10.48047/jocaaa.2023.31.02.35

Automatic Essay Score Prediction in Bengali

Piyal Roy[1], Rajat Pandit[1*] and Sudip Kumar Naskar[2]

[1]Computer Science, West Bengal State University, Barasat, West Bengal, India.
[2]Computer Science and Engineering, Jadavpur University, Jadavpur, West Bengal, India.

*rajatpandit123@gmail.com

Abstract:
Automatic essay scoring (AES) is a crucial task in educational technology that aims to evaluate student essays objectively and accurately. While numerous AES models have been proposed over the years for languages with large spoken populations, there has been very little research on this topic for low-resource and morphologically complex languages, such as Bengali. To address this gap, we developed a custom dataset in Bengali designed explicitly for this task, on which we carried out comprehensive experiments using various machine learning and language processing. The essays were preprocessed and analyzed to extract features relevant to the quality of the essays. We employed multiple regression models, including simple linear regression, Decision Tree Regression, Ridge Regression, LASSO Regression, Elastic Net Regression, and Support Vector Regression (SVR) models, to predict the scores for Bengali essays. Furthermore, we also conducted experiments with state-of-the-art large language models, such as ChatGPT and Gemini, to assess their performance on this task. They were evaluated using the Pearson Coefficient. In our experiments, the Decision Tree Regressor outperforms other models, achieving a Pearson correlation coefficient of 0.905. However, we found that LLMs are not well-suited for this task, as they do not achieve the same performance as traditional regression models. Only ChatGPT performed well, with a Pearson correlation of 0.605. The paper's findings suggest that integrating LLMs into the system's regression-based AES may lead to an overall increase in the system's accuracy for Bengali.

Keywords: AES, Regression, LLM, Perplexity, NLP.

1. INTRODUCTION

1.1. Background and Motivation

Essays play a significant role in the curricula of schools and universities worldwide. It serves as a good instrument to evaluate students' understanding, competence, and critical thinking abilities. Essay evaluation is a problem that has been machine-done since it has manual conduct. It is time-consuming, expensive, and prone to human error as the manual essay scoring process is comprehensive but is still in use today. Believing that there is a growing demand for essays that are to be evaluated, it becomes necessary to automate the scoring system. There are also systems, such as AES, based on machine learning algorithms that evaluate and score human-written texts; an example of the contribution the field of Natural Language Processing brings. There are several benefits to Automated Essay Scoring: it has high reliability, is scalable, scoring is consistent, and this translates to reduced time spent scoring. For both theoretical and experimental works on AES for English and other overt languages, research has been conducted in the literature. However, research on low-resource languages, such as Bengali, is still in its nascent stages and needs to be explored further.

1.2. Importance of Essay Scoring in Educational Domain

Essays are an essential part of the educational system, as they test a student's ability to write, identify ideas, analyze, and present a coherent argument. They are extensively used in tests such as standardized tests, classroom evaluations, and admissions processes. Particularly in the widely spoken language of Bengali, a system incorporating the use of an automated essay scoring system is of the utmost importance. These systems also enhance fairness and affordability and allow wider audience to evaluate. The AES is needed in countries where Bengali is one of the official or regional languages. The reason for this is that many students will be used for formal tests and standard assessments. A Bengali essay would be scored with an automated system because this would not only reduce manual load but would also provide students with timely performance feedback, and as a result promote a culture of continuous learning and development.

10.48047/jocaaa.2023.31.02.35

1.3. Challenges Specific to Bengali Essay Scoring

The presentation of the development challenges of an Automated Essay Scoring (AES) system for Bengali, due to the scarcity of annotated resources and linguistic properties of Bengali are given. Bengali is an Indo-Aryan, or Indo-European, language. It has a great deal of inflection: several suffixes that can change the meaning of words, and complex verb conjugation. The feature extraction and model training become much more complicated due to this linguistic complexity. In addition, we face an insurmountable hurdle – there are no well-annotated Bengali essay datasets that we need to map and evaluate in order to match the students' work. While large corpora with standardized scores are available for English, this is not the case for Bengali. This lack of data also negatively affects the training and evaluation of reliable AES models. Since the essays may have a different nature and writing style due to cultural and contextual factors prevalent in Bengali-speaking regions, it may be necessary to develop feature extraction and model training for this region through targeted adjustments.

