAIN TRANSFORMER
    Initialize model parameters θ
    for each batch in training data **do**
        Compute loss $\mathscr{L}(\theta)$ using cross-entropy
        Update θ using backpropagation and optimizer
    end for
end function

$$s(Y,X) = \frac{\log(P(Y|X))}{lp(Y)} + cp(X;Y) \tag{1}$$

$$lp(Y) = \frac{(5+|Y|)^{\alpha}}{(5+1)^{\alpha}} \tag{2}$$

$$cp(X;Y) = \beta \sum_{i=1}^{|X|} \sum_{j=1}^{|Y|} \log(\min(p_{ij}, 1.0)) \tag{3}$$

Here, $P(Y|X)$ is the probability of the hypothesis Y given the source sentence X, $lp(Y)$ is the length penalty, and $cp(X;Y)$ is the coverage penalty. The parameter β controls the strength of the coverage penalty, and p_{ij} represents the attention probability of the j-th target word on the i-th source word.

mBART-50 translation

mBART-50 model is notable for its ability to handle 50 languages, doubling the capacity of the original mBART (*Liu et al., 2020*). A key technical aspect is the use of a denoising autoencoder approach for pretraining, leveraging large amounts of monolingual data, especially beneficial for low-resource languages. The model is then fine-tuned on parallel text for translation tasks (*Tang et al., 2020*). The core of the model's training objective is summarized in the Eq. (4):

$$L(\theta) = \sum_{D_i \in D} \sum_{x \in D_i} \log P(x|g(x);\theta) \tag{4}$$

where,

D represents the training data, D_i denotes the data in language i, x is a text instance, $g(x)$ is a noise function applied to x, and θ are the model parameters. The noise function g

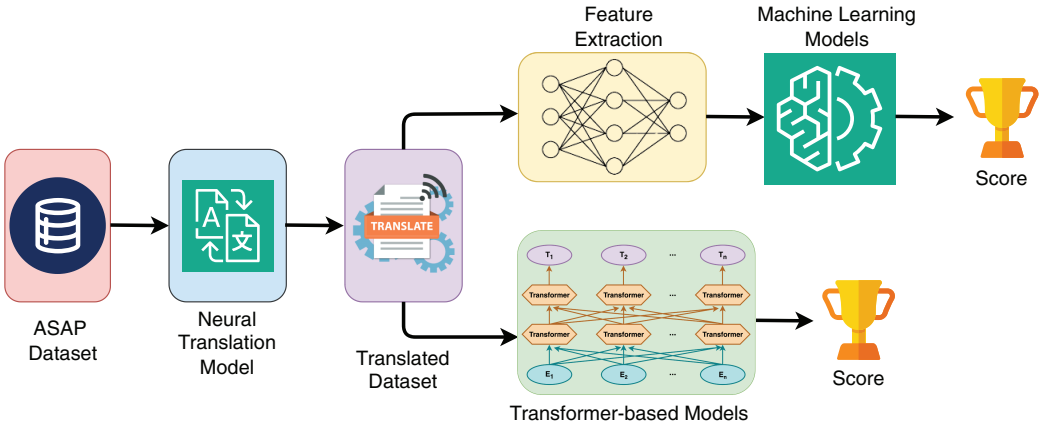

Figure 1 Workflow of automated essay scoring system. Full-size DOI: 10.7717/peerj-cs.3253/fig-1

includes random span masking and order permutation, and P is the probability distribution defined by the model.

METHODOLOGY

In this section, we describe the process of extracting features from various sources, including essay quality dimensions. We also provide information on the baseline models used for comparison. The Fig. 1 provides a comprehensive illustration of the overall workflow of the AES system, detailing each step from the initial submission of an essay to the final assignment of a score.

Manual feature extraction

Our feature extraction draws concepts from various sources, incorporating elements from the essay quality dimensions proposed by *Ke & Ng (2019)*, the readability metrics outlined by *Sinha et al. (2012)*, and the attributes described in the *Mathias & Bhattacharyya (2018)*. This comprehensive technique enables us to examine multiple aspects when analyzing our system's performance. We deliberately excluded spelling-related parameters from our selection method due to neural translation algorithms' tendency to automatically correct spelling problems. We systematically extracted and employed six distinct features:

- *Unique words count:* This analyzes and quantifies the occurrence of infrequently used words with a frequency of 1 in a given text after pre-processing and filtering out common stop words and punctuation.
- *Overlap score:* The grade of an essay is determined by how well its ideas are linked and how smoothly they flow together. To learn this, we focus on computing semantic similarity scores using the MuRIL model (*Khanuja et al., 2021*) and cosine similarity. We measured the semantic similarity between two sentences, considering them at an interval of four sentences. This allows us to grasp how well the essay maintains coherence and semantic overlap. Ultimately, we calculated the average scores by comparing pairs of sentences.

- *Essay length:* The total number of words in the essay determines its length.
- *Average sentence length:* The calculation of the average sentence length involves the summation of the individual lengths of all sentences within the text, followed by dividing this sum by the total number of sentences present.
- *Average word length:* The determination of the average word length entails the aggregation of the individual lengths of all words within the given text, and subsequently dividing this cumulative length by the total number of words present.
- *Readability:* In the work by *Sinha et al. (2012)*, a readability metric for Hindi and Bangla is introduced, wherein various structural features are studied. Both Hindi and Nepali utilize the Devanagari script in their written expressions. For the assessment of readability, we adopted the linguistic concepts and methodologies inherent in the Hindi language, considering the shared script and linguistic similarities between the two languages. These features encompass parameters such as average sentence length (ASL), average word length (AWL), number of polysyllabic words (PSW), number of jukta-akshars (JUK), among others. Spearman's rank correlation coefficient was employed to analyze the relationships among these structural features. Our readability scores are calculated using a formula based on the results of their regression analysis, which includes the observed structural features.

$$(0.01 * PSW) + (2.14 * AWL) - (2.34). \tag{5}$$

Machine learning baselines

In this section, we present a wide range of baseline models meticulously selected for comparison.

Naive Bayes

Naive Bayes (NB) analyzes the probabilities of each potential value within the targeted range, resulting in a thorough probability distribution for the target variable. The resulting distribution represents the probability of each possible outcome (*Frank et al., 2000*). The model calculates the probability of each score category given the essay features, using Bayes' theorem. The category with the highest probability is assigned as the predicted score (*Rudner & Liang, 2002*).

