Automation of Answer Script Evaluation

Ganesh Prasad Tamminedi[1]; Sri Abhirama Maganti[2]; Tarush Chandra[3]
G.I.T.A.M. University- Hyderabad 502329, Rudraram, Telangana, India

Abstract:-The goal of this study, "Automation of Answer Scripts Evaluation," is to create an end-to-end automated process that can quickly and fairly evaluate answer scripts and grade students. Optical Character Recognition (OCR), Artificial Intelligence (AI), Machine Learning (ML), Natural Language Processing (NLP) are brought together to build a workflow for automating this tedious, time taking, subjective activity. The paper discusses failures and successes of various models applied in our endeavour.

Keywords:- *OCR Model, Bert Model, NLP, GPT Model, Optimization, Cosine Similarity, Vectorization, Rubric Model, Evaluating Model, Datasets, Ensemble, Majority Voting, Gradient Descent.*

I. INTRODUCTION

In the world of education, the persistent struggle to understand messy handwriting and meet strict grading deadlines remains a challenge. Traditional grading systems are currently facing problems like giving subjective grades, dealing with different student answers, and struggling with many evaluations. These issues, involving personal opinions and difficulties in handling a lot of papers, highlight the need for new ways to make grading simpler and better. Essentially, we need to find creative solutions using AI, NLP, ML.

The first hurdle is ability to translate handwritten answers into computer analysable text. OCR models are explored for this purpose. Current OCR models [Ref] have given limited success in our experiments. Once the manageable text is available, interpreting the answer and evaluating has shown even more difficulties. Similarity, BERT and GPT models of AI have been tried to find the feasibility of automated evaluation.

➢ *What are we Expecting the Proposed Process to Accomplish? The Process Needs to*

- Reduce the time and effort required for grading.
- Apply predefined criteria consistently, minimizing the potential for subjective grading biases.
- Handle a large volume of answer scripts for specific subjects
- Utilize predefined algorithms and models to evaluate answers objectively.

➢ *Issues Identified in Our Exploration*

- OCR model struggles with accurately deciphering unclear or unconventional handwriting.
- Struggle with subjective questions that require nuanced understanding and context.
- Handling diverse ways in which students express their answers.
- Evaluating open-ended answers those requiring creative thinking.
- Models like Similarity, BERT and GPT, face limitations due to a lack of real-world datasets.
- To accommodate various subjects and exams without compromising accuracy.

II. LITERATURE REVIEW

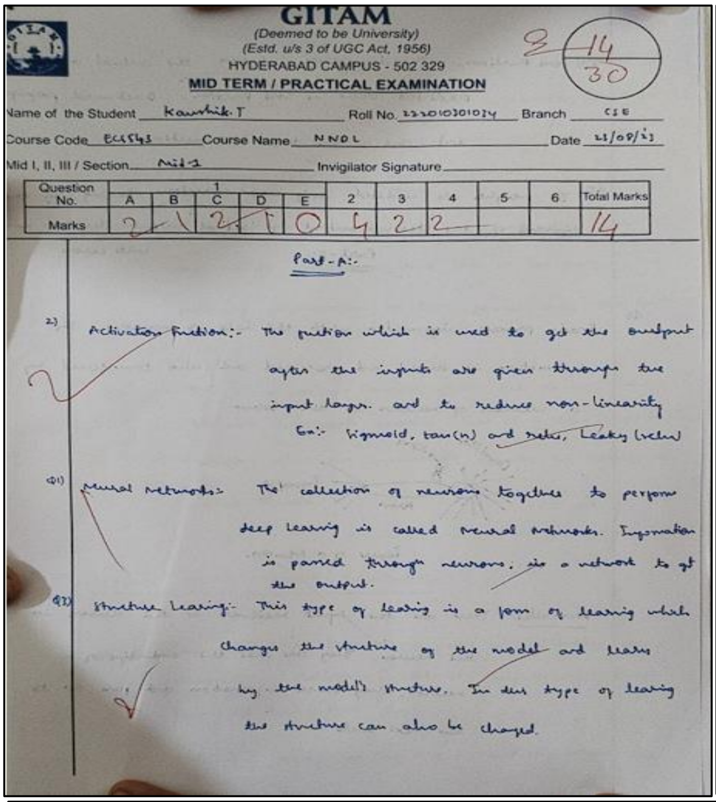
Previous models for automating the evaluation of answer scripts, such as those by Ravikumar et al., del Gobbo et al., and Rahman and Siddiqui, have shown notable limitations in terms of time complexity, correctness in assessing student responses, and handling the variability of subjective content in educational contexts. While efforts have been made to create frameworks and apply machine learning and NLP-based techniques, these models struggled to replicate the nuanced judgment of human graders, particularly in processing diverse handwriting styles and contextual understanding. Consequently, this paper aims to address these shortcomings by developing a new model from scratch that focuses on enhancing the speed, accuracy, and adaptability of automated grading systems, building upon the insights and limitations highlighted in prior research.

The following sections discuss methodologies tried, results and analysis.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24OCT205

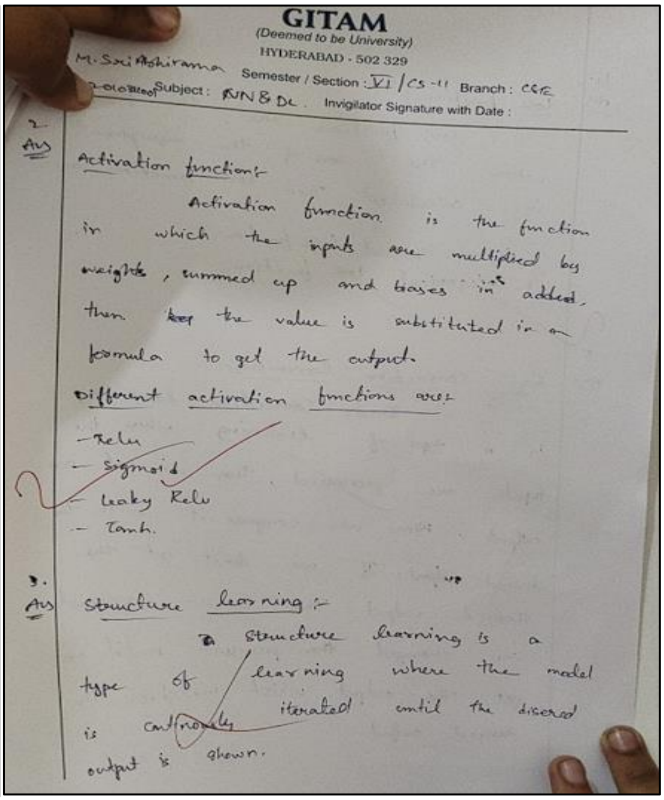
III. METHODOLOGY

Data sets have been created by us from the real-world examinations.


GITAM
(Deemed to be University)
(Estd. u/s 3 of UGC Act, 1956)
HYDERABAD CAMPUS - 502 329
MID TERM / PRACTICAL EXAMINATION
Name of the Student Kaushik. T Roll No. 2210103010?4 Branch CSE
Course Code ECS543 Course Name NNDL Date 21/09/23
Mid I, II, III / Section Mid-1 Invigilator Signature

Question No.	A	B	C	D	E	2	3	4	5	6	Total Marks
Marks	2	1	2	1	0	4	2	2			14

Part-A:-
Activation Function:- The function which is used to get the output after the inputs are given through the input layer and to reduce non-linearity Ex:- Sigmoid, tan(h) and Relu, Leaky (relu)
Neural Networks:- The collection of neurons together to perform deep learning is called neural networks. Information is passed through neurons, is a network to get the output.
Structure Learning:- This type of learning is a form of learning which changes the structure of the model and learns by the model's structure. In this type of learning the structure can also be changed.


GITAM
(Deemed to be University)
HYDERABAD - 502 329
M. Sai Abhirama
Semester / Section : VI / CS-11 Branch : CSE
Subject : ANN & DL Invigilator Signature with Date :
Activation function:-
Activation function is the function in which the inputs are multiplied by weights, summed up and biases is added, then the value is substituted in a formula to get the output.
Different activation functions are
– Relu
– Sigmoid
– leaky Relu
– Tanh.
Structure learning:-
Structure learning is a type of learning where the model is continuously iterated until the desired output is shown.