1.4. Objectives and Scope of the Research

The purpose of this research is to develop an efficient and secure AES system for Bengali essays. Compiling a comprehensive dataset of Bengali essays with standardized scorers' evaluations is what this amounts to. To train and evaluate the AES models, the annotated essay dataset from Open Assistant will be used. The task in this work is to identify and extract meaningful linguistic, semantic, and syntactical features of the Bengali essays and then train models to score essays being evaluated by essay scoring. As I mentioned earlier, we will utilize various deep learning and machine learning algorithms to score an essay. The standard metrics will be used to train and test the models for selection and validation. However, we hope that our study can contribute to the educational evaluation of low-resource languages and should be relevant to both the educational technology community and natural language processing.

2. LITERATURE REVIEW

Because of its ability to provide easily comparable, accurate, and exhaustive assessment of written language, the task of AES has attracted a great deal of interest within NLP. The AES roots can be traced back to the early 1960s when Page introduced “Project Essay Grade” (PEG), which used simple statistical indices to serve as an indicator of essay quality (Page, 1966). The centuries of AES have progressed from initial methods of data handling over the 1990s to today’s learning systems. Despite the evolution of computational linguistics over the years, there have been many techniques and algorithms offered in literature, including latent semantic analysis (LSA) (Landauer, Foltz, & Laham, 1998), support vector machine (SVM) (Dikli, 2006), and the recent one such as deep learning such as convolutional neural networks (CNN) and recursive neural networks (RNN) (Dong & Zhang, 2016).

While a considerable amount of research has been conducted on AES in English, a problem arises when implementing AES in other languages, specifically the Bengali language. To cite an instance, Mahjabin et al. (2024) acknowledge that Bengali is characterized by high linguistic density and morphological variety, which necessitates personalization at the essay rater level for accurate calculation. Training accurate models is made challenging by the lack of Bengali corpora with annotations, as noted by Chakraborty et al. (2021). Although feature engineering has always been an important part of constructing AES systems, it remains an integral component of these systems to this date. Shermis and Burstein (2013) refer to them as the first generation, and they were primarily based on basic count statistics, such as word frequencies, syntactic density, and lexical variety. However, due to the latest advancements in neural networks, embeddings, and attention mechanisms (Vaswani et al., 2017), feature extraction has mainly been replaced. Recently, some works using Bengali have employed character-level embedding in conjunction with bidirectional LSTMs (Karim et al., 2020). It has been proven that these methods are fruitful, and in these evaluations, there is

10.48047/jocaaa.2023.31.02.35

potential to conduct further research on these models to optimize them for the required characteristics of Bengali essays.

The results generated by the models (whether correct or not) are then compared to the assessment scores provided by human annotators for the AES systems. As suggested by Williamson (2009), the commonly used assessment tools are Pearson correlation, quadratic weighted kappa, and accuracy. However, the debate on the use of Bengali has arisen due to similar use caused by the absence of a standard evaluation dataset. This has led to variations in the outcomes of different studies (Islam & Hoque, 2013). This will help avoid these discrepancies and determine the models that would be useful for rating. Uniform feedback provided to students from remote and low-income areas can be equally effective in promoting equity among Bengali learners through an effective AES system, as noted by Saha and Pal (2020). In addition, the use of If AES is integrated with other technologies as a means of delivering effective learning appropriate for the student's needs.

Advanced limes have already been developed for the AES systems; however, their implementation in Bengali is a significant issue, both in terms of development and prospects. The Bengali language is characterized by its unique nature, for which current models need to be modified or new techniques are required to create a language that is specifically appropriate for the Bengali language.

3. METHODOLOGY

We developed an automated essay scoring system for the Bengali language with a systematic approach. To handle the unique features of the Bengali language, such as its limited word inventory, compound and camouflage words, and a large number of tags, a rigorous and reproducible framework for data collection, preprocessing, feature extraction, and model selection, including both machine learning and deep learning models, was employed. The set of evaluation metrics was used to evaluate the efficacy of these models. These evaluation metrics are used to assess model performance against human grading standards.