Support vector regression

The support vector regression (SVR) model predicts the score of an essay by mapping it to the hyperplane that has been trained within the feature space. The hyperplane represents the correlation between the features and the target scores that the model has acquired through training. The predicted score is calculated by finding the value on the hyperplane that is closest to the essay's feature vector. This ensures that the prediction aligns closely with the patterns identified in the training data (*Li & Yan, 2012*).

Linear regression

Linear regression (LR) is a relatively simple and efficient model for AES. The model is provided with a set of essays, each accompanied by its corresponding scores, enabling it to

learn the coefficients of the linear equation that optimally fits the data. Features used can include simple statistics like word count, vocabulary diversity, grammar errors, *etc*. as well as more complex NLP-based features (*Song & Zhao, 2013*).

Decision tree

In the decision tree (DT), models are represented as tree structures, where nodes represent features and edges indicate the possible values of those features (*Suthaharan & Suthaharan, 2016*). At each internal node, a decision is made based on the value of a specific feature, while the leaf nodes represent the predicted essay scores. This structure allows DTs to capture complex, non-linear relationships between the features and scores effectively (*De Ville, 2013*).

Random forest

Random forest (RF) is an ensemble learning technique that uses multiple decision trees to perform regression tasks. Each individual decision tree in the forest makes a prediction, and the final output is the average of all the tree predictions. Bagging and feature randomness are used to introduce randomness. The hyperparameters for a RF regressor include the number of trees, the maximum tree depth, and the number of features used for each split (*Cutler, Cutler & Stevens, 2012*).

XGBoost

Extreme gradient boosting (XGB) is a robust ML algorithm that uses the gradient boosting technique to optimize model performance and speed. It is particularly popular due to its ability to handle large datasets efficiently and accurately. The main features of XGB are parallelization, tree pruning, regularization, and sparsity awareness. These features make it a versatile tool that is widely used for various tasks, including classification, regression, and ranking problems (*Chen & Guestrin, 2016*).

AdaBoost (ADA)

Adaptive Boosting, or AdaBoost (ADA) (*Schapire, 2013*), is a popular ensemble learning algorithm that operates by training a series of weak learners. ADA can be employed in regression tasks to improve predictive performance by iteratively adjusting the weights of training samples according to the errors of prior models.

Transformers-based models

Transformer-based models (TBMs) use self-attention to capture relationships between distant words in a text effectively (*Vaswani et al., 2017*). Bidirectional Encoder Representations from Transformers (BERT) utilizes a bidirectional context, enabling it to analyze the complete input sequence during training, unlike standard language models that handle text in a unidirectional manner (*Devlin et al., 2018*). BERT's bidirectional attention mechanism allows it to understand complex contextual details and relationships in language, leading to better performance on a range of NLP tasks like question answering, sentiment analysis, and named entity recognition. The model achieves this through unsupervised pre-training on vast amounts of text data, followed by fine-tuning on specific downstream tasks (*Devlin et al., 2018*). BERT's architecture comprises multiple

transformer layers, incorporating self-attention mechanisms to efficiently capture long-range dependencies. Its success has resulted in extensive use and has been the foundation for advanced models in the field of NLP.

DistilBERT (Nepali)

DistilBERT (*Sanh et al., 2019*) presents a technique for compressing BERT-based language models. DistilBERT uses knowledge distillation to train a smaller model to mimic the behavior of the original BERT model. DistilBERT's decrease in model size by about 40% and fewer parameters promotes quicker inference and lowers computational costs. We used a special Nepali DistilBERT (*Shrestha, 2021*) model for our testing, and it can be found in the Hugging Face library. This model was pre-trained using the OSCAR Nepali (*Suárez, Sagot & Romary, 2019*) collection of texts.

RoBERTa (Nepali)

RoBERTa (*Liu et al., 2019*) introduces a better pretraining technique for BERT models. It utilizes larger batch sizes, dynamic masking patterns, and longer training periods, resulting in increased language representation capabilities. For our benchmark study, we used a special model called *roberta-base-ne* (*Chaudhary, 2021*). This model was trained from a huge collection of Nepali sentences called the 'Nepali CC-100' dataset (*Conneau et al., 2019*; *Wenzek et al., 2019*), which comprises about 12 million sentences.

NepBERTa

NepBERTa which is a BERT-based model specifically designed for the Nepali language. This model is unique in its training on an extensive *corpus* of 0.8 billion words, sourced from a variety of popular news sites in Nepal. NepBERTa shows its proficiency across several NLP tasks, including named-entity recognition and content classification, and establishes the first Nepali Language Understanding Evaluation benchmark (Nep-gLUE) (*Timilsina, Gautam & Bhattarai, 2022*).

NepaliBERT

NepaliBERT model (*Ghimire, 2022*) finds its usage in various NLP tasks related to the Devanagari language and at the time of its training, it was considered a state-of-the-art model for the Devanagari dataset. The model has been carefully trained on a large dataset consisting of 6.7 million lines of unprocessed Nepali texts. The extensive training dataset was carefully created by combining a large Nepali *corpus* (*Lamsal, 2020*) with the OSCAR Nepali *corpus* (*Suárez, Sagot & Romary, 2019*). The combination of these two datasets enhances the strength and linguistic flexibility of NepaliBERT, establishing it as a significant asset for a wide range of NLP tasks and research projects focused on the Nepali language.

NepNewsBERT

NepNewsBERT (*Pudasaini, 2021*) stands as an advanced masked language model (MLM) created specifically to address the complex structure of the Nepali language. The advanced model has been extensively trained using a well-curated dataset from well-known Nepali

Table 3 Experimental settings for the transformer-based models. All models were trained with the same batch sizes, epochs, and learning rates.

Models	Batch size	Epoch	Learning rate	Maximum sequence length
DistillBERT (Nep)	8	10	2e−5	128
RoBERTa (Nep)	8	10	2e−5	512
NepBERTa	8	10	2e−5	512
NepaliBERT	8	10	2e−5	512
NepNewsBERT	8	10	2e−5	512

news websites. The training dataset consists of almost 10 million sentences in Nepali, reflecting diverse linguistic styles and contexts found in news articles.