Questions	Answers
What is meant by activation function. list any 3 activation functions?	activation function is used to determine the output of the neuron after summation of weights and inputs takes place.
What is meant by activation function. list any 3 activation functions?	Activation Function:- The function which is used to get the output after the inputs are given through the input layer and to reduce non-linearity. Ex:- Sigmoid, tan(h)
Define neural networks?	Neural Network:- The collection of neurons together to perform deep learning is called neural networks. Information is passed through neurons is a network to get
Define structure learning?	Structure Learning:- This type of learning is a form of learning which changes the structure of the model and learns by the model's structure. In this type of learning
What is meant by activation function. list any 3 activation functions?	Activation Function:-
Define loss functions. List any two loss functions?	It is the difference between the predicted output and the desired output.
Define loss functions. List any two loss functions?	Loss function:-
Define neural networks?	Neural Networks:-
What is meant by activation function. list any 3 activation functions?	Activation function:-
Define neural networks?	Neural Networks:-
What is meant by activation function. list any 3 activation functions?	Activation Functions:-
Define structure learning?	The learning that takes place with the help of structure of the neural network. The neural network forms on the basis of the structure present. They have input laye

➢ *Challenges Encountered During Real Time Exam Answer Sheet to CSV File Conversion:*

- Insufficient data.
- OCR model scanning accuracy is less to fully identify the text on answer sheets.
- Diagrams cannot be converted to CSV.
- Illegible handwriting from students.
- All the best OCR models are pay to use, huge number of resources are consumed to train a new model.

Starting with handwritten to text that can be analysed with computers, the following OCR models have been tried.

A. OCR Model

➢ *OCR Model 1*

We made two of our own OCR models - one from examples on GitHub and the other following instructions from YouTube. Unfortunately, both the models only worked well with the specific data we provided.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24OCT205

First was the model was made on the instructions given by PythonLessons on YouTube [https://www.youtube.com/watch?v=WhRC31SlXzA&t=5s] . It was trained on words from IAM datasets (sample images in fig- in section 3) which consisted of words, sentences and forms. The model used TensorFlow with different layers consisting of CNN and LSTM with loss functions such as CTC. The model despite showing high accuracy for the input from the dataset it is trained on, gives output which doesn't match with the expected output for an input outside of the dataset. So, the model was discontinued.

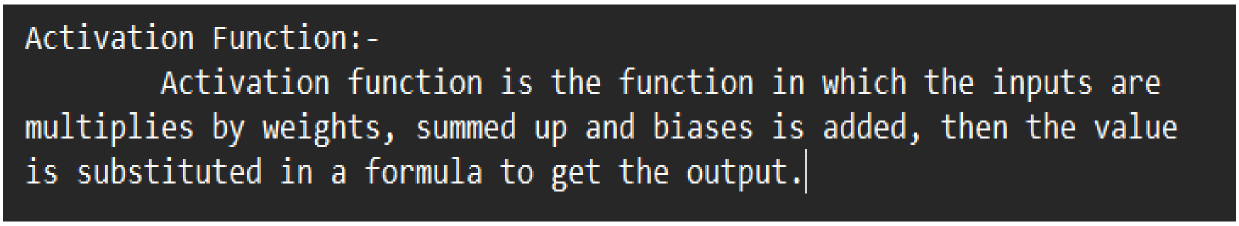
➢ *OCR Model 2*

The model which was built on idea taken from GitHub was trained on EMNIST dataset which consists of alphabets using sequential model and layers such as Conv2D, MaxPooling2D, Dense, Flatten, Dropout with metrics such as confusion matrix, classification report. Optimizers such as Adam, SGD were also used. This model also was giving output that doesn't match with the expected output for an input outside of the dataset it is trained on. This model was suspended due to the significant demand of resources and started experimenting with free and trail versions of paid OCR models.

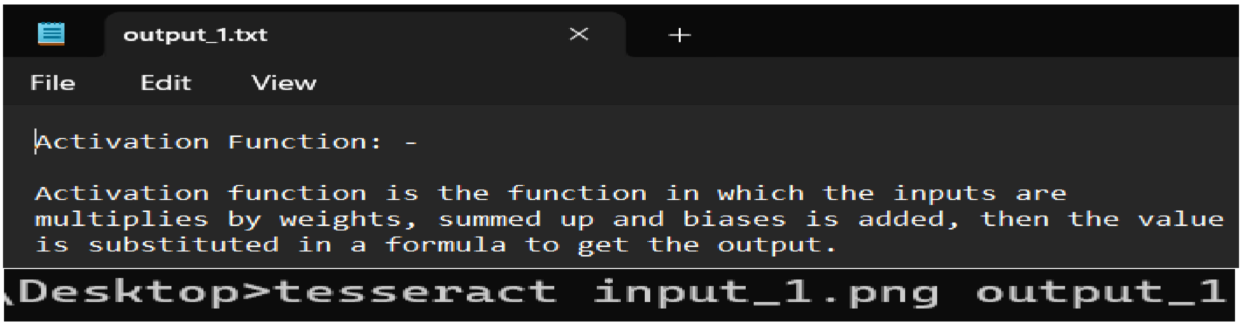
➢ *Open-Source OCR Models*

We experimented with OCR models like Tesseract open-source OCR model[https://tesseract-ocr.github.io/tessdoc/Installation.html], google lens[https://lens.google/] which are readily available. However, these models were limited to recognizing computer-generated text and struggled with handwritten text.

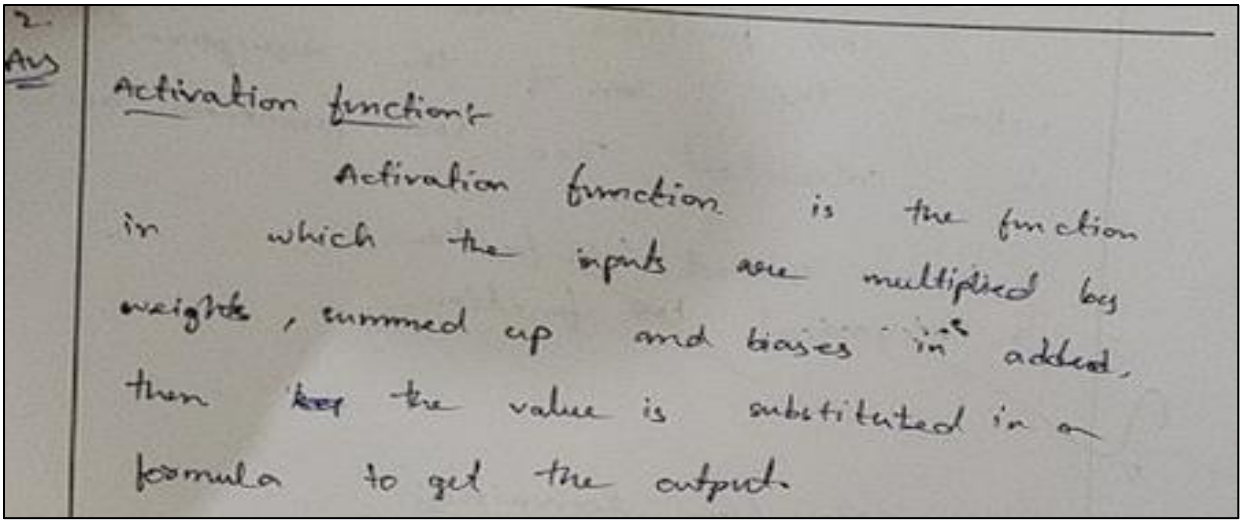
- *Tesseract Input_1.png (Computer Generated Text): -*



- *Tesseract Output_1.txt (Output for Computer Generated Text): -*

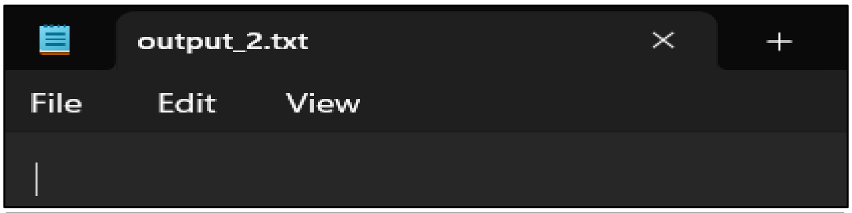


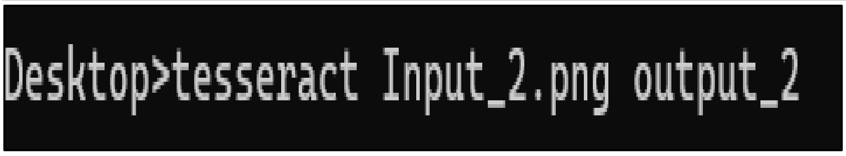
- *Tesseract Input_2.png (Handwritten Text): -*



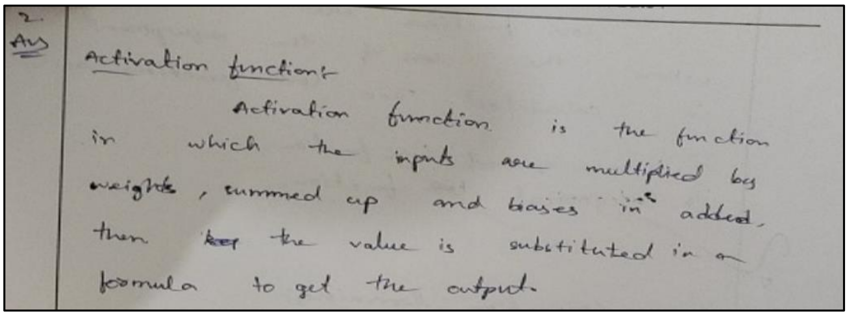
ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24OCT205