3.1. Description of the Dataset

For the Bengali language, the dataset was carefully prepared to verify the impact of AES on the system's performance. The following subsections describe the process of data collection, preprocessing, and annotation.

3.1.1. Source of Essays

Our dataset comprises essays written by undergraduate (UG) and postgraduate (PG) students of Bengali origin from a university. The students were taught on a standardized basis to write an essay on the topic of "একটি ভ্রমণের অভিজ্ঞতা" [A Tour Experience], as this will ensure consistency and relevance. This also meant that the students were given a 10-minute time limit to complete their essays, ensuring a relatively uniform length and depth of their answers. Figure 1 is a blank form in the format given to students; Figure 2 is the completed form after a student has filled it out.

10.48047/jocaaa.2023.31.02.35

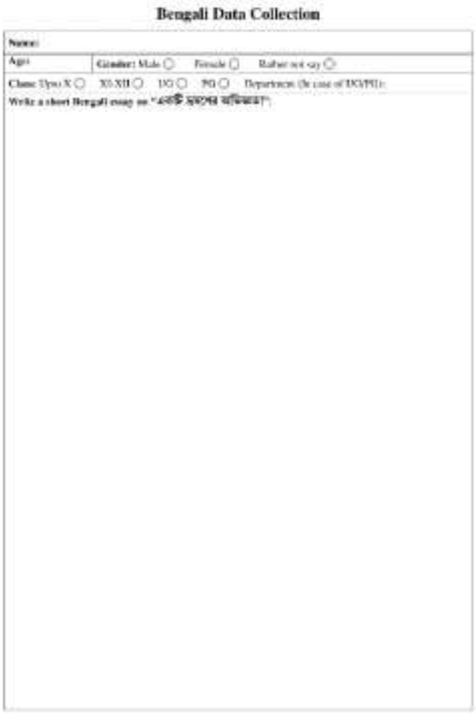

Figure 1: Blank essay collection form.

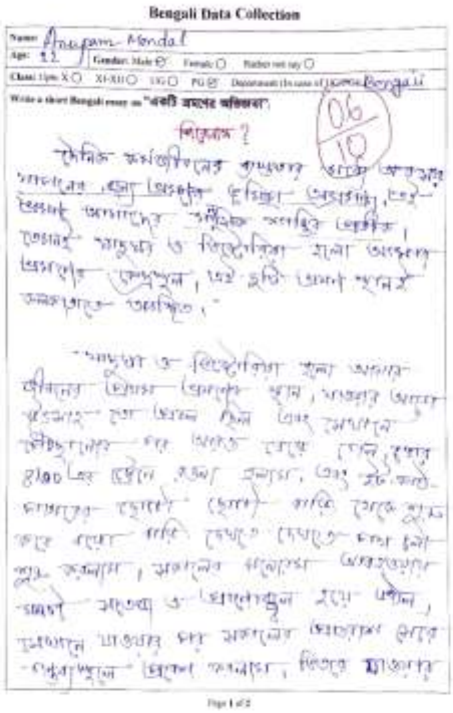

Figure 1: Blank essay collection form. **Figure 2:** Filled essay collection form.

A total of 68 essays were collected during this process. The diversity in the academic levels of the participants (UG and PG) provided a range of writing abilities, making the dataset more representative and comprehensive for the task of essay scoring.

3.1.2. Data Preprocessing Steps

Firstly, the handwritten texts were digitized in the data preparation process. In this, I had to copy each essay as exact text and save it in a CSV (Comma Separated Values) file. Converted in the form of text for the digitizing ones, the handwritten essays. After the transcription, several preprocessing steps were taken.

Verification of Transcriptions: All the transcribed text was carefully compared to handwritten responses to confirm the accuracy of the digitization. Hence, this effort was made to obtain a 'careful' comparison, intended to minimize transcription errors (such as misreading handwriting) and, therefore, ensure high reliability in the data.

Formatting Consistency: All the essays were thoroughly formatted to have correct punctuation, spacing, and overall structure. This process was designed to address some of the inevitable issues that could arise during the transcription process and maintain a consistent formatting style. The accuracy and integrity of the data were to be maintained by ensuring uniformity.