All four BERT-based models except DistilBERT (Nepali) were only available as fill masks. Since we needed the models for regression tasks, we modified each model for downstream regression and used them accordingly.

EXPERIMENTAL SETUP

This section provides details about how the experiments were set up and conducted. In 'Experimental settings', it explains how we improved the reliability of our models by normalizing custom feature scores and preparing the text data for analysis. In 'Evaluation metrics', we talk about using QWK, which is a well-known measure, to evaluate and compare different AES methods.

Experimental settings

To enhance the reliability of the model, the custom feature scores were subjected to a process of normalization. In the context of feature-based approaches, the text underwent pre-processing, involving the removal of stopwords, named-entities, and mentions (indicated by '@' symbols). This pre-processing step aims to refine the input data for improved performance. We employed the AdamW optimizer (*Loshchilov & Hutter, 2017*) to fine-tune pre-trained transformer models. The model parameters for all transformer-based models are illustrated in Table 3. This involved carefully tuning these parameters to optimize the performance and reaching an optimal state of each model during the training process. The selection of appropriate learning rates and epochs is crucial for achieving optimal results in the training phase. Our normalization procedure guarantees standardized score ranges between 0 and 1. Normalized scores are converted back to the original prompt-specific scale for prediction when calculating QWK scores. In our experimental setup, we employ a split strategy of 70-15-15 for the training, validation, and test datasets across all prompts.

Evaluation metrics

In our research, we used the QWK, a widely recognized metric in the field, to assess and compare AES methods (*Cohen, 1968*). In QWK, it accounts for the agreement between raters while considering the possibility of agreement occurring by chance. QWK is particularly effective in situations where raters are ranking items on an ordinal scale, like

grading essays. QWK score typically ranges from −1 to 1. A score of 1 indicates perfect agreement between raters, while a score of −1 signifies perfect disagreement. A score of 0 would indicate that the agreement is no better than chance.

$$w_{ij} = \frac{(i-j)^2}{(N-1)^2}. \tag{6}$$

Here, i and j are the ratings assigned by the two raters, and N is the number of possible ratings. w_{ij} means that the weight increases with the square of the distance between the two ratings. This quadratic weighting penalizes larger disagreements more heavily than smaller ones, reflecting the intuition that a larger difference in ratings indicates a more significant disagreement. QWK score is calculated according to Eq. (7):

$$\text{QWK} = 1 - \frac{\sum_{i,j} w_{ij} o_{ij}}{\sum_{i,j} w_{ij} e_{ij}} \tag{7}$$

where:

- w_{ij} is the weight assigned to the disagreement between raters for the i^{th} and j^{th} items.
- o_{ij} is the observed frequency of ratings in which one rater rates an item as i and the other rates it as j.
- e_{ij} is the expected frequency of ratings for the i^{th} and j^{th} items, based on the assumption that ratings are given randomly.

RESULTS AND ANALYSIS

In this section, we discuss the results for both the NepAES datasets. For ML algorithms, we employed a feature-based approach, incorporating six distinct features for our experiments. For NepAES *corpus* (Google translation), we utilized a total of seven ML algorithms, with NB and XGB both achieving a QWK score of 0.658 for prompt P2. RF attained a QWK score of 0.647, while SVR achieved the highest QWK score of 0.784 for prompts P3, and P8, respectively. Similarly, the transformer-based model exhibited better performance for the remaining prompts, achieving QWK scores of 0.796, 0.774, 0.812, 0.762, and 0.774 for the prompts P1, P4, P5, P6, and P7, respectively. Table 4 illustrates the performance of all the models. The comprehensive results reveal that NB, an ML model, attains the highest average score of 0.712.

We also employed the mBART-50 NMT model for generating another NepAES dataset. Similar to the NepAES dataset created using Google translation, we employed the same strategy for the ML model applied to the NepAES *corpus* (mBART-50 translation). For prompts P1, P2, and P5, the ML models outperformed the BERT-based model, achieving QWK scores of 0.784, 0.693, and 0.808, respectively. Similarly, for the remaining prompts P3, P4, P6, and P8, the BERT-based model produced promising results with QWK scores of 0.673, 0.736, 0.829, and 0.637, respectively. The NepaliBERT, a BERT-based model, exhibited the highest QWK average score across all prompts, as depicted in Table 5.

Table 4 Outcomes of experiments for all models are presented in terms of QWK on the NepAES *corpus* (Google translation). The most ideal performance for each prompt is highlighted with bold numbers.

Model family	Models	P1	P2	P3	P4	P5	P6	P7	P8
ML Models	NB	0.787	**0.658**	0.605	0.681	0.804	0.710	0.719	0.733
	SVR	0.775	0.647	0.644	0.645	0.810	0.659	0.722	**0.784**
	LR	0.764	0.596	0.596	0.605	0.791	0.670	0.713	0.781
	RF	0.760	0.625	**0.647**	0.623	0.791	0.698	0.662	0.736
	DT	0.660	0.590	0.453	0.620	0.673	0.549	0.618	0.674
	XGB	0.767	**0.658**	0.534	0.609	0.765	0.689	0.708	0.763
	ADA	0.754	0.600	0.496	0.528	0.787	0.559	0.668	0.759
TBMs	DistillBERT (Nep)	0.602	0.529	0.631	0.623	0.700	0.656	0.697	0.343
	RoBERTa (Nep)	0.611	0.433	0.408	0.758	0.783	**0.762**	0.686	0.282
	NepBERTa	0.742	0.568	0.538	**0.774**	**0.812**	0.685	**0.774**	0.558
	NepaliBERT	**0.796**	0.568	0.579	0.733	0.705	0.721	0.624	0.590
	NepNewsBERT	0.727	0.589	0.410	0.751	0.748	0.751	0.737	0.475

Table 5 Outcomes of experiments for all models are presented in terms of QWK on the NepAES *corpus* (mBART-50 translation). The most ideal performance for each prompt is highlighted with bold numbers.