- *Tesseract Output_2.txt (Output for Handwritten Text): -*

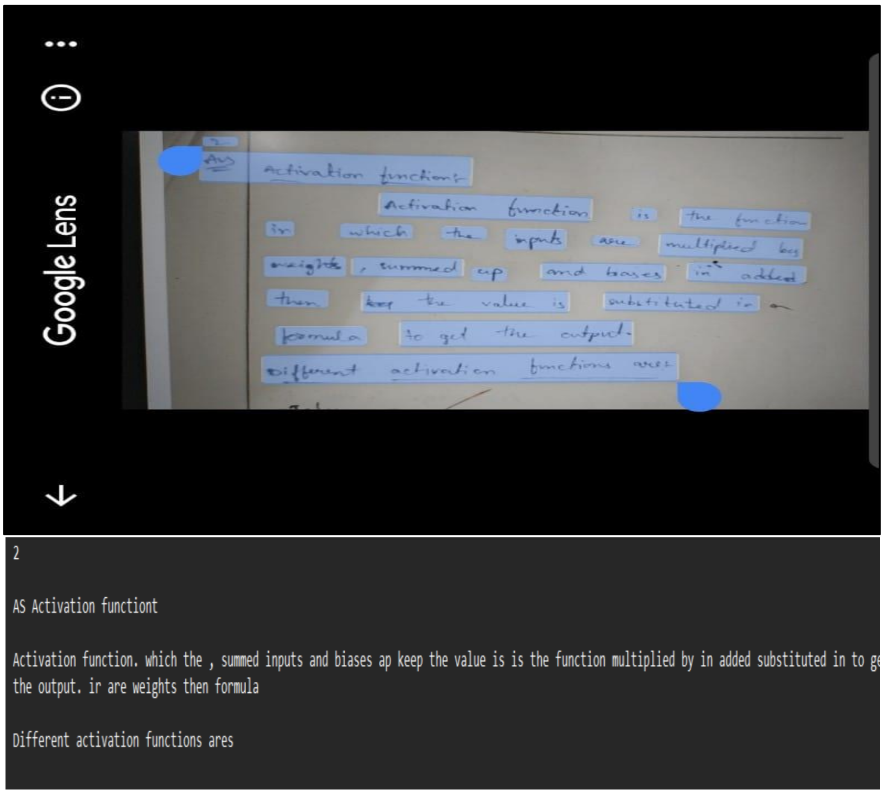




- *Google Lens Input_1.png:*



- *Googlee Lens Output_1.txt(Text Reading from Image):-*



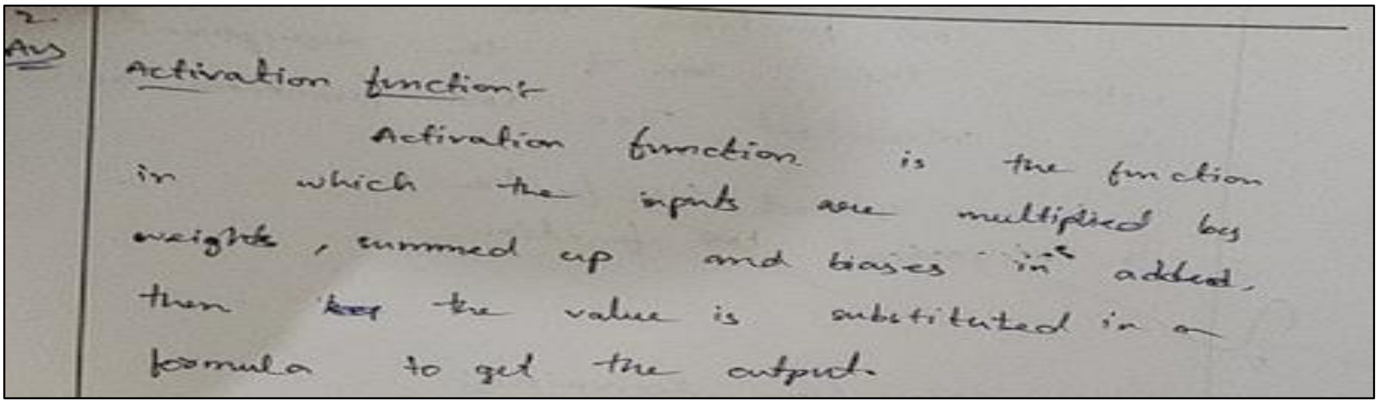
➢ *Models Offered by Entrepreneurial Companies*

For a more advanced approach, we explored paid OCR models such as Google cloud vision, Nanonets and Vision Studio-Azure. While these models performed better, especially with handwritten text, their consistency varied across different styles of handwriting.

