Automated Evaluation of Speech Responses using Deep Learning

Nikith Birru
ID: 11654339
College of Information
University of North Texas
nikhithbirru@my.unt.edu

Naquibuddin Shaik
ID: 11602741
College of Information
University of North Texas
NaquibuddinShaik@my.unt.edu

Pavan Mahindra Varma Penmethsa
ID: 11597110
College of Information
University of North Texas
Pavanmanindravarmapenmethsa@my.unt.edu

Bhanu Prasad Kommula
ID: 11590771
College of Information
University of North Texas
Bhanuprasadkommula@my.unt.edu

Eshwitha Naini
ID: 11609231
College of Information
University of North Texas
eshwithanaini@my.unt.edu

***Abstract*—Deep learning model for development of an automated grading of english text is the focus of our project. Our objective is to address the very time consuming and very subjective nature of manual grading, in particular the educational and language assessment contexts to that of an automated process. By using the advancements and leveraging the deep learning models potential which can be used for automated grading of english text with a higher accuracy and exact same consistency every time. For our methodology we will be collecting data, then preprocess it, development of model, training and then evaluation. With this extensive testing and validation, our model's effectiveness can be assessed on various criterias which will include grammar, coherence and relevance. All of these findings will highlight the potential of automated grading and will also streamline the assessment process. Also this will save a lot of time and give timely feedback to the learners immediately. There is a wide range of implementations for this research. It can be extended to various industries that will require automation in textual analysis and by using content moderation and grammatical analysis. Hence the model that we have created will open many doors for future development in the process of automated language evaluation and in textual analysis tasks by leveraging the use of deep learning and NLP.**

***Index Terms*—Automated grading, Language assessment, Textual analysis, Natural language processing (NLP)**

I. Introduction and Problem Statement

An evaluation of a test taker's ability to utilize the English language, especially for non-native speakers, is conducted globally through the administration of an English proficiency test, or EPT. In many academic and professional contexts, the EPT is essential since it is a prerequisite for entrance to many colleges and because it opens up job chances. Through language proficiency evaluation, it also facilitates immigration procedures. Various activities that evaluate a candidate's capacity to articulate thoughts, synthesize data, and participate in intellectual discourse within a designated time span are commonly included in the EPT. Our main concern with this project is the English proficiency test's manual grading system. Speaking answers must be evaluated by human reviewers according to predetermined standards for grammar, vocabulary, fluency, and pronunciation. Because of the labor-intensive nature of this method and its susceptibility to human bias, assessments from various examiners and test administrations may differ from one another. Questions concerning the general dependability of the exam are raised by the variability introduced by human grading, which is subjective. Our goal is to showcase the capabilities of deep learning models in order to tackle these issues and improve language competence evaluation. Test takers can receive immediate feedback from these models, which are capable of analyzing spoken replies at any scale. We want to construct an automated method for assessing english exam scores by utilizing deep learning techniques. Our mission is to guarantee impartial and equitable assessment while enhancing the general effectiveness and dependability of tests of English proficiency.

II. Literature Review

There is a log of significant attention in recent years for the automation of language assessments. This is due to the advancement in NLP and ML techniques. Many researchers had been exploring a lot of different and various approaches to automate language proficiency assessments. The aim of these systems is to use deep learning models and streamline the assessment process to improve grading accuracy. There are many better models in deep learning like RNNs and CNNs also called recurrent neural networks and Convolutional neural networks then there are also transformer based models that have emerged as a good choice for automated language assessment tasks. There are a lot of studies that have specifically been focused on the integration of deep learning models and also AI or artificial intelligence for language proficiency assessments this will also include English Proficiency Test. (Gupta et al. 2018) proposed a deep learning approach for automated spoken language assessment, demonstrating the feasibility of using neural networks to evaluate spoken English proficiency. Lu and Xiong (2018) explored the application of deep learning models in automatic scoring of spoken English language proficiency tests and have highlighted the potential in AI-based systems to provide reliable and consistent evaluations. The bidirectional LSTM-RNN can significantly improve the recognition performance of non-native children's speech, which in turn improves automated scoring performance (Wang et al., 2018).