Model family	Models	P1	P2	P3	P4	P5	P6	P7	P8
ML Models	NB	0.733	0.680	0.612	0.664	0.806	0.658	0.713	0.498
	SVR	**0.784**	0.668	0.637	0.618	**0.808**	0.674	0.734	0.568
	LR	0.781	0.595	0.553	0.601	0.797	0.680	0.705	0.554
	RF	0.736	**0.693**	0.588	0.642	0.793	0.663	0.685	0.499
	DT	0.674	0.518	0.527	0.558	0.667	0.558	0.480	0.438
	XGB	0.763	0.630	0.584	0.632	0.777	0.592	0.656	0.499
	ADA	0.759	0.616	0.416	0.537	0.776	0.588	0.628	0.536
TBMs	DistillBERT (Nep)	0.502	0.429	0.633	0.711	0.715	0.509	**0.754**	0.376
	RoBERTa (Nep)	0.424	0.439	0.615	0.728	0.738	0.742	0.745	0.350
	NepBERTa	0.717	0.601	0.609	0.697	0.746	**0.829**	0.693	0.555
	NepaliBERT	0.770	0.636	0.613	**0.736**	0.697	0.775	0.745	0.577
	NepNewsBERT	0.685	0.543	**0.673**	0.602	0.777	0.691	0.708	**0.637**

Analysis

It can be observed that a few ML-based models gave promising results for both datasets due to their feature-based approach. Feature-based techniques depend on distinct features or attributes of the data to generate accurate predictions. This fundamental focus on discrete features enables them to consistently perform well in tasks because they are highly tailored to the input data. Nevertheless, feature-based techniques have a drawback in that they have a restricted capacity to understand the complex nuances and context of language. They often struggle to interpret the content accurately, which can hinder their performance in tasks that require a deeper understanding of language. On the other hand,

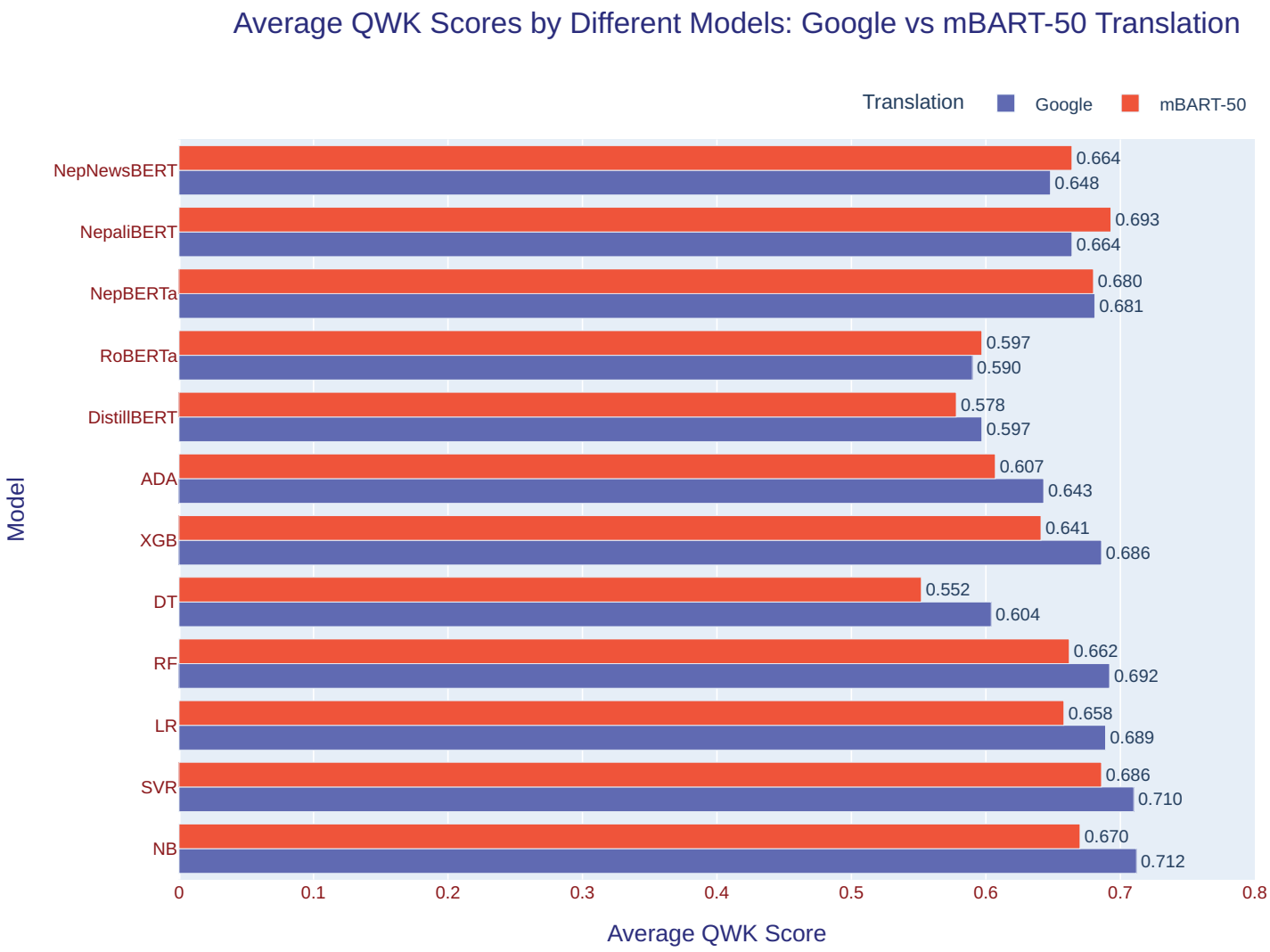

Figure 2 **The bar chart compares the average performance values of various machine learning models and transformer-based models for the NepAES *corpus* for Google and mBART-50 translation.** Full-size DOI: 10.7717/peerj-cs.3253/fig-2

large pre-trained language models like BERT have undergone training using extensive text data and possess the ability to understand complex language patterns and context. As a result, they can perform exceptionally well in tasks that require nuanced interpretation of content.

In comparison to the performance on other prompts, P8 did not meet expectations for the BERT-based model. This could be attributed to the size of the dataset, as prompt P8 consisted of only 723 essays. Consequently, the model was trained on a relatively small amount of data. RoBERTa (Nep) exhibited poorer performance compared to other BERT-based models on both datasets. The performance of various models differed depending on the method used to translate the dataset. The quality of translation plays a significant role in determining the effectiveness of the AES systems. In summary, the accuracy and precision of the translation method directly influence the performance of the

AES system. Therefore, ensuring high-quality translations is crucial for achieving reliable results, especially in AES applications for low-resource languages. Figure 2 shows the average QWK score of various models.