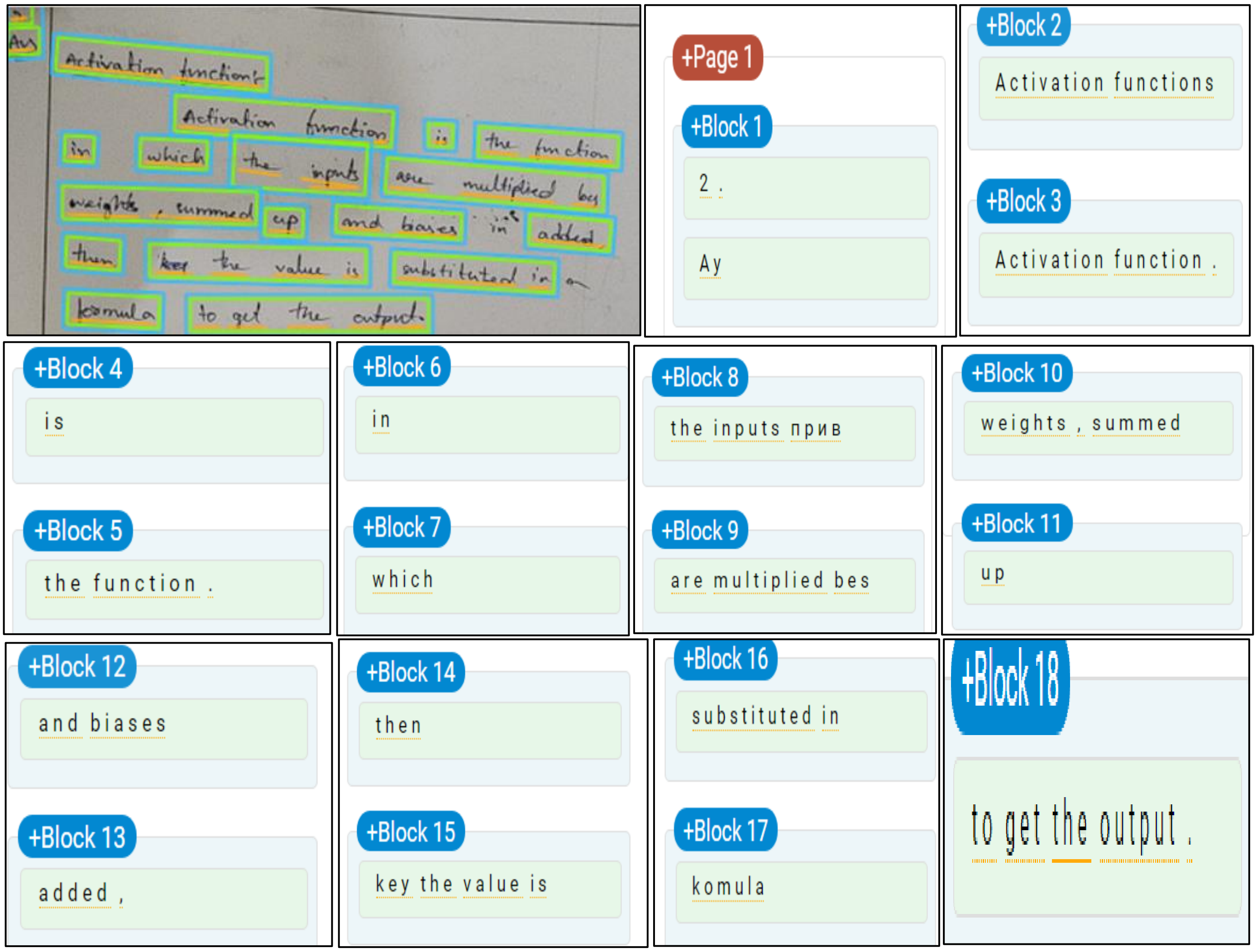
ISSN No:-2456-2165

https://doi.org/10.38124/ijisrt/IJISRT24OCT205

- *Google cloud vision input_1.jpg:*



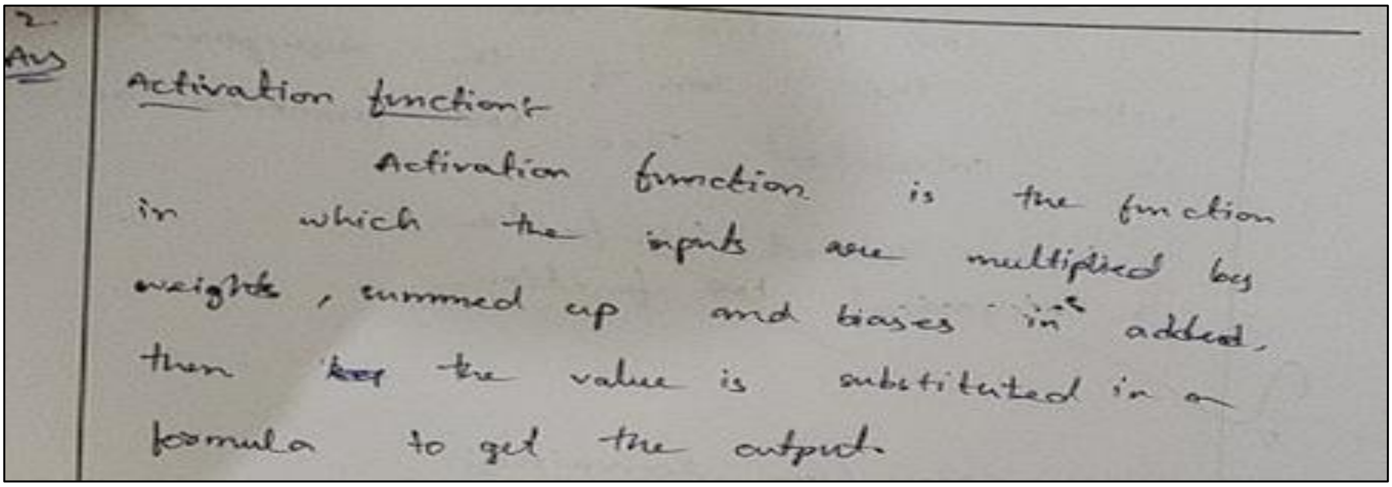
- *Google cloud vision output_1(reading text from handwritten document):*



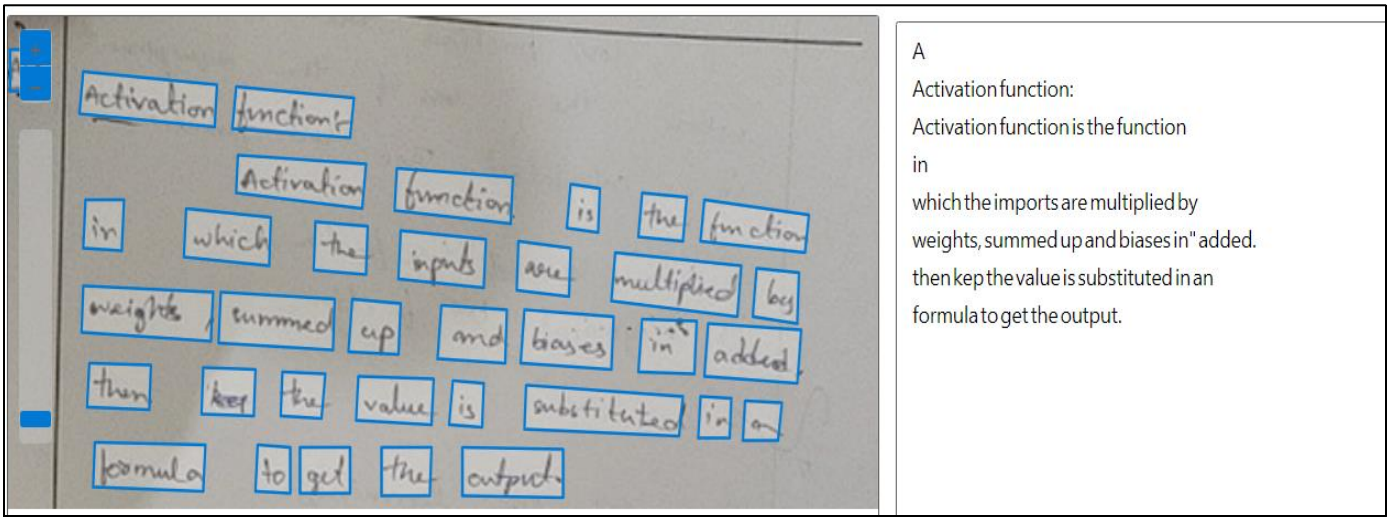
Using the output of Vision Studio azure’s OCR model, we tried the interpretation and analysis with Similarity models, BERT models, GPT models. Discussion on these models follows below.

Volume 9, Issue 10, October – 2024 International Journal of Innovative Science and Research Technology
ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24OCT205

- *Vision Studio Azure Input:*



- *Vision Studio Azure Output (Text Reading from Images):*



As you can see it is accurate, we decided to use outputs of this OCR for developing the NLP models.

B. Similarity Model

document1 = "Technology is rapidly advancing, and this progress is undoubtedly affecting traditional values. I believe that in today's technological era, traditional values are likely to fade away. To begin with, there are several reasons why these age-old values are diminishing in the modern world. Firstly, in our fast-paced society, mobile phones have become everyone's companion for staying connected with family and friends. In contrast, in the past, people used to send letters and wait in long lines for a single telephone call using STD and ISD services. The evolution of communication methods highlights that traditional practices hold little value today. Secondly, technology has revolutionized the realm of fashion. People used to engage in manual activities like knitting, stitching, and designing, but now machines have simplified every task. As an illustration, fashion design students now swiftly assess color compatibility using advanced software instead of manually portraying themselves as models. Furthermore, the advent of refrigerators has drastically reduced the need for traditional water pitchers. In conclusion, as modern technology continues to evolve, traditional methods that are time-consuming will struggle to keep up with the latest trends. Thus, preserving these methods becomes futile and a waste of time."

query_doc = "Technology is flourishing by leaps and bounds and undoubtedly this advancement is taking a toll on the traditional values. I do believe that in this technological world traditional values are bound to disappear. To initiate with, there are many reasons why these conventional values have no existence in this modern world. First, in this fast-paced world, everyone is assisted with the mobile phone to stay connected with their family and friends. However, in olden times people used to send letters and stand in the long queues on S.T.D and I.S.D just for a maturity of one call. This advancement in the modes of communication has proved that traditional skills are worth for nothing. Secondly, technology has transformed the world of fashion. Earlier people used to do knitting, stitching, and designing manually but now machines have made every task easier and comfortable. To substantiate my view, many of fashion designing students

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24OCT205

were seen portraying pictures of model themselves for checking the compatibility of colors but at present, this work is done in seconds on multifarious advanced software. Apart from this, hardly anyone is seen purchasing pitcher for cool water because of the invention of refrigerators. To conclude, an evolution of modern technology is an ongoing process, so, the time-consuming traditional methods will not be able to maintain their pace with these latest trends. hence, it is useless and wastage to time to preserve them."

Document1 and query_doc are the inputs given to every model below.

❖ *Similarity Checker:*

➢ *Model 1.1: Word2Vec Basic Implementation*

- Tokenization: Splits documents into words.
- Word2Vec Embeddings: Converts words into vectors capturing meaning.
- Cosine Similarity: Measures similarity based on vector angles. This is used to compare the documents

✓ Output:

```
Similarity Score = 0.9180
------------------------
Query Document:
Technology is flourishing by leaps and bounds and undoubtedly this advancement is taking a toll on the traditional values.
------------------------
Corpus Document:
Technology is rapidly advancing, and this progress is undoubtedly affecting traditional values. I believe that in today's
------------------------
```

➢ *Model 1.2: Preprocessing Enhancement*

- Stop Word Removal: Eliminates common words.
- Lemmatization: Reduces words to base form.
- Enhanced Preprocessing: Combines stop word removal and lemmatization.

✓ Output:

```
Similarity Score = 0.6245
------------------------
Query Document:
['technology', 'flourishing', 'leap', 'bound', 'undoubtedly', 'advancement', 'taking', '
------------------------
Corpus Document:
['technology', 'rapidly', 'advancing', 'progress', 'undoubtedly', 'affecting', 'traditio
------------------------
```

➢ *Model 1.3: Stemming Integration*

- Stemming: Reduces words to root forms.
- Stem-Based Preprocessing: Integrates stemming with tokenization and stop word removal.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24OCT205

✓ Output:

```
Similarity Score = 0.6565
-------------------------
Query Document:
['technolog', 'flourish', 'leap', 'bound', 'undoubtedli', 'thi', 'advanc', 'tak
-------------------------
Corpus Document:
['technolog', 'rapidli', 'advanc', 'thi', 'progress', 'undoubtedli', 'affect',
-------------------------
```

➢ *Model 1.4: Dual Similarity Metrics*

- Jaccard Similarity: Measures token overlap.
- Dual Metrics: Provides Jaccard and cosine similarity scores.