Graves & Jaitly(2014) proposes a novel speech recognition system that directly transcribes audio into text, bypassing the need for an intermediate phonetic representation. It leverages recurrent neural networks and a modified objective function to achieve state-of-the-art word error rate on a standard benchmark. The authors also explore future directions, including applying the system to different speech data and integrating the language model more deeply into the network.

There are still continuous efforts that have been taken directly towards developing a better system that is AI powered and is capable of correctly analyzing spoken responses in a real time environment. This will give the test takers immediate assessment of their test and the results can be displayed immediately. Hence helping the test takers to improve their proficiency and this is possible by utilizing NLP techniques and all the other techniques like sentiment analysis or voice recognition. Automated Language Essay Scoring (AES) systems, leveraging NLP and machine learning, address time-consuming grading issues (Hussein et al., 2019).

III. Objectives of the Study

Automation of Assessment Criteria: There is another key aspect to automate the assessment criteria for grading English text. This will involve defining and then implementing scoring rubrics for evaluating various aspects of the spoken language proficiency which includes pronunciation with fluency, grammar and vocabulary usage. By bringing automation to these criterias. The Assessment Use Argument (AUA) outlines the interconnectedness of test taker performance, assessment records, interpretations of ability, decision-making, and consequences (Bachman & Adrian, 2022).

Provision of Instant Feedback: In addition our study will aim to provide instant feedback to the people that took the test. Instant feedback is very valuable for those learners as it helps them to understand their strengths and their weaknesses, it also helps identify areas of improvement, and will track their progress over a time period. Every evaluated criterion will receive thorough feedback, along with recommendations for development and further practice.

Evaluation of User Experience: In the end our study will be seeking to evaluate the user experience of interaction with the model prediction. Our objective is to ensure that the given feedback provides a clean and seamless engagement and great experience for test takers, educators, and other stakeholders involved in language assessment. The applied linguistics field has explored language assessment literacy, examining its definition and stakeholder variations (Giraldo, 2018).

IV. Data Collection

Our primary dataset for this study has a collection of English essays collected from various sources, including speaking test samples, educational institutions, online forums, and language learning platforms. A wide range of topics are covered by these essays in our dataset. This provides a diverse set of linguistic data for the model training and evaluation. Each and every essay is accompanied by the corresponding score or grade, assigned by human evaluators based on predefined assessment criteria. The dataset is then broken down into three parts the training file, the test file and the val file the dataset that we collected is in the form of a tsv or also called as tab separated values where every tab has one essay and then it has the grades that were given by a human evaluator.

A. Dataset

ASAP-AES Dataset: Automated Student Assessment Prize - Automated Essay Scoring dataset from OSP which is a free and open-source for research and collaboration. This dataset is a collection of essays written by students along with corresponding scores assigned by human raters. Dataset contains 28 features in total. Key features of the dataset are as follows:

1. essayid: Unique identifier for each essay
2. topic: Topic Identifier
3. essay: Text of the essay
4. rater1domain1, rater2domain1, rater3domain1: Scores assigned by human raters for domain 1.
5. rater1domain2, rater2domain2: Scores assigned by human raters for domain 2.
6. targetscore: Target score based on the essay topic.

Reference: https://osf.io/9fdrw/

V. Exploratory Data Analysis (EDA) and Hypotheses for the Study

Exploratory Data Analysis, also called EDA, plays an important role in understanding all the possible characteristics of the dataset and then formulating hypotheses for any further investigation. In this section, we show how we tried to use EDA into various aspects of the dataset to gain insights into the distribution of essays, word counts, and scores across different topics.

A. Exploratory Data Analysis

Our dataset contains essays that are categorized into eight different topics corresponding to various prompts in the english proficiency test. Initially, we looked at essay counts for each topic and the results showed that there were differences in the total number of essays for each topic. In contrast to the previous themes, Topic 8 seems to feature the fewest essays, which might provide difficulties for model evaluation and training.