Data organization: The transcribed essay was matched with the score that was assigned to it by the human grader and placed into a structured CSV file. The whole essay text file was stored in one column of each row of the file, while the associated score was stored in another adjacent column. The organization ensured that, in the later stages of model training, the setup provided seamless access and analysis.

The prep work for this dataset was essential, as these preprocessing steps helped ensure that the essays comprising the data were adequately digitized and formatted for the development of this essay scoring model.

3.1.3. Annotation and Scoring

10.48047/jocaaa.2023.31.02.35

The digitized and preprocessed essays were evaluated by a domain expert with extensive knowledge of the Bengali language and literature. The expert graded each essay based on a set of predefined criteria, which included coherence and organization, assessing the essay's structural flow and logical presentation of ideas; language and grammar, focusing on accuracy, complexity, and proper use of grammar, syntax, and vocabulary; content relevance, evaluating the alignment of the essay with the given topic and the depth of the ideas presented; and creativity and originality, considering the uniqueness and creativity in expressing ideas. It was these criteria that guaranteed a thorough and uniform evaluation of the essay.

Each essay was scored by an expert on a scale of 1 to 10, with 1 indicating a poorly written essay with little or no relevance or coherence and 10 indicating an excellently written essay that met all the parameters to the highest extent. In the next column of the corresponding CSV file, the essay text was aligned to the recorded score for each essay.