✓ Output:

```
Jaccard Similarity Score: 0.2812
-------------------------
Query Document:
['technology', 'flourishing', 'leap', 'bound', 'undoubtedly', 'advancement', 'taking', 'toll', 'traditional', 'value', 'believe', 'technological', 'world', 'traditional', 'value', 'bound', 'initiate', 'many', 'reason', 'conventio
-------------------------
Corpus Document:
['technology', 'rapidly', 'advancing', 'progress', 'undoubtedly', 'affecting', 'traditional', 'value', 'believe', 'today', 'technological', 'era', 'traditional', 'value', 'likely', 'fade', 'begin', 'several', 'reason', 'value', '
-------------------------
```

➢ *Model 1.5: TF-IDF Vectorization*

- TF-IDF: Highlights term importance. Hence, I gave multiple best answers to find the best matching model answer for all those answers
- Cosine Similarity with TF-IDF: Considers weighted term importance.
- The best matching model answer is used to improve accuracy again

✓ Output:

```
Tokens before removing stopwords: ['technology', 'is', 'flourishing', 'by', 'leaps', 'and', 'bounds', 'and', 'undoubtedly', 'this', 'advancement', 'is', 'taking', 'a', 'toll'
Filtered tokens: ['technology', 'flourishing', 'leaps', 'bounds', 'undoubtedly', 'advancement', 'taking', 'toll', 'traditional', 'values', 'believe', 'technological', 'world'
Preprocessed text: technology flourishing leaps bounds undoubtedly advancement taking toll traditional values believe technological world traditional values bound initiate ma
Tokens before removing stopwords: ['technology', 'is', 'flourishing', 'by', 'leaps', 'and', 'bounds', 'and', 'undoubtedly', 'this', 'advancement', 'is', 'taking', 'a', 'toll'
Filtered tokens: ['technology', 'flourishing', 'leaps', 'bounds', 'undoubtedly', 'advancement', 'taking', 'toll', 'traditional', 'values', 'believe', 'technological', 'world'
Preprocessed text: technology flourishing leaps bounds undoubtedly advancement taking toll traditional values believe technological world traditional values bound initiate ma
Tokens before removing stopwords: ['technology', 'is', 'rapidly', 'advancing', ',', 'and', 'this', 'progress', 'is', 'undoubtedly', 'affecting', 'traditional', 'values', '.',
Filtered tokens: ['technology', 'rapidly', 'advancing', 'progress', 'undoubtedly', 'affecting', 'traditional', 'values', 'believe', 'today', 'technological', 'era', 'traditio
Preprocessed text: technology rapidly advancing progress undoubtedly affecting traditional values believe today technological era traditional values likely fade begin several
Best matching model answer: Technology is flourishing by leaps and bounds and undoubtedly this advancement is taking a toll on the traditional values. I do believe that in th
```

➢ *Model 1.6: Simple Text Similarity*

- Stem-Based Preprocessing: Simplifies token variations based on stem.
- Cosine Similarity: Measures similarity based on document vectors.

✓ Output:

```
Student's answer is incorrect.
0.3157894736842105 Your answer is not quite there.
```

➢ *Model 1.7: Detailed Approach with Word2Vec*

- Preprocessing: Converts text to lowercase, tokenizes, removes stop words, and stems.
- Word Embedding Generation: Uses Word2Vec for semantic understanding.
- Document Vectorization: Computes average vector for document representation.
- Similarity Computation: Calculates cosine similarity between document vectors.

✓ Output:

```
similarity_score:0.8846630454063416
Student's answer is correct.
Great job! Your answer is very similar to the model answer.
```

➢ *Model 1.8: Jaccard Similarity with N-Grams*

- Preprocessing: Converts text to lowercase, tokenizes, and removes stop words.
- N-gram Generation: Captures sequential patterns for similarity assessment.
- Similarity Computation: Calculates Jaccard similarity based on n-gram overlap.

✓ *Output:*

```
0.09454545454545454
Student's answer is not very similar to the model answer.
```

➢ *Model 1.9: Word2Vec with Cosine Similarity*

- Preprocessing: Converts text to lowercase, tokenizes, removes stop words, and stems.
- Word Embedding Generation: Uses Word2Vec for semantic understanding.
- Document Vectorization: Computes average vector for document representation.
- Similarity Computation: Calculates cosine similarity between document vectors.

✓ *Output:*

```
Similarity Score: 0.8206197619438171
Student's answer is similar to the model answer.
```

➢ *Analysis:*

After extensive exploration, we concluded that none of the similarity models were suitable for evaluating answer scripts. Despite our efforts in incorporating various NLP methods such as stemming, lemmatization, Word2Vec, N-Gram, different types of vectorization, different types of checking similarities, the results were not feasible for our intended application, Eg: a student writing Hi instead of Hello the meaning same but hello is not in our corpus document with will significantly affect the similarity score, as you can see in Model 1.8. The complexity of evaluating subjective content in answer scripts posed challenges beyond the capabilities of our researched similarity models.

Our exploration led us to BERT models, a discussion follows here.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24OCT205

C. BERT Model

➢ *BERT Model-2.0: BERT Exploration*

In the BERT Exploration phase, we began by understanding how the BERT model, which reads text in both directions (forward and backward), could be applied to recognize text within our OCR system. We were essentially testing if BERT could understand written words in images as it does with plain text.

➢ *BERT Model-2.1: BERT Ensemble Approach*

To enhance performance, we tried combining multiple BERT models into an ensemble. This involved merging their predictions to create a more robust and accurate OCR system.

✓ Input:

```
# Inference
# Student's answer
student_answer = "Photosynthesis is how plants make food. It takes light energy and turns it into chemical energy."

# Synthetic data generation (Step 1)
sample_answers = [
    "The process of photosynthesis involves the conversion of light energy into chemical energy.",
    "Photosynthesis is a process used by plants to convert light energy into chemical energy.",
    "Through photosynthesis, plants transform light energy into chemical energy.",
    "Plants convert light energy to chemical energy via photosynthesis."
]
```

✓ Output:

```
Mean Squared Error: 0.11101569687676
Predicted Grade: 2.24780158278144
Feedback:
Your answer needs improvement. Please provide more details and clarity.
```

➢ *BERT Model-2.2: Text Processing and Feature Engineering*

For the Text Processing and Feature Engineering phase, we fine-tuned the way our model reads and understands the text. It's like sharpening the tools before starting the work—cleaning the data and picking out the key parts of the text that would help the model learn better and make more accurate predictions.

✓ Input:

```
# Inference
# Student's answer
student_answer = "Photosynthesis is how plants make food. It takes light energy and turns it into chemical energy."

# Synthetic data generation (Step 1)
sample_answers = [
    "The process of photosynthesis involves the conversion of light energy into chemical energy.",
    "Photosynthesis is a process used by plants to convert light energy into chemical energy.",
    "Through photosynthesis, plants transform light energy into chemical energy.",
    "Plants convert light energy to chemical energy via photosynthesis."
]
```

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24OCT205

✓ Output:

```
Mean Squared Error: 0.12981201075416
Predicted Grade: 2.85758298662654
Feedback:
Your answer needs improvement. Please provide more details and clarity.
```

➢ *BERT Model-2.3: Data Augmentation and Ensemble*

Facing limitations with the amount and variety of text our model had seen, we used data augmentation, specifically back translation (translating text to another language and back to the original). This expanded our model's exposure, making it akin to reading more books to learn more about the world, thus improving its ability to understand and process text.

✓ Input:

```
# Synthetic data generation
sample_answers = [
    "The process of photosynthesis involves the conversion of light energy into chemical energy.",
    "Photosynthesis is a process used by plants to convert light energy into chemical energy.",
    "Through photosynthesis, plants transform light energy into chemical energy.",
    "Plants convert light energy to chemical energy via photosynthesis."
]

grades = [4.5, 4.0, 4.2, 4.6]
answers = []
grades_dataset = []
```

```
# Inference
student_answer = "Photosynthesis is how plants make food. It takes light energy and turns it into chemical energy."
preprocessed_student_answer = preprocess_text(student_answer)
tokenized_student_answer = tokenizer(preprocessed_student_answer, padding=True, truncation=True, return_tensors="pt", max_length=128)
```

✓ Output:

```
Mean Squared Error: 0.05898155755930
Predicted Grade: 2.82190196467519
Feedback:
Your answer needs improvement. Please provide more details and clarity.
```

➢ *BERT Model-2.4: BERT Model Optimization*

During the BERT Model Optimization phase, we adjusted the model's settings like tuning hyper parameters for better performance so that it could learn faster and more effectively from the OCR tasks we gave it. We wanted to ensure the model was running at its best to handle the complex task of reading handwriting from our images.