Further considerations and limitations

In the NepAES study, the dataset was created by translating the widely used ASAP dataset into Nepali using translation models such as Google Translate and mBART-50. The accuracy of these translations played a critical role in the effectiveness of the scoring models, as translation errors could introduce artifacts that impact model performance. While we conducted a structured human evaluation where bilingual experts assessed fluency, coherence, and semantic fidelity to identify overt errors and inconsistencies, this process is inherently limited in detecting subtle linguistic nuances that may affect downstream scoring behavior. As a result, even though human evaluators judged the translations to be generally reliable, undetected translation artifacts may still have influenced model outcomes, highlighting the need for more robust, organically sourced datasets for low-resource languages like Nepali. The low availability of comprehensive and diverse linguistic datasets in Nepali inhibits the capacity to construct and fine-tune robust models for AES. Furthermore, the research reveals that different types of models (standard ML and advanced language models) operate differently depending on how the dataset is translated. This means it is vital to think about how the dataset is prepared and which model is chosen. Essays can be different types, like narrative or argumentative, and each type might demand a different technique for the best results in AES.

The study offers significant insights and demonstrates the feasibility of AES for Nepali essays. It also highlights several challenges and constraints that need to be addressed in future research. This includes the development of more advanced models, the creation of larger datasets specifically for Nepali, and exploring a combination of approaches to improve the accuracy and reliability of AES in the context of low-resource languages.

CONCLUSION

The research primarily aimed to investigate AES for Nepali essays, which is a novel attempt in the realm of educational technology. The study focused on addressing the challenges related to evaluating essays in real-time, particularly taking into account the linguistic complexities and the subjective aspect of essay scoring. An essential component of the research involved utilizing diverse ML and NLP models, such as BERT-based models, to carry out the scoring procedure. The NepAES *corpus*, a crucial component of this study, was created by translating the ASAP dataset into the Nepali language. This translation was conducted utilizing various models such as mBART-50 and Google Translate, with the aim of exploring the feasibility of AES in a low-resource language setting, specifically Nepali. The study showed that the efficiency of different models varied depending on the essay prompts and the translation methods employed.

The study not only offered useful insights into AES for Nepali essays but also indicated various issues and limitations that demand attention in future research. This includes the advancement of complex models, the generation of extensive and varied datasets

specifically for the Nepali language, and the investigation of a combination of methods to improve the precision and dependability of AES in situations when there are limited linguistic resources. The NepAES study represents an important step in the sector of educational technology and language processing, notably in the context of the Nepali language and low-resource languages in general.

ACKNOWLEDGEMENTS

We acknowledge the use of generative AI tools (ChatGPT) to support the editing and refinement of language in this manuscript. All conceptual, experimental, and analytical contributions were developed independently by the authors.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

The authors received no funding for this work.

Competing Interests

The authors declare that they have no competing interests.

Author Contributions

- Sweta Poudel conceived and designed the experiments, performed the experiments, analyzed the data, performed the computation work, prepared figures and/or tables, authored or reviewed drafts of the article, and approved the final draft.
- Kritesh Rauniyar conceived and designed the experiments, performed the experiments, analyzed the data, performed the computation work, prepared figures and/or tables, authored or reviewed drafts of the article, and approved the final draft.
- Ashish Acharya performed the experiments, analyzed the data, performed the computation work, authored or reviewed drafts of the article, and approved the final draft.
- Junaid Rashid conceived and designed the experiments, performed the experiments, authored or reviewed drafts of the article, and approved the final draft.
- Surabhi Adhikari conceived and designed the experiments, analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the article, and approved the final draft.
- Usman Naseem conceived and designed the experiments, performed the experiments, authored or reviewed drafts of the article, and approved the final draft.
- Surendrabikram Thapa conceived and designed the experiments, analyzed the data, performed the computation work, prepared figures and/or tables, authored or reviewed drafts of the article, and approved the final draft.

Data Availability

The following information was supplied regarding data availability:

The code and data are available in the Supplemental File.

The data is available at Zenodo: Thapa, S., Rauniyar, K., & Naseem, U. (2025). Dataset for article "NepAES: Exploring the promise of automated essay scoring for Nepali essays" [Data set]. Zenodo. https://doi.org/10.5281/zenodo.16760033.

Supplemental Information

Supplemental information for this article can be found online at http://dx.doi.org/10.7717/peerj-cs.3253#supplemental-information.

REFERENCES

Acharya BN, Rana K. 2024. How students and teachers voyaged from physical classroom to emergency remote teaching in COVID-19 crisis: a case of Nepal. *E-Learning and Digital Media* **21(2)**:125–140 DOI 10.1177/20427530231156166.

Adhikari A, Ghimire S. 2019. Nepali image captioning. In: *2019 Artificial Intelligence for Transforming Business and Society (AITB)*. Vol. 1, Piscataway: IEEE.

Adhikari S, Thapa S, Naseem U, Singh P, Huo H, Bharathy G, Prasad M. 2022. Exploiting linguistic information from Nepali transcripts for early detection of alzheimer's disease using natural language processing and machine learning techniques. *International Journal of Human-Computer Studies* **160(2)**:102761 DOI 10.1016/j.ijhcs.2021.102761.

Atkinson J, Palma D. 2025. An LLM-based hybrid approach for enhanced automated essay scoring. *Scientific Reports* **15(1)**:14551 DOI 10.1038/s41598-025-87862-3.

Bahdanau D, Cho K, Bengio Y. 2014. Neural machine translation by jointly learning to align and translate. ArXiv DOI 10.48550/arXiv.1409.0473.

Beseiso MH. 2021. Essay scoring tool by employing RoBERTa architecture. In: *International Conference on Data Science, E-learning and Information Systems 2021*, 54–57.

Brusilovsky P. 2024. AI in education, learner control, and human-AI collaboration. *International Journal of Artificial Intelligence in Education* **34(1)**:122–135 DOI 10.1007/s40593-023-00356-z.

Chaudhary A. 2021. RoBERTa(Nepali). *Available at https://huggingface.co/amitness/roberta-base-ne* (accessed 25 February 2024).

Chen T, Guestrin C. 2016. XGBoost: a scalable tree boosting system. In: *Proceedings of the 22nd ACM Sigkdd International Conference on Knowledge Discovery and Data Mining*, 785–794.