	A	B
1	**essay**	**marks**
2	দীঘা ভ্রমন দীঘা একটি পর্যটন কেন্দ্র। এটি বঙ্গোপ সাগরের তীরে অবস্থিত। দীঘা একটি বিখ্যাত পর্যটন কেন্দ্রের মধ্যে একটি। এখানে বহু মানুষ ছুটি পেলে সময় কাটাতে আসে। সমুদ্র দেখতে আসে এবং স্বাস্থ্যকর পরিবেশে কিছুদিন সময় কাটায়। এখানে অনেক নামিদামি হোটেল আছে। অনেক পর্যটকরা এসে হোটেলে থাকে তাদের হোটেল থেকে আয় করে জীবিকা সংগ্রহ করে। এছাড়া অনেক ছোট ছোট বোট, লঞ্চ, স্টিমার দেখা যায়। ভারতবর্ষের বহু মানুষ এখানে বসবাস করে। অন্য সকলের মত আমিও বেড়াতে ভালোবাসি। দীঘা সমুদ্র দেখার আমার খুব ইচ্ছা ছিল। গত সরস্বতী পূজার ছুটিতে আমার মা, বাবা, দিদি আর আমি দীঘা বেড়াতে গিয়েছিলাম। হাওড়া থেকে কালকা মেলে চেপে দিল্লি বেড়াতে গিয়েছিলাম। সেই প্রথমবার আমি দূরপাল্লার ট্রেনের চেপে যেতে মনে দারুন আনন্দ হচ্ছিল। বৈকাল পাঁচটায় ট্রেনে চেপে সারারাত ট্রেনে কাটিয়ে ভোর তিনটি দিঘা পেছনে পৌঁছালো। নিউ দীঘা স্টেশনে। ঝকঝকে একটা সুন্দর স্টেশন দেখলাম। সেখান থেকে আমরা একটি অটো নিয়ে একটা হোটেলে পৌঁছালাম। প্রাকৃতিক সৌন্দর্য:- খুব সুন্দর এই হোটেল। অনেক ছোট ছোট ঘর। হোটেলের দুতলা ঘরে আমাদের জায়গা হল। ভোরের আবছা আলোয় সমুদ্র দেখতে পেয়ে মন আনন্দে ভরে গেল। বেরিয়ে পরলাম হোটেল থেকে সূর্য দেখতে। অদ্ভুত এক দৃশ্য দেখলাম। প্রথমে মনে হল সূর্য সমুদ্রের জল থেকে উঠছে। আমার সামনে পিছনে দুদিকে দুরবাম জলের রং। সূর্য যেন খুব তাড়াতাড়ি উঠে পড়ল। এখনো শুধু সেই স্মৃতি চিহ্নটা মনে পড়ে। মনে হয় আবার যাবো সমুদ্র সৈকত দেখতে।	6
3	বকখালি ভ্রমণ আমার প্রিয় উইকেন্ড ডেস্টিনেশন গুলির মধ্যে পশ্চিমবঙ্গের বকখালি অন্যতম। যদিও এখানকার সমুদ্র দীঘার মত যদিও খুব উত্তাল নয় এবং সব সময় ঢেউ পাওয়া যায় না, জোয়ারের সময় ছাড়া। বিচ এর সাইড এ অনেক দোকান আছে ইচ্ছা হলে টুকি টাকি ঘর সাজানোর জিনিস কিনতে পারা যায়। ট্রেনের টিকিট বুকিং নিয়ে কোন চিন্তা নেই। এবার আমার বকখালি যাত্রা শুরুর পরিচয় দেই। তখন সময় সকাল 6:40। হাসনাবাদ শিয়ালদহ লোকাল ধরে আমি শিয়ালদহ স্টেশনে পৌঁছালাম। যেখান থেকে নামখানে লোকাল ধরে বকখালি গিয়ে পৌঁছালাম। মোটামুটি সময় হবে 11:30 মতো। সেখানে গিয়েই সমুদ্রের দৃশ্যটা অতি সুন্দর। তো আমরা কয়েকজন বন্ধু মিলে গিয়েছিলাম। ওখানে গিয়ে পরে আমরা আগে একটা Hotel খুজলাম কারণ আমাদের ইচ্ছা ছিল যে দুই রাত আমরা ওখানে কাটাবো। যথারীতি সেই দিন আমরা সকালে দোকান থেকে খাওয়া দাওয়া করে ঘুরতে বেরিয়ে পড়লাম। বকখালি তে সন্ধ্যের থেকে দৃশ্যটা আরো সুন্দর লাগে তার কারণ সমস্ত দোকান শারি সারি লাইন দিয়ে থাকে। সমস্ত রকমের ঘর সাজানোর জিনিস পাওয়া যায় এখানে। এক হিসেবে ওখানে যারা নিত্য যাতায়াত করে বা ওখানকার বাসিন্দা তারা খুবই ভালো মনের মানুষ তার কারণ হল আমার একটা বন্ধুর শরীর খারাপ হওয়ায় আমরা কি করব সেটা বুঝতে না পেরে একটা দাদাকে ডেকেছিলাম সেই দাদাটা শুধু আমাদের কে পরামর্শই দেয়নি, আমাদেরকে নিজে Doctor এর কাছে নিয়ে গিয়েছিলেন। এই থেকে তাদের আচার-আচরণের অনেকটা পরিচয় মিলে এবং সেখানকার পরিবেশ শুধু মৌসুমী প্রকৃতির হয়ে থাকে। তাই খুব ভালোভাবেই মানুষ সেখানে মিশে যেতে পারে। শীতকালেই বেশি যেতে ভালো লাগে। সেখানকার লোক সংস্কৃতির দিক দিয়ে দেখতে গেলে দেখা যায় যে কোন রকম কোলা হল নেই। শান্ত পরিবেশ হয়ে থাকে। অনেক ভালো গাছ,নারকেল গাছ দেখতে পাওয়া যায় এবং নানা ধরনের ফুল গাছ লাগানো থাকে যা একটা গ্রাম্য সুন্দর পরিবেশের বার্তা বহন করে আর আমি ওখানে গিয়ে অনেক খুশি।	6

Figure 3: Sample of the CSV file.

The essay scores were predicted by taking the raw text of each essay as input and using the LLMs, ChatGPT, and Gemini. They deployed their advanced deep learning architectures, based on layers of transformers, attention mechanisms, and other processing techniques, on the text to generate a score. For this specific task, we employed few-shot learning to improve the performance of these models, that is, to learn to produce essay scores given only a small sample of example essays with their corresponding accurate scores. By applying this approach, models were enabled to learn more effectively from the pre-training task, utilizing their skills while also enhancing their understanding of the scoring criteria used in this work.

3.3.3. Training Process and Hyperparameter Tuning

This dataset was divided into training and test sets on an 80:20 basis. The models were trained on the training set and tested on the test set.