Language Correction and Tokenization: We use packages like TextBlob and LanguageTool ('en-US') to correct the grammatical mistakes and increase the language quality for training the model. Next, a cutting-edge natural language

processing (NLP) library called SpaCy was used to perform tokenization. This allowed text to be segmented into meaningful units while removing stop words and punctuation artifacts. We made use of pre-trained English language processing models, such as ”en_core_web_sm,” ”en_core_web_md,” and ”en_core_web_lg,” to enhance our comprehension of the text's linguistic features. With the help of these models, we were able to extract a wide range of linguistic features, with a particular focus on nouns: sentences, named entities recognition (NER), tokens, lemmatized tokens, and parts of speech (POS) tags. On further exploration which involved analyzing the distribution of token counts in essays across different topics. Plotting histograms allowed for the visualization of the word count distribution and revealed information about the average essay length for each subject. We also noted that variances in the word count distributions, suggesting that the topic prompts may have an impact on how long responses take.
Feature Engineering: We also derived extra features that are essential in modeling neural networks. Using SpaCy's POS tagging capabilities, we specifically computed features like token count, capitalization patterns, verb frequency, and adverb frequency, as well as the frequency of commas, questions, exclamation marks, and quotations.
Topic Modeling: We used a probabilistic generative model called Latent Dirichlet Allocation (LDA) to try and find latent topical structures in the text corpus. First, we created a document-term matrix by converting the essays into a numerical representation using CountVectorizer. The LDA model was then fitted to this matrix, allowing the probability distribution of each word associated with the corresponding topics to be estimated. (Fig. 1.) Explain how the essays can be related to respective topic names after performing topic modeling using LDA.

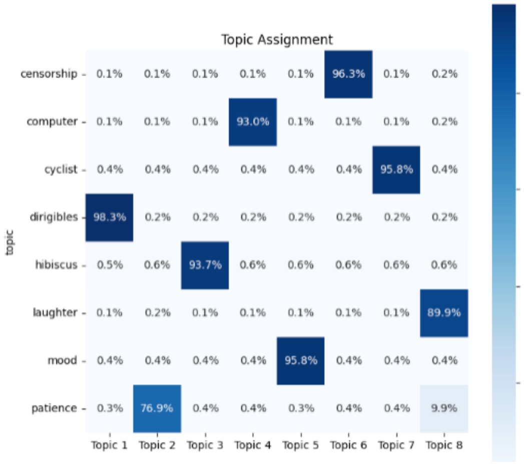


Fig. 1. Topic Modeling

B. Hypotheses for the Study

On the basis of the findings from the EDA that we did on our dataset we have formulated hypothesis regarding all the effectiveness of this AI model in assessing various criteria Effect of Essay Length on Score: we have a hypothesis that if the essays are longer then our model will give higher scores due to the potential that the user has provided a more detailed and coherent response. However, this relationship may vary from topic to topic, with some topics prioritizing length over verbosity. Topic-Specific Assessment Challenges: We have also anticipated that there are certain topics that may present unique challenges in automated assessment, such as nuanced language usage, topic relevance, and argument coherence. Consequently, The essay subject prompt and related assessment criteria may have a huge impact on how well the model will grade submissions. Consistency in Score Allocation: We expect the model to demonstrate consistency in score allocation across different essays with similar linguistic features and thematic content. However, due to the subjective nature of language evaluation and the complexity of human language interpretation, score differences may arise.