Chen H, He B. 2013. Automated essay scoring by maximizing human-machine agreement. In: *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 1741–1752.

Cohen J. 1968. Weighted kappa: nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin* **70(4)**:213–220 DOI 10.1037/h0026256.

Conneau A, Khandelwal K, Goyal N, Chaudhary V, Wenzek G, Guzmán F, Grave E, Ott M, Zettlemoyer L, Stoyanov V. 2019. Unsupervised cross-lingual representation learning at scale. ArXiv DOI 10.48550/arXiv.1911.02116.

Cordery C, Arora B, Manochin M. 2023. Public sector audit and the state's responsibility to "leave no-one behind": the role of integrated democratic accountability. *Financial Accountability & Management* **39(2)**:304–326 DOI 10.1111/faam.12354.

Cutler A, Cutler DR, Stevens JR. 2012. Random forests. In: *Ensemble Machine Learning: Methods and Applications*. New York: Springer, 157–175.

De Ville B. 2013. Decision trees. *Wiley Interdisciplinary Reviews: Computational Statistics* **5(6)**:448–455 DOI 10.1002/wics.1278.

Devlin J, Chang M-W, Lee K, Toutanova K. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. ArXiv DOI 10.48550/arXiv.1810.04805.

Frank E, Trigg L, Holmes G, Witten IH. 2000. Naive Bayes for regression. *Machine Learning* **41(1)**:5–25 DOI 10.1023/a:1007670802811.

Gaheen MM, ElEraky RM, Ewees AA. 2021. Automated students Arabic essay scoring using trained neural network by e-jaya optimization to support personalized system of instruction. *Education and Information Technologies* **26(1)**:1165–1181 DOI 10.1007/s10639-020-10300-6.

Gayed JM, Carlon MKJ, Oriola AM, Cross JS. 2022. Exploring an AI-based writing assistant's impact on English language learners. *Computers and Education: Artificial Intelligence* **3(1)**:100055 DOI 10.1016/j.caeai.2022.100055.

Ghimire R. 2022. NepaliBERT. *Available at https://huggingface.co/Rajan/NepaliBERT* (accessed 25 February 2023).

Gierl MJ, Latifi S, Lai H, Boulais A-P, De Champlain A. 2014. Automated essay scoring and the future of educational assessment in medical education. *Medical Education* **48(10)**:950–962 DOI 10.1111/medu.12517.

Guilherme A. 2019. AI and education: the importance of teacher and student relations. *AI & Society* **34**:47–54 DOI 10.1007/s00146-017-0693-8.

Gupta CP, Bal BK. 2015. Detecting sentiment in Nepali texts: a bootstrap approach for sentiment analysis of texts in the Nepali language. In: *2015 International Conference on Cognitive Computing and Information Processing (CCIP)*. Piscataway: IEEE.

Hirao R, Arai M, Shimanaka H, Katsumata S, Komachi M. 2020. Automated essay scoring system for nonnative Japanese learners. In: *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 1250–1257.

Holmes W, Tuomi I. 2022. State of the art and practice in AI in education. *European Journal of Education* **57(4)**:542–570 DOI 10.1111/ejed.12533.

Hu S, Yang Q, Yang Y. 2022. A new handwritten essay dataset for automatic essay scoring with a new benchmark. In: *Proceedings of the 2022 5th International Conference on Algorithms, Computing and Artificial Intelligence*.

Huang AY, Lu OH, Yang SJ. 2023. Effects of artificial intelligence–enabled personalized recommendations on learners' learning engagement, motivation, and outcomes in a flipped classroom. *Computers & Education* **194(6)**:104684 DOI 10.1016/j.compedu.2022.104684.

Hussein MA, Hassan H, Nassef M. 2019. Automated language essay scoring systems: a literature review. *PeerJ Computer Science* **5(2)**:e208 DOI 10.7717/peerj-cs.208.

Hwang G-J, Tu Y-F. 2021. Roles and research trends of artificial intelligence in mathematics education: a bibliometric mapping analysis and systematic review. *Mathematics* **9(6)**:584 DOI 10.3390/math9060584.

Ikram A, Castle B. 2020. Automated essay scoring (AES); a semantic analysis inspired machine learning approach: an automated essay scoring system using semantic analysis and machine learning is presented in this research. In: *Proceedings of the 12th International Conference on Education Technology and Computers*, 147–151.

Jiang D, Liu Y, Liu S, Zhao J, Zhang H, Gao Z, Zhang X, Li J, Xiong H. 2023. From CLIP to DINO: visual encoders shout in multi-modal large language models. ArXiv DOI 10.48550/arXiv.2310.08825.

Kamalov F, Santandreu Calonge D, Gurrib I. 2023. New era of artificial intelligence in education: towards a sustainable multifaceted revolution. *Sustainability* **15(16)**:12451 DOI 10.3390/su151612451.

Ke Z, Ng V. 2019. Automated essay scoring: a survey of the state of the art. In: *IJCAI*, Vol. 19, 6300–6308.

Khanuja S, Bansal D, Mehtani S, Khosla S, Dey A, Gopalan B, Margam DK, Aggarwal P, Nagipogu RT, Dave S, Gupta S, Gali SCB, Subramanian V, Talukdar PP. 2021. Muril: multilingual representations for indian languages. ArXiv DOI 10.48550/arXiv.2103.10730.

Koirala P, Niraula NB. 2021. NPVec1: word embeddings for Nepali-construction and evaluation. In: *Proceedings of the 6th Workshop on Representation Learning for NLP (RepL4NLP-2021)*, 174–184.

Lagakis P, Demetriadis S. 2021. Automated essay scoring: a review of the field. In: *2021 International Conference on Computer, Information and Telecommunication Systems (CITS)*. Piscataway: IEEE.

Lamsal R. 2020. A large scale Nepali text corpus. ieeedataport. DOI 10.21227/jxrd-d245.

Li W, Liu H. 2024. Applying large language models for automated essay scoring for non-native Japanese. *Humanities and Social Sciences Communications* **11(1)**:1–15 DOI 10.1057/s41599-024-03209-9.

Li Y, Yan Y. 2012. An effective automated essay scoring system using support vector regression. In: *2012 Fifth International Conference on Intelligent Computation Technology and Automation*. Piscataway: IEEE, 65–68.