Fitting the regression models to the training data by minimizing the loss function was a key part of the training process. Based on this, hyperparameter tuning was performed on models such as Ridge,

10.48047/jocaaa.2023.31.02.35

LASSO, and Elastic Net to determine the optimal regularization parameters, which define the strength of the penalty applied to the model coefficients.

ChatGPT and Gemini, the large language models, were not trained or tuned in this setting, as they were pre-trained models applied pointwise to the task. Finally, their performance was evaluated on the test set together with the regression models.

3.4. Evaluation Metrics

To evaluate the effectiveness of the AES system, we must assess the model's performance. In our experiment, we used the Pearson correlation coefficient as the primary measure to evaluate model predictions against the human expert's actual scores.

The Pearson correlation coefficient (r) is a statistical measure that indicates the strength and direction of the linear relation between the predicted essay scores and the human grader-assigned scores (Rodgers & Nicewander, 1988). The value of the Pearson correlation ranges from -1 to +1, where:

- +1 indicates a perfect positive linear relationship, meaning that as the predicted scores increase, the actual scores also increase proportionally.
- 0 indicates no linear relationship between the predicted and actual scores.
- -1 indicates a perfect negative linear relationship, meaning that as the predicted scores increase, the actual scores decrease proportionally.

In our work, we prefer a higher positive Pearson correlation coefficient because it indicates that our model is more closely aligned with human measures. This is a simple, interpretable metric for which a scoring system is commonly evaluated (Cohen et al., 2002). We calculated the predictions for each model on the test set and computed the Pearson correlation for each model's predictions to compare their direct performance. Important for the deployment of an AES was a metric that was both transparent and interpretable of the model's ability to reproduce the scoring patterns of the human expert.

4. EXPERIMENTS AND RESULTS

4.1. Experimental Setup

This research relates to experiments running to investigate the effectiveness of different regression models and large language models (LLMs) in predicting essay scores. We used Simple Linear Regression, Decision Tree Regression, Ridge Regression, LASSO Regression, Elastic Net Regression, Support Vector Regression regression models. Also, with these new large language models (LLMs), like ChatGPT and Gemini, cutting edge language processing techniques were applied to discover new use cases to be used for essay scoring with the technologies.

All models were built using Scikit-learn, one of the popular machine learning tools of the language Python (Pedregosa et al., 2011). The experimentation was carried on the Google Colab platform with sufficient computational power available in hand, and it was the best way to do machine learning on a large scale.

Moreover, we used perplexity score as another metric to evaluate how well model performs. Perplexity gave deeper insights into how the models could understand and process language.

4.2. Results of Different Models

We evaluated the performance of each model using the Pearson correlation coefficient, a standard metric for determining the linear relationship between the actual essay scores and the predicted ones. To assess the impact of the perplexity score on the models' performance, the raw Pearson correlation coefficients are compared to those obtained after the metric is introduced (Table 1).

Model	Without Perplexity	With Perplexity
Simple Linear Regression	0.507	0.598
Decision Tree Regression	0.875	0.905

10.48047/jocaaa.2023.31.02.35

Ridge Regression	0.514	0.536
LASSO Regression	0.514	0.588
Elastic Net Regression	0.536	0.530
Support Vector Regression	0.488	0.250

Table 1: Pearson Correlation with and without Perplexity

Table 2 presents the Pearson correlation coefficients for each model, indicating their respective abilities to predict essay scores, including those of the LLMs accurately.

Model	**Pearson Correlation**
Simple Linear Regression	0.598
Decision Tree Regression	0.905
Ridge Regression	0.536
LASSO Regression	0.588
Elastic Net Regression	0.530
Support Vector Regression	0.250
ChatGPT	0.605
Gemini	0.266

Table 2: Pearson Correlation of Different Models.