✓ Input:

```
# Inference
student_answer = "Photosynthesis is how plants make food. It takes light energy and turns it into chemical energy."
```

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24OCT205

```
# Synthetic data generation
sample_answers = [
    "The process of photosynthesis involves the conversion of light energy into chemical energy.",
    "Photosynthesis is a process used by plants to convert light energy into chemical energy.",
    "Through photosynthesis, plants transform light energy into chemical energy.",
    "Plants convert light energy to chemical energy via photosynthesis."
]

grades = [4.5, 4.0, 4.2, 4.6]
answers = []
grades_dataset = []
```

✓ Output:

```
Downloading (…)solve/main/vocab.txt: 100%  232k/232k [00:00<00:00, 2.54MB/s]
Downloading (…)okenizer_config.json: 100%  28.0/28.0 [00:00<00:00, 482B/s]
Downloading (…)lve/main/config.json: 100%  570/570 [00:00<00:00, 13.5kB/s]
Downloading model.safetensors: 100%  440M/440M [00:02<00:00, 159MB/s]
Some weights of BertForSequenceClassification were not initialized from the model checkpoint at bert-base-uncased and are newly initialized: ['classifier.bias', 'classifier.weight']
You should probably TRAIN this model on a down-stream task to be able to use it for predictions and inference.
Some weights of BertForSequenceClassification were not initialized from the model checkpoint at bert-base-uncased and are newly initialized: ['classifier.bias', 'classifier.weight']
You should probably TRAIN this model on a down-stream task to be able to use it for predictions and inference.
/usr/local/lib/python3.10/dist-packages/transformers/optimization.py:411: FutureWarning: This implementation of AdamW is deprecated and will be removed in a future version. Use the PyTorch implementation torch.optim.AdamW instead, or set `no_deprecation_warning=True` to disable this warning
  warnings.warn(
Mean Squared Error: 0.07947183674982145
Predicted Grade: 4.142022132873535
Feedback:
Excellent! Your answer is very well-written and comprehensive.
```

➢ *BERT Model-2.5: Finalizing BERT Model Features*

After trying out various adjustments, we picked the best features that helped our BERT model recognize text most accurately—like choosing the best ingredients for a recipe. We combined text processing, feature engineering, and methods to handle lots of data at once, and used the ensemble approach where all the 'expert' models we created worked together to give their best prediction.

✓ Input: -

```
sample_answers = [
    "Supervised learning requires labeled data, which means each example in the dataset is paired with the correct output. It uses this labeled data to train the model and ma
    "Supervised learning uses labeled data for training, whereas unsupervised learning uses unlabeled data.",
    "In supervised learning, we have an algorithm that learns from labeled data, while unsupervised learning is where the algorithm learns from unlabeled data.",
    "Supervised is when the data is labeled. Unsupervised is when it's not.",
    "Supervised learning is where you train the model by presenting it with input and the correct output, and it learns by comparing its actual output with the correct output
]

grades = [5.0, 3.5, 4.0, 2.5, 4.5]
```

```
student_answer = input("Enter your response: ")
```

```
Enter your response: Unsupervised learning:- No help at all is provided, the model should learn everything on its own, Supervised learning:- Humans oversee and help the model to learn
```

✓ Output:

```
Predicted Grade: 1
Feedback: Good job! You've covered some of the main points.
```

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24OCT205

➢ *BERT Model-2.6: Testing with Custom Dataset*

Finally, we put our BERT model to the test with a custom dataset—essentially a tailored exam, including content that we expected the model to recognize. This was the ultimate test to see if all our fine-tuning paid off, and if our model could indeed understand and process the variety of handwriting styles and texts that students might use in their answers.

✓ Input:

question	perfect_answer	keywords	point_based
Define neural networks?	["neural networks is a network which consists of a Set of neuron\'s connected togetl	artifical emulation, humain brain, interconnection, layers,	0
"What is meant by activation	["Activation function is used to determine the Output of the neuron after getting sum	mathematical funcations, standardization, activation, out	0
Define structure learning?	["Structure Leaving:- This type of learning is a form of leaving which Changes the stru	self-deciding, user burden, accuracy. Weights, biases, up	0
Define loss functions. List any	["Loss function:- It is function that compare the target and predicted ouput Values:\	comparison, error computation, loss calculation, minimiz	1

```
Enter your response: Neural networks are computing systems vaguely inspired by the biological neural networks that constitute animal brains. Such systems 'learn' to perform tasks by considering examples, generally without being programmed with task-specific rules.
```

✓ Output:

```
Question 1: Define neural networks?
Enter your response: Neural networks are computing systems vaguely inspired by the biological neural networks that constitute animal brains. Such systems 'learn' to perform tasks by considering examples, generally without being programmed with task-specific rules.
Feedback: Your answer is out of context or not relevant to the question.
Score: 0
Question 2: "What is meant by activation function?
list any 3 activation functions"
Enter your response: Activation functions are crucial components in neural networks, particularly in artificial neural networks, where they define the output of a node, or 'neuron,' based on its inputs. They help decide whether a neuron should be activated or not, e
Feedback: Your answer is out of context or not relevant to the question.
Score: 0
```

Table 1: Analysis: Real-World Data Limitations

Model	Improvement Over Previous Model	Advantages	Disadvantages
BERT Model-2.1	Introduced ensemble approach, combining multiple BERT models for enhanced performance.	Enhanced performance through ensemble learning.	Increase complexity in model integration
BERT Model-2.2	Implemented text processing and feature engineering to optimize input for BERT, aiming for improved accuracy.	Improved accuracy through optimized input preprocessing.	Require additional preprocessing steps.
BERT Model-2.3	Explored data augmentation techniques like back translation and ensemble methods to diversify dataset and enhance performance.	Enhanced dataset diversity and performance through augmentation.	May introduce noise or bias with augmentation techniques.
BERT Model-2.4	Fine-tuned BERT model parameters and configuration for optimized performance.	Optimized performance through parameter tuning.	Time and resource intensive tuning process.
BERT Model-2.5	Integrated features from text processing, feature engineering, data augmentation, time complexity optimization, and ensemble for a more robust OCR model.	Comprehensive approach for improved OCR model robustness.	Increased complexity and resource requirements.
BERT Model-2.6	Tested the final BERT model with a custom dataset to evaluate its effectiveness in recognizing and processing varied content.	Real-world validation of model effectiveness.	Requires access to diverse and representative datasets

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24OCT205

GPT Model-3.1	Tailored the GPT model specifically for single-answer scripts, aiming to enhance its performance in handling this type of data.	Enhanced performance in handling single answer scripts.	Limited applicability to other types of data.
GPT Model-3.2	Implemented the GPT model using a custom dataset to improve its adaptation to specific research needs.	Improved adaptation to specific research needs.	Dependency on availability and quality of custom dataset.

Despite our efforts, the lack of a sufficiently diverse real-world dataset for training hindered the success of the BERT model. Obtaining more varied and representative data could be crucial for future improvements in OCR model performance.

Further on, GPT models have been explored and the outcomes are discussed below.

D. GPT Model

The GPT (Generative Pre-trained Transformer) models are a series of language processing AI designed by OpenAI that use deep learning to produce human-like text. Here's how each version was utilized:

➢ *GPT Model 3.0*

- **Engine Used**: text-davinci-003
- **Purpose and Application:** This version of the GPT model was employed to leverage its advanced language processing capabilities to enhance our research outputs. The text-davinci-003 engine is known for its ability to understand and generate human-like text, making it suitable for complex language tasks.
- **Special Features:** The model is built on a transformer architecture that prioritizes context and coherence, which allows it to perform a wide range of text-based tasks effectively.

➢ *GPT Model 3.1*

- **Customization:** Tailored specifically for handling single-answer scripts
- **Purpose and Application:** The focus was to refine the GPT model to improve its performance specifically in scenarios where precise, single-answer responses were needed. This adaptation makes it ideal for applications such as automated FAQs, where direct answers are preferred over elaborate text.
- **Special Features:** Enhancements in this model version include better handling of specific query types and improved accuracy in short-answer predictions. Input and Output:

```
Question: Define neural networks?
Answer: Neural Network:- The collection of neurons together to perform deep learning is called neural networks. Information is passed through neurons is a network to get the output.
Evaluation Marks: 0
Grammar Correction: Neural networks are a collection of neurons that work together to perform deep learning. Information is passed through neurons in a network to get the output.
Too many grammatical errors.
```

➢ *GPT Model 3.2*

- **Customization:** Implemented using a custom dataset
- **Purpose and Application:** This iteration of the GPT model was customized with a particular dataset tailored to the specific needs of our research. Using a custom dataset allows the model to better understand and generate text that is more aligned with the thematic elements of the research.
- **Special Features:** The ability to train on specific data enhances the model's relevance to the user's needs, potentially improving both the quality and applicability of its outputs in tailored scenarios.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24OCT205

✓ Input:

A1 fx Questions

	A	B	C
1	Questions	Answers	Grades
2	Define neural networks?	Neural netowrk are the connections of millions or billions of nuron stacked together mimicing like a hum	1
3	What is meant by activation function	activation function is used to determine the output of the neuron after summation of weights and inputs t Three activation functions are: -sigmoid -tanh -Relu	2
4	Define loss functions. List any two lc	Loss function:-	0
5	Define structure learning?	structure learning means to determine shape and structure of the network. It dosen't require weights anc	0
6	Define neural networks?	neural networks is a network which consists of a set of neuron's connected together to transfer and proc	2
7	What is meant by activation function	Activation function is a kind of function which is only used in hidden layer and output layer and we shoul There are many activation functions. some of them are -Sigmoid function -ReLu -softmax -reverse Relu	2
8	Define structure learning?	Structured learning is a type of learning where the information is in a particular format and it follows a or	0
9	Define loss functions. List any two lc	Loss function is a technique which is used to calculate the loss occurred during the execution. Which is t	0
10	What is meant by activation function	Activation Function:- The function which is used to get the output after the inputs are given through the i	2
11	Define neural networks?	Neural Network:- The collection of neurons together to perform deep learning is called neural networks.	1
12	Define structure learning?	Structure Learning:- This type of learning is a form of learning which changes the structure of the model	2
13	Define loss functions. List any two lc	Loss Functions:- The difference between the actual and the predicted value is the loss function. Backwa	1
14	Define neural networks?	Neural networks is a interconnection of small units called nodes. It takes inputs assigned with some wei	2
15	What is meant by activation function	Activation Function:- Activation function is used to predict the value of the data. It is applied to get the output of the network. I 1) Sigmoid activation fuinction 2)RELU 3) Leaky RELU 4) Tan h	2
16	Define loss functions. List any two lc	It is the difference between the predicted output and the desired output. Absolute mean error and mean square error are two loss function	1
17	Define loss functions. List any two lc	Loss function:- it is function that compares the target and predicted output values. The two loss functions are: -mean square error -Mean absolute error	2
18	Define neural networks?	Neural Networks:- A computer system modelled on the human brain and nervous system. It is also called artifical neuron.	0
19		Activation function:- Activation function decides whether a neuron should be activated or not by calculating the weighted sur List of 3 activation function are -Rectifier function	

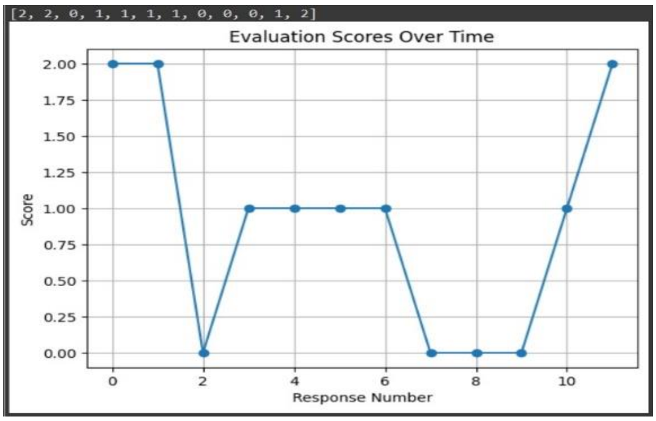
ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24OCT205

✓ Output:

```
Evaluating: 100%|██████████████████████████████| 20/20 [00:20<00:00,  1.00s/it]

Question: Define neural networks?
Answer: Neural netowrk  are the connections of millions or billions of nuron stacked together mimicing like a human brain
Evaluation Marks: 1
Grammar Correction: Neural networks are the connections of millions or billions of neurons stacked together mimicking a human brain.

Evaluating: 100%|██████████████████████████████| 20/20 [00:20<00:00,  1.00s/it]

Question: What is meant by activation function. list any 3 activation functions?
Answer: activation function is used to determine the output of the neuron after summation of weights and inputs takes place.
Three activation functions are:
-sigmoid
-tanh
-Relu
Evaluation Marks: 2
Grammar Correction: The activation function is used to determine the output of the neuron after summation of weights and inputs takes place. Three activation functions are: -sigmoid -tanh -Relu

Evaluating: 100%|██████████████████████████████| 20/20 [00:20<00:00,  1.00s/it]

Question: Define loss functions. List any two loss functions?
Answer: Loss function:-
Evaluation Marks: 0
The response is too short.

Evaluating: 100%|██████████████████████████████| 20/20 [00:20<00:00,  1.00s/it]

Question: Define structure learning?
Answer: structure learning means to determine shape and structure of the network. It dosen't require weights and biases like parametrices learning and only outward structure is learned
Evaluation Marks: 0
The response is out of context.

Evaluating: 100%|██████████████████████████████| 20/20 [00:20<00:00,  1.00s/it]

Question: Define neural networks?
Answer: neural networks is a network which consists of a set of neuron's connected together to transfer and process information
Evaluation Marks: 0
Grammar Correction: Neural networks are networks which consist of a set of neurons connected together to transfer and process information.

Evaluating: 100%|██████████████████████████████| 20/20 [00:20<00:00,  1.00s/it]

Question: What is meant by activation function. list any 3 activation functions?
Answer: Activation function is a kind of function which is only used in hidden layer and output layer and we should use the same activation function to all the neuron's in a respective layer.
There are many activation functions. some of them are
-Sigmoid function
-ReLu
-softmax
-reverse Relu
Evaluation Marks: 2
Grammar Correction: Activation function is a kind of function which is only used in hidden layer and output layer and we should use the same activation function to all the neuron's in a respective layer.
There are many activation functions. Some of them are:
-Sigmoid function
```

➢ *Analysis*

After extensive experimentation, we achieved a 60% accuracy in validating our custom dataset using the GPT model, as compared to evaluations done by human experts. This indicates that while the models showed promise, there's room for improvement in their performance, especially when dealing with unique datasets and tasks.



IV. DATASETS

For our project focused on the "Automation of Answer Scripts Evaluation" using OCR (Optical Character Recognition) technology, we have specifically chosen handwritten text images from college mid-term papers as our primary dataset. Here's a comprehensive explanation of how these datasets were collected and prepared, their granularity, and the annotations process:

A. Data Collection Methodology

- Source: The datasets were sourced from an array of college mid-term examination papers available at our institution. These papers typically contain answers written by students in a handwritten format, making them ideal for our OCR model training.
- Selection Criteria: Papers with diverse handwriting styles were selected to ensure the model's robustness in recognizing different handwriting patterns.

B. Granularity of Data

- Granularity Level: The data was processed at the word level. Each word from the handwritten texts was treated as a single unit for OCR processing.
- Input Size: The OCR model was trained on individual words extracted from sentences. This approach allows the OCR technology to focus on recognizing each word independently, enhancing the accuracy of text recognition.

C. Annotations and Data Preprocessing

- Annotation Process: Each word in the handwritten scripts was manually annotated to match its corresponding text in typed form. This step is crucial as it serves as the ground truth for training the OCR model.
- Annotation Quality Control: Specific attention was paid to ensure high-quality annotations. Any samples with illegible handwriting or unclear annotations were omitted to maintain the integrity of the training data.

D. Challenges with Font and Annotation

Font Issues: Initially, some of the handwritten samples used fonts or styles that were not conducive to accurate OCR recognition (e.g., cursive or highly stylized handwriting). This led to complications in training the OCR model effectively.

Annotation Standards: It was observed that inconsistent annotations could potentially skew the model's learning process. To counter this, we set strict guidelines for how annotations should be formatted, focusing on clarity and uniformity in the text.

E. Improving Data and Annotation Quality

Revising Data Collection: To enhance the OCR model's performance, we are continually looking to diversify the handwriting samples in our dataset. This involves including more varied handwriting styles and ensuring that even subtle nuances in text are accurately annotated.

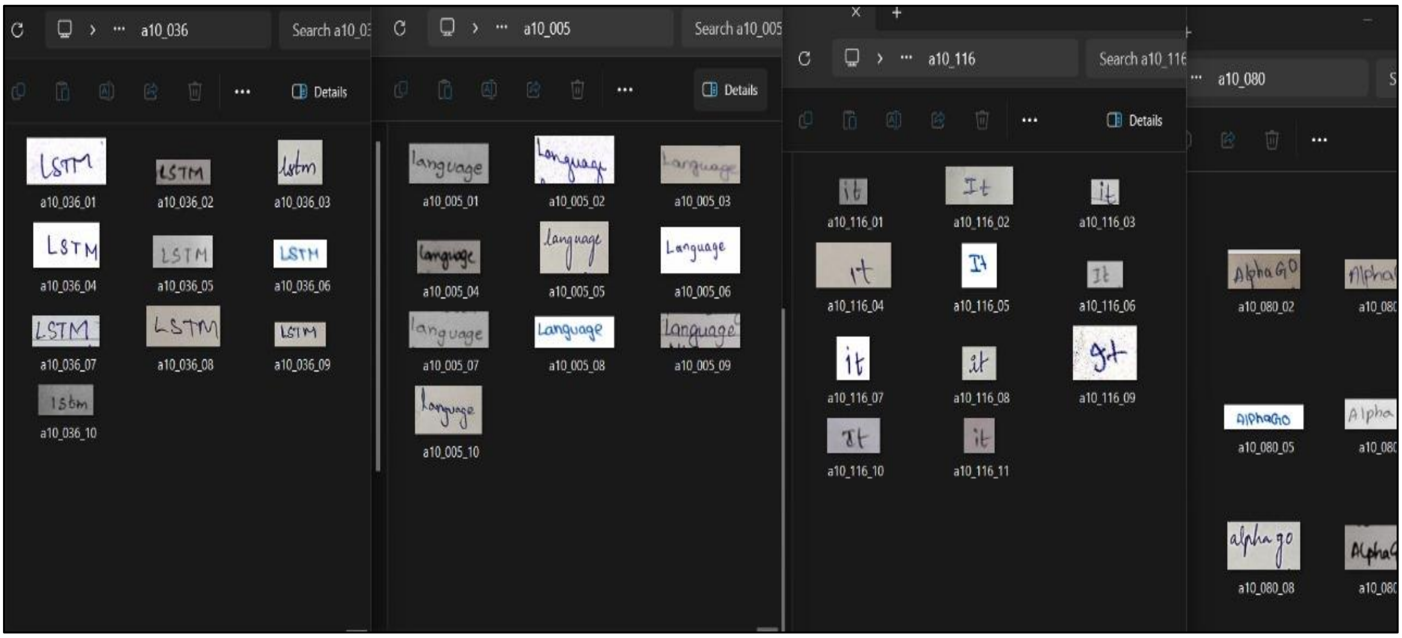
Enhanced Annotation Guidelines: Annotations are now rigorously checked for consistency and legibility. A standardized font guideline was introduced for annotators to follow, which helps in minimizing errors during data entry and improves the model's learning accuracy.