VI. Deep Learning Model Architecture

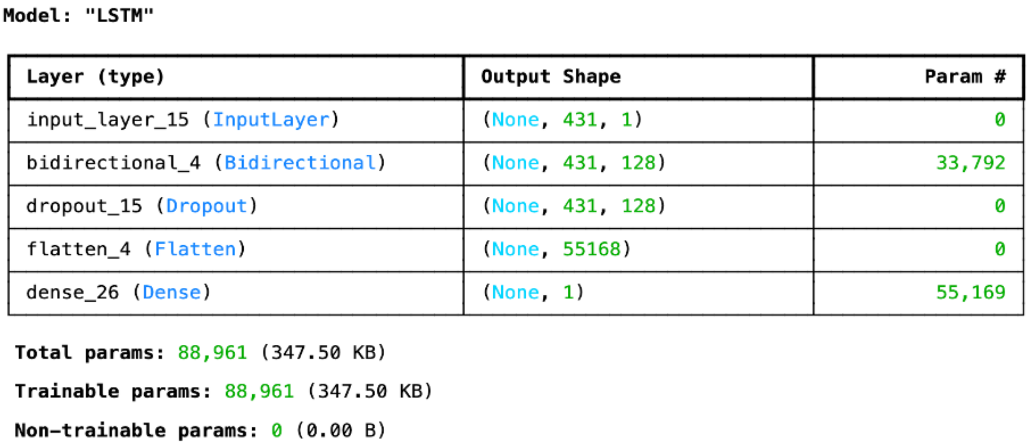
The architecture of our neural network used is a Bidirectional LSTM with a few hidden layers. The model consists of an input layer, a hidden layer with 64 neurons, a dropout layer to prevent overfitting, a flatten layer to convert multi-dimensional array to single-dimensional array, and an output layer. The output layer has a single neuron for regression, which will be used for predicting the essay scores. (Fig. 2.) Displays the model summary which is being used for this project.

A. Data Preprocessing

Preprocessing of the data is a crucial part in preparing the raw essay data for input into the neural network models. It involves a lot of sub-steps, including:
The loading of raw essay data from a tab separated values or TSV file with that associated metadata such as topic or essay set and target scores. We need to apply spelling and grammar corrections to all the essays using various tools like LanguageTool to improve the quality of our text data. We need to Generate word embeddings using Gensim's Word2Vec model to represent each word in the essays as dense vectors in a continuous vector space. In order to make model training and evaluation easier, the goal scores should be scaled to a common range.

B. Model Selection

Selection of an appropriate model deep learning neural network with the right architecture is required for automatic essay scoring. In the provided script, two main approaches are explored Individual Model Approach here, we have constructed separate models for each essay topic or we can call that essay set to capture topic-specific patterns and characteristics. Combined Model Approach: Here we have used combined data from all essay topics which will merge into a single model to leverage information from multiple topics simultaneously. Sutskever et al. (2014) propose a sequence-to-sequence learning approach with deep Long Short-Term



```
Model: "LSTM"

┏━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━┳━━━━━━━━━━━━━━━━━━━━━━━━━━━━━┳━━━━━━━━━━━━━━━━━┓
┃ Layer (type)                          ┃ Output Shape                ┃         Param # ┃
┡━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━╇━━━━━━━━━━━━━━━━━━━━━━━━━━━━━╇━━━━━━━━━━━━━━━━━┩
│ input_layer_15 (InputLayer)           │ (None, 431, 1)              │               0 │
├───────────────────────────────────────┼─────────────────────────────┼─────────────────┤
│ bidirectional_4 (Bidirectional)       │ (None, 431, 128)            │          33,792 │
├───────────────────────────────────────┼─────────────────────────────┼─────────────────┤
│ dropout_15 (Dropout)                  │ (None, 431, 128)            │               0 │
├───────────────────────────────────────┼─────────────────────────────┼─────────────────┤
│ flatten_4 (Flatten)                   │ (None, 55168)               │               0 │
├───────────────────────────────────────┼─────────────────────────────┼─────────────────┤
│ dense_26 (Dense)                      │ (None, 1)                   │          55,169 │
└───────────────────────────────────────┴─────────────────────────────┴─────────────────┘

 Total params: 88,961 (347.50 KB)

 Trainable params: 88,961 (347.50 KB)

 Non-trainable params: 0 (0.00 B)
```

Fig. 2. Model Architecture

Memory (LSTM) networks. Their model achieves competitive translation performance on English-to-French tasks, even surpassing a state-of-the-art statistical machine translation system. The authors also demonstrate the LSTM's ability to learn meaningful phrase and sentence representations. Bidirectional LSTM, it is a neural network which can predict depending on data sequences,this is used to handle the data. Dropout is used to stop the model from learning the training set, preventing overfitting. An input in three dimensions is generated by the LSTM and converted into a one-dimensional array for additional examination. The projections are shown in the output layer. The model's performance is enhanced by the Adam optimizer, and the model's prediction accuracy is evaluated by the Kappa score. In light of the possibility that some matches are accidental, the Kappa score assesses the dependability of the model.