4.3. Result Analysis

The results of the experiments demonstrate several important results. The best model resembles Decision Tree Regression, with a Pearson correlation coefficient of 0.905, indicating a robust relationship between the predicted and actual essay scores. This model most likely outperformed its counterparts due to its ability to effectively capture nonlinear relationships in the data. Among all the large language models, ChatGPT demonstrated superior performance compared to other models, including Support Vector Regression and traditional regression methods such as Ridge and LASSO Regression, as indicated by a Pearson correlation coefficient of 0.605. What this means is that ChatGPT and such advanced language models have good performance in essay scoring. Typically, adding the perplexity score to the list of metrics improved the Pearson correlation coefficients of all of these models, except for a few. In addition, Decision Tree Regression and Simple Linear Regression models also received an important contribution from perplexity, which means that its usage is indeed beneficial to predicting the essay score. However, in all, we see that the additional metric of perplexity does little for the performance of our Support Vector Regression model, not a good indication that this is a good metric to use with for this model. This means that when constructing automatic essay scoring systems, it is crucial to choose in which type of models and evaluation metrics carefully, especially in cases where the nuance of the language matters. We find with the use of advanced language processing techniques and more metrics (perplexity, among other things) the accuracy and reliability of a system using the traditional regression models suffer little if at all in performance.

5. CONCLUSION & FUTURE WORK

In this work, we study the comparison of regression models and large language models (LLMs) in this task, where learning to predict essay scores (on a scale of 1 to 100) of essays in the Bengali language is considered. Also among the traditional models the results indicated that Decision Tree Regression better showed a high Pearson correlation with actual scores than the others. Furthermore, the grown up ChatGPT also outperformed many classic models and also among LLMs. On the whole, adding the perplexity score as an additional neccesary metric may improve model performance for Decision Tree Regression and Simple Linear Regression. These findings results show that it is also

10.48047/jocaaa.2023.31.02.35

possible to join the traditional machine learning techniques with the best of the class of language models to create robust and the accurate automatic essay scoring systems.

The use of regression models and large language models (LLMs) for essay scoring in Bengali as presented by our work could also be of interest to the larger field. In the context of automated assessment, this language has been relatively little noticed. Finally, for future research, it is suggested that one will study further validation of the models by fine-tuning their LLMs on much more extensive and diverse datasets.

REFERENCES

[1] Mahjabin, R., Farzana, S., Zerin, N., Chowdhury, S. S., & Hossain, I. (2024). Automated essay scoring for Bangla (Doctoral dissertation, Brac University).

[2] Chakraborty, S., Nayeem, M. T., & Ahmad, W. U. (2021, May). Simple or complex? Learning to Predict Readability of Bengali Texts. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 35, No. 14, pp. 12621-12629).

[3] Islam, M. M., & Hoque, A. L. (2013, May). Automated Bangla essay scoring system: ABESS. In 2013 International Conference on Informatics, Electronics, and Vision (ICIEV) (pp. 1-5). IEEE.

[4] Dikli, S. (2006). "An Overview of Automated Scoring of Essays." The Journal of Technology, Learning, and Assessment, 5(1), 4-28.

[5] Dong, F., & Zhang, Y. (2016). "Automatic Features for Essay Scoring." Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 1071-1080.

[6] Landauer, T. K., Foltz, P. W., & Laham, D. (1998). "An Introduction to Latent Semantic Analysis." Discourse Processes, 25(2-3), 259-284.

[7] Karim, M. R., Chakravarthi, B. R., McCrae, J. P., & Cochez, M. (2020, October). Classification benchmarks for the under-resourced Bengali language based on a multichannel convolutional LSTM network. In 2020, IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA) (pp. 390-399). IEEE.

[8] Page, E. B. (1966). "The Imminence of… Grading Essays by Computer." The Phi Delta Kappan, 47(5), 238-243.

[9] Shermis, M. D., & Burstein, J. (2013). Handbook of Automated Essay Evaluation: Current Applications and New Directions. Routledge.

[10] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). "Attention is All You Need." Advances in Neural Information Processing Systems, 30, 5998-6008.

[11] Williamson, D. M. (2009). "A Framework for Implementing Automated Scoring." Journal of Educational Measurement, 46(2), 135-150.

[12] Rodgers, J. L., & Nicewander, W. A. (1988). Thirteen ways to look at the correlation coefficient. The American Statistician, 42(1), 59-66.

[13] Cohen, L., Manion, L., & Morrison, K. (2002). Research methods in education. routledge.

[14] Heafield, K. (2011, July). KenLM: Faster and more minor language model queries. In Proceedings of the sixth workshop on statistical machine translation (pp. 187-197).

[15] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825-2830.