For the BERT and the GPT model we have taken the dataset as our college mid term papers as the input so that they can help the model to train on the student point of view and also understand how the teachers are correcting and based on what criteria are the marks allocated so that the model can be trained similarly to replicate the teachers corrections on the student answer script when it is fed to the model as input for the grading part.

F. BERT Model and GPT Model Datasets Gathering Process

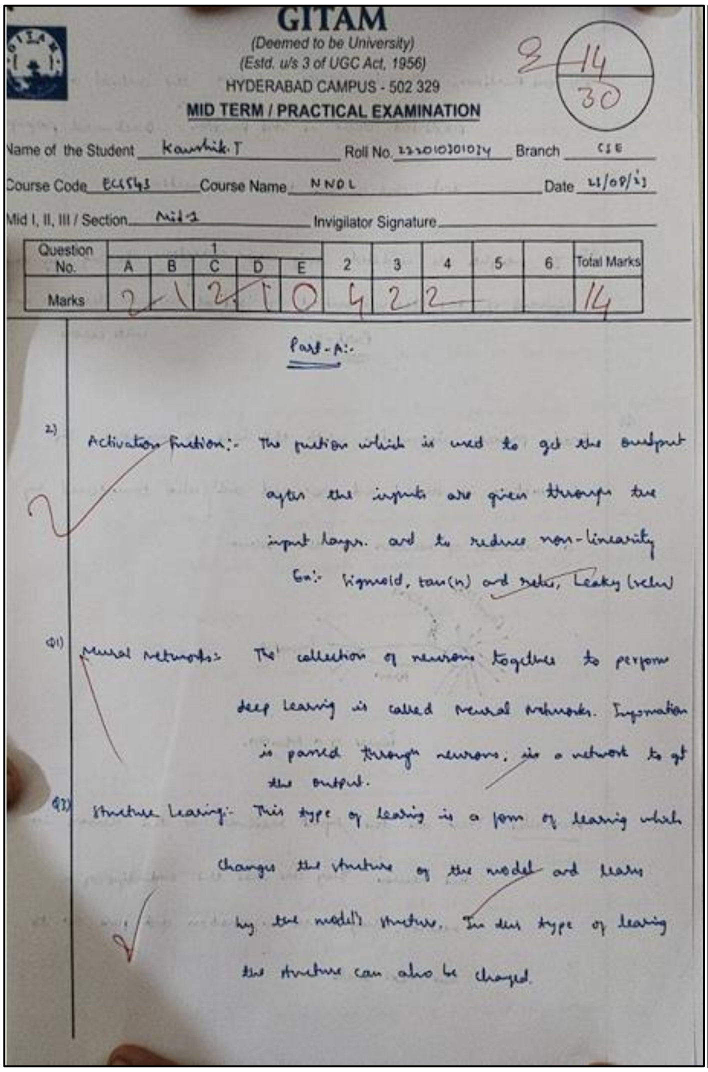
For the BERT and GPT models, we chose our college's mid-term exam papers as the training material. Think of these models like students learning to grade papers just like teachers do. The mid-term papers are full of varied answers from students these are the 'lessons' for our models. They study how teachers check these answers and what reasons they give for the marks they award. This way, our models are learning to grade by understanding the 'teacher's way' of scoring. The goal is for them to get so good that they can look at new answers they've never seen before and grade them just like a teacher would, using the same logic and attention to detail that a real teacher applies when marking a student's work.

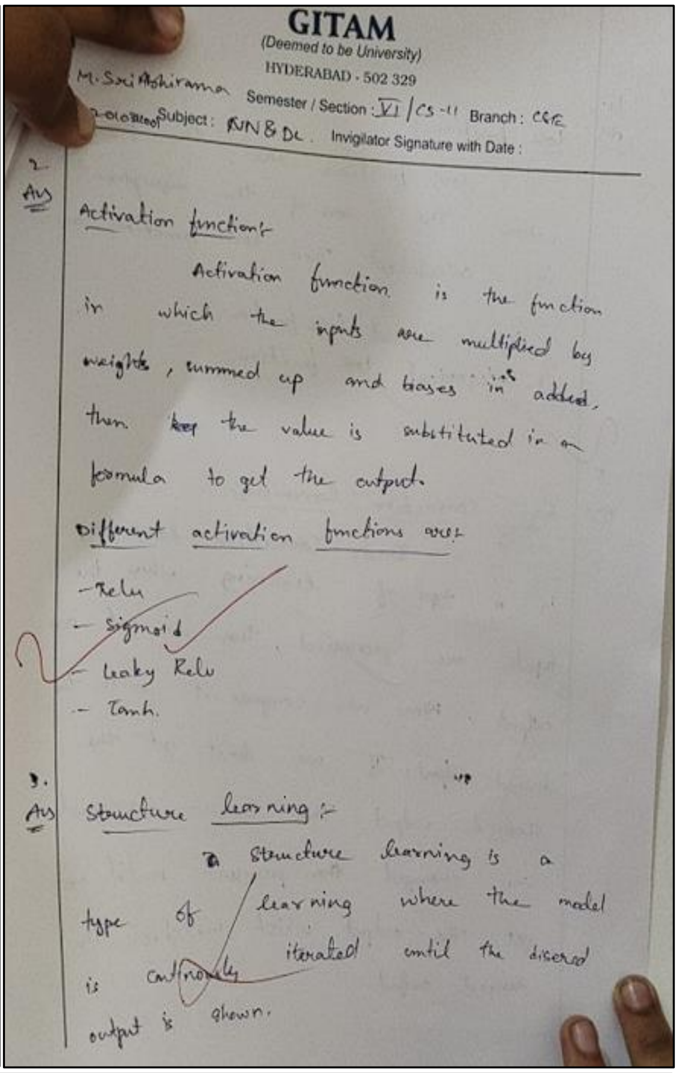
➢ *Handwritten Text Image(Words): -*



Volume 9, Issue 10, October – 2024 International Journal of Innovative Science and Research Technology
ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24OCT205

➢ *Mid-Term Papers from our Teachers:*





Questions	Answers
Define neural ne	Neural netowrk are the connections of millions or billions of nuron stacked together mimicing like a human brain
What is meant b	activation function is used to determine the output of the neuron after summation of weights and inputs takes place. Three activation functions are: -sigmoid -tanh -Relu
Define loss funct	Loss function:-
Define structure	structure learning means to determine shape and structure of the network. It dosen't require weights and biases like parametrices learning and only outward structure is learned
Define neural ne	neural networks is a network which consists of a set of neuron's connected together to transfer and process information
What is meant b	Activation function is a kind of function which is only used in hidden layer and output layer and we should use the same activation function to all the neuron's in a respective layer. There are many activation functions. some of them are -Sigmoid function -ReLu -softmax -reverse Relu
Define structure	Structured learning is a type of learning where the information is in a particular format and it follows a order throughout the set.
Define loss funct	Loss function is a technique which is used to calculate the loss occurred during the execution. Which is the difference between expected and the normal outcomes.
What is meant b	Activation Function:- The function which is used to get the output after the inputs are given through the input layer and to reduce non-linearity. Ex:- Sigmoid, tan(h) and relu, leaky(relu).
Define neural ne	Neural Network:- The collection of neurons together to perform deep learning is called neural networks. Information is passed through neurons is a network to get the output.
Define structure	Structure Learning:- This type of learning is a form of learning which changes the structure of the model and learns by the model's structure. In this type of learning the structure can also be changed.
Define loss funct