Liang J-C, Hwang G-J, Chen M-RA, Darmawansah D. 2023. Roles and research foci of artificial intelligence in language education: an integrated bibliographic analysis and systematic review approach. *Interactive Learning Environments* **31(7)**:4270–4296 DOI 10.1080/10494820.2021.1958348.

Liu Y, Gu J, Goyal N, Li X, Edunov S, Ghazvininejad M, Lewis M, Zettlemoyer L. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics* **8**:726–742 DOI 10.1162/tacl_a_00343.

Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, Levy O, Lewis M, Zettlemoyer L, Stoyanov V. 2019. RoBERTa: a robustly optimized BERT pretraining approach. ArXiv DOI 10.48550/arXiv.1907.11692.

Loshchilov I, Hutter F. 2017. Decoupled weight decay regularization. ArXiv DOI 10.48550/arXiv.1711.05101.

Lotfy N, Shehab A, Elhoseny M, Abu-Elfetouh A. 2023. An enhanced automatic Arabic essay scoring system based on machine learning algorithms. *Computers, Materials & Continua* **77(1)**:1227–1249 DOI 10.32604/cmc.2023.039185.

Lu C, Cutumisu M. 2021. Integrating deep learning into an automated feedback generation system for automated essay scoring. *International Educational Data Mining Society* 573–579.

Maharjan G, Bal BK, Regmi S. 2019. Named entity recognition (NER) for Nepali. In: *Creativity in Intelligent Technologies and Data Science: Third Conference, CIT&DS 2019, Volgograd, Russia, September 16–19, 2019, Proceedings, Part II 3*. Cham: Springer, 71–80.

Maity S, Sahu TN, Sen N. 2021. Panoramic view of digital education in COVID-19: a new explored avenue. *Review of Education* **9(2)**:405–423 DOI 10.1002/rev3.3250.

Marinho JC, Anchiêta RT, Moura RS. 2022. Essay-BR: a Brazilian corpus to automatic essay scoring task. *Journal of Information and Data Management* **13(1)**:65–76 DOI 10.5753/jidm.2022.2340.

Mathias S, Bhattacharyya P. 2018. ASAP++: enriching the ASAP automated essay grading dataset with essay attribute scores. In: *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)*.

Mizumoto A, Eguchi M. 2023. Exploring the potential of using an AI language model for automated essay scoring. *Research Methods in Applied Linguistics* **2(2)**:100050 DOI 10.1016/j.rmal.2023.100050.

Niraula NB, Dulal S, Koirala D. 2021. Offensive language detection in Nepali social media. In: *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, 67–75.

Norton LS. 1990. Essay-writing: what really counts? *Higher Education* **20(4)**:411–442 DOI 10.1007/bf00136221.

Onesi-Ozigagun O, Ololade YJ, Eyo-Udo NL, Ogundipe DO. 2024. Revolutionizing education through ai: a comprehensive review of enhancing learning experiences. *International Journal of Applied Research in Social Sciences* **6(4)**:589–607 DOI 10.51594/ijarss.v6i4.1011.

Ouahrani L, Bennouar D. 2020. AR-ASAG an Arabic dataset for automatic short answer grading evaluation. In: *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 2634–2643.

Paul A, Purkayastha BS, Sarkar S. 2015. Hidden markov model based part of speech tagging for Nepali language. In: *2015 International Symposium on Advanced Computing and Communication (ISACC)*. Piscataway: IEEE, 149–156.

Piryani R, Piryani B, Singh VK, Pinto D. 2020. Sentiment analysis in Nepali: exploring machine learning and lexicon-based approaches. *Journal of Intelligent & Fuzzy Systems* **39(2)**:2201–2212 DOI 10.3233/JIFS-179884.

Pradhan A, Yajnik A. 2021. Probabilistic and neural network based POS tagging of ambiguous Nepali text: a comparative study. In: *2021 International Symposium on Electrical, Electronics and Information Engineering*, 249–253.

Pudasaini S. 2021. NepNewsBERT. *Available at https://huggingface.co/Shushant/NepNewsBERT* (accessed 25 February 2023).

Ramalingam V, Pandian A, Chetry P, Nigam H. 2018. Automated essay grading using machine learning algorithm. In: *Journal of Physics: Conference Series.* Vol. 1000. Bristol, UK: IOP Publishing, 012030.

Ramesh D, Sanampudi SK. 2022. An automated essay scoring systems: a systematic literature review. *Artificial Intelligence Review* **55(3)**:2495–2527 DOI 10.1007/s10462-021-10068-2.

Rauniyar K, Poudel S, Shiwakoti S, Thapa S, Rashid J, Kim J, Imran M, Naseem U. 2023. Multi-aspect annotation and analysis of Nepali tweets on anti-establishment election discourse. *IEEE Access* **11**:143092–143115 DOI 10.1109/ACCESS.2023.3342154.

Reafat M, Ewees A, Eisa M, Ab Sallam A. 2012. Automated assessment of students Arabic free-text answers. *International Journal of Cooperative Information Systems* **12**:213–222.

Rodriguez PU, Jafari A, Ormerod CM. 2019. Language models and automated essay scoring. ArXiv DOI 10.48550/arXiv.1909.09482.

Rodríguez ML, Pulido-Montes C. 2022. Use of digital resources in higher education during COVID-19: a literature review. *Education Sciences* **12(9)**:612 DOI 10.3390/educsci12090612.

Rudner LM, Liang T. 2002. Automated essay scoring using Bayes' theorem. *The Journal of Technology, Learning and Assessment* **1(2)**.

Sanh V, Debut L, Chaumond J, Wolf T. 2019. Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. ArXiv DOI 10.48550/arXiv.1910.01108.

Schapire RE. 2013. Explaining adaboost. In: *Empirical Inference*. Cham: Springer, 37–52.

Seßler K, Fürstenberg M, Bühler B, Kasneci E. 2025. Can AI grade your essays? a comparative analysis of large language models and teacher ratings in multidimensional essay scoring. In: *Proceedings of the 15th International Learning Analytics and Knowledge Conference*, 462–472.

Shahi TB, Pant AK. 2018. Nepali news classification using Naive Bayes, support vector machines and neural networks. In: *2018 International Conference on Communication Information and Computing Technology (ICCICT)*. Piscataway: IEEE.