Sak et al. (2014) propose deep Long Short-Term Memory (LSTM) recurrent neural networks for large-scale acoustic modeling in speech recognition. Their approach outperforms conventional RNNs and deep feedforward networks, achieving state-of-the-art performance. Additionally, the paper introduces a distributed training method using asynchronous stochastic gradient descent, enabling efficient training of LSTMs on large machine clusters.

Hochreiter & Schmidhuber (1997) introduce Long Short-Term Memory (LSTM), a novel recurrent neural network architecture that tackles the vanishing gradient problem hindering traditional RNNs. By incorporating constant error flow through special units within the network, LSTM facilitates faster learning and enables handling of longer time lags compared to previous architectures.

C. *Model Optimization Techniques*

Our neural network model will always need to be optimized and the performance must be enhanced by various optimization techniques they are:

- Dropout Regularization: we have added a dropout layer in the MLP architecture so that it would prevent any overfitting problem and randomly drop fraction of inputs during training.
- Hyperparameter Tuning: there are different parameters like learning rate or batch size or number of hidden units. All these parameters are tuned using various techniques like grid search or random search to optimize our model performance.
- Cross-Validation: Cross-validation is used to evaluate the generalization performance of the model and address data variability-related problems. This entails dividing the dataset into several subgroups and repeatedly training and assessing the model on various subsets.
- Evaluation Metrics: there are various metrics like the MSE or mean squared error matrix, MAE also called Mean absolute error which was used for the quantification of the performance of our model and comparing it with the baseline approaches.
- Visualization and Analysis: plotting various histograms can be an example of data visualization. This technique is used to evaluate the robustness of our model and to also uncover any hidden potential biases or any other anomalies and then provide insights to the distribution of true and projected scores.

There are further more areas that need to be explored and improved this include: There is a need for alternate exploration of neural network architectures like RNNs and CNNs also called Recurrent neural networks and convolutional neural networks. There are newer deep learning models like Latent Dirichlet Allocation (LDA), which is used for capturing sequential dependencies and information that is contextual in essays. Hence by checking for alternate feature engineering techniques and adding external knowledge we can enrich the input representations and enhance model performance.

VII. Data Visualization and Results Report

A. Graphical Representation

For graphical representation we will be using line charts and bar graphs that will help us visualize the performance of metrics of our automated speech scoring model over the time and across various experimental conditions. In this project we have plotted various metrics for different parameters like for accuracy and prediction so that we get a comprehensive overview of the models performance across various evaluation criteria. (Fig. 4.) Depicts the MSE loss of training data vs validation data, which settles down after the first few epochs and fits properly.

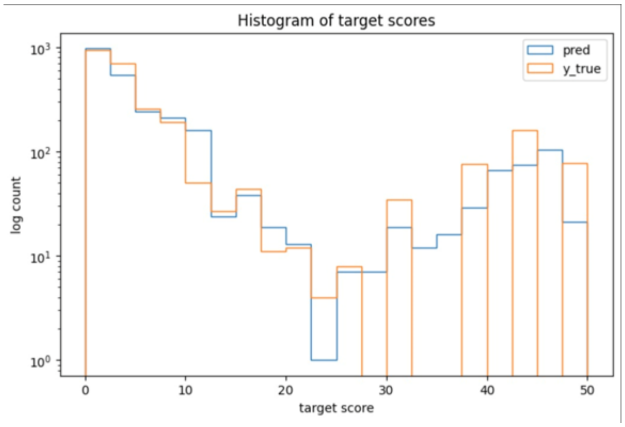


Fig. 3. Histogram of Target Score and Actual Score

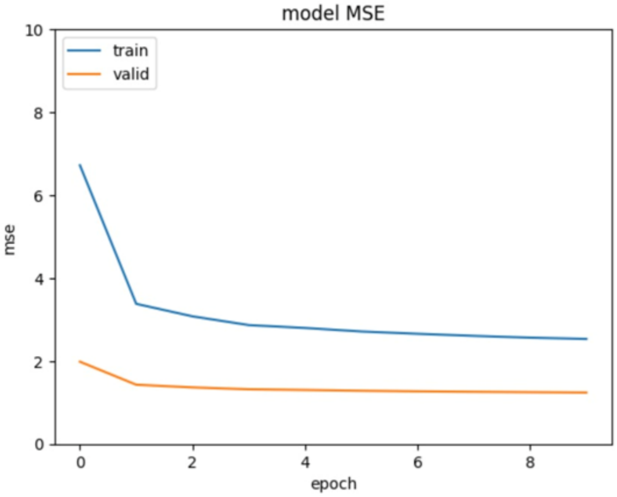


Fig. 4. Model MSE Loss

B. Comparison with Human Evaluations

We will be comparing our model with the human judgements to check how weel the model will align with a humans evaluation We will also be able to show the amount of agreement we can get between an automated score and that of a human assessment by kappa scores.

C. Implications for Automated Speech Scoring

Provide insights and interpretations based on the visualized results, discussing the implications for automated speech scoring systems in educational or assessment contexts. Identify strengths and weaknesses of the model based on the visualized metrics and analyze potential factors influencing performance, such as dataset characteristics, grammatical complexity, or test taker variability.

VIII. Conclusion

From this project in the field of automating language assessment, we were not able to show any significant improvements using the latest models. But, we were able to achieve 70.5% accuracy with our model close enough to the previous research. To combine and conclude we addressed the very need for an efficient and very reliable evaluation matrix in this field of education and professional settings. We have used a very rigorous and comprehensive approach which used data processing, model selection and optimisation to increase the accuracy of the system. We have also plotted metrics and visualized them so that we can see any insights to our data and increase the performance to our advantage. Although we acknowledge the inherent limitations of our approach, including the need for additional research into multimodal inputs and model refinements, our results highlight the revolutionary potential of automated language assessment in advancing evaluation processes' efficiency, fairness, and dependability. In the end, our research advances the conversation about automated language assessment and opens the door to more equal and accessible evaluation procedures in a variety of linguistic situations and fields.

IX. Bibliography

References

[1] Hussein, M. A., Hassan, H., & Nassef, M. (2019). Automated language essay scoring systems: A literature review. PeerJ Computer Science, 5, e208.

[2] Bachman, L., & Adrian, P. (2022). Language assessment in practice: Developing language assessments and justifying their use in the real world. Oxford University Press.

[3] Giraldo, F. (2018). Language assessment literacy: Implications for language teachers. Profile Issues in Teachers Professional Development, 20(1), 179-195.

[4] Wang, Y., Wong, J. H. M., Gales, M. J., Knill, K. M., & Ragni, A. (2018, December). Sequence teacher-student training of acoustic models for automatic free speaking language assessment. In 2018 IEEE Spoken Language Technology Workshop (SLT) (pp. 994-1000). IEEE.

[5] Gupta, P., Andrassy, B., & Schütze, H. (2018). Replicated siamese LSTM in ticketing system for similarity learning and retrieval in asymmetric texts. arXiv preprint arXiv:1807.02854.

[6] Graves, A., & Jaitly, N. (2014, June). Towards end-to-end speech recognition with recurrent neural networks. In International conference on machine learning (pp. 1764-1772). PMLR.

[7] Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. Advances in neural information processing systems, 27.

[8] Sak, H., Senior, A. W., & Beaufays, F. (2014). Long short-term memory recurrent neural network architectures for large scale acoustic modeling.

[9] Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural computation, 9(8), 1735-1780.