Shahi T, Sitaula C, Paudel N. 2022. A hybrid feature extraction method for Nepali COVID-19-related tweets classification. *Computational Intelligence and Neuroscience* **2022(3)**:1–11 DOI 10.1155/2022/5681574.

Shermis MD, Burstein JC. 2003. *Automated essay scoring: a cross-disciplinary perspective*. England, UK: Routledge.

Shin D, Shim J. 2021. A systematic review on data mining for mathematics and science education. *International Journal of Science and Mathematics Education* **19(4)**:639–659 DOI 10.1007/s10763-020-10085-7.

Shrestha D. 2021. DistillBERT(Nepali). *Available at https://huggingface.co/dexhrestha/Nepali-DistilBERT* (accessed 25 February 2024).

Shrestha BB, Bal BK. 2020. Named-entity Based sentiment analysis of Nepali news media texts. In: *Proceedings of the 6th Workshop on Natural Language Processing Techniques for Educational Applications*, 114–120.

Shrestha S, Haque S, Dawadi S, Giri RA. 2022. Preparations for and practices of online education during the COVID-19 pandemic: a study of Bangladesh and Nepal. *Education and Information Technologies* **27(1)**:243–265 DOI 10.1007/s10639-021-10659-0.

Singh S, Pupneja A, Mital S, Shah C, Bawkar M, Gupta LP, Kumar A, Kumar Y, Gupta R, Shah RR. 2023. H-AES: towards automated essay scoring for Hindi. ArXiv DOI 10.48550/arXiv.2302.14635.

Singh OM, Timilsina S, Bal BK, Joshi A. 2020. Aspect based abusive sentiment detection in Nepali social media texts. In: *2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. Piscataway: IEEE, 301–308.

Sinha M, Sharma S, Dasgupta T, Basu A. 2012. New readability measures for Bangla and Hindi texts. In: *Proceedings of COLING 2012: Posters*, 1141–1150.

Sitaula C, Shahi TB. 2024. Multi-channel CNN to classify Nepali COVID-19 related tweets using hybrid features. *Journal of Ambient Intelligence and Humanized Computing* **15(3)**:2047–2056 DOI 10.1007/s12652-023-04692-9.

Song W, Zhang K, Fu R, Liu L, Liu T, Cheng M. 2020. Multi-stage pre-training for automated Chinese essay scoring. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 6723–6733.

Song S, Zhao J. 2013. *Automated essay scoring using machine learning*. California, USA: Stanford University.

Song Y, Zhu Q, Wang H, Zheng Q. 2024. Automated essay scoring and revising based on open-source large language models. *IEEE Transactions on Learning Technologies* **17(2)**:1880–1890 DOI 10.1109/tlt.2024.3396873.

Stahl M, Biermann L, Nehring A, Wachsmuth H. 2024. Exploring LLM prompting strategies for joint essay scoring and feedback generation. ArXiv DOI 10.48550/arXiv.2404.15845.

Suthaharan S, Suthaharan S. 2016. Decision tree learning. In: *Machine Learning Models and Algorithms for Big Data Classification: Thinking with Examples for Effective Learning*, 237–269.

Suárez PJO, Sagot B, Romary L. 2019. Asynchronous pipeline for processing huge corpora on medium to low resource infrastructures. In: *7th Workshop on the Challenges in the Management of Large Corpora (CMLC-7)*. Germany: Leibniz-Institut für Deutsche Sprache.

Tang Y, Tran C, Li X, Chen P-J, Goyal N, Chaudhary V, Gu J, Fan A. 2020. Multilingual translation with extensible multilingual pretraining and finetuning. ArXiv DOI 10.48550/arXiv.2008.00401.

Thapa S, Adhikari S, Naseem U, Singh P, Bharathy G, Prasad M. 2020. Detecting alzheimer's disease by exploiting linguistic information from Nepali transcript. In: *Neural Information Processing: 27th International Conference, ICONIP 2020, Bangkok, Thailand, November 18–22, 2020, Proceedings, Part IV 27*. Cham: Springer, 176–184.

Timilsina S, Gautam M, Bhattarai B. 2022. NepBERTa: Nepali language model trained in a large corpus. In: *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing*, 273–284.

Vajjala S. 2018. Automated assessment of non-native learner essays: investigating the role of linguistic features. *International Journal of Artificial Intelligence in Education* **28**:79–105 DOI 10.1007/s40593-017-0142-3.

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I. 2017. Attention is all you need. *Advances in Neural Information Processing Systems* **30**:6000–6010.

Wang Y, Ma C, Dong Q, Kong L, Xu J. 2023. A challenging benchmark for low-resource learning. ArXiv DOI 10.48550/arXiv.2303.03840.

Wang Y, Wang C, Li R, Lin H. 2022. On the use of BERT for automated essay scoring: joint learning of multi-scale essay representation. ArXiv DOI 10.48550/arXiv.2205.03835.

Weigle SC. 2013. English language learners and automated scoring of essays: critical considerations. *Assessing Writing* **18(1)**:85–99 DOI 10.1016/j.asw.2012.10.006.

Wenzek G, Lachaux M-A, Conneau A, Chaudhary V, Guzmán F, Joulin A, Grave E. 2019. CCNet: extracting high quality monolingual datasets from web crawl data. ArXiv DOI 10.48550/arXiv.1911.00359.

Wu Y, Schuster M, Chen Z, Le QV, Norouzi M, Macherey W, Krikun M, Cao Y, Gao Q, Macherey K, Klingner J, Shah A, Johnson M, Liu X, Kaiser L, Gouws S, Kato Y, Kudo T, Kazawa H, Stevens K, Kurian G, Patil N, Wang W, Young C, Smith J, Riesa J, Rudnick A, Vinyals O, Corrado G, Hughes M, Dean J. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. ArXiv DOI 10.48550/arXiv.1609.08144.

Xiao C, Ma W, Song Q, Xu SX, Zhang K, Wang Y, Fu Q. 2025. Human-AI collaborative essay scoring: a dual-process framework with LLMs. In: *Proceedings of the 15th International Learning Analytics and Knowledge Conference*, 293–305.

Xie J, Cai K, Kong L, Zhou J, Qu W. 2022. Automated essay scoring via pairwise contrastive regression. In: *Proceedings of the 29th International Conference on Computational Linguistics*, 2724–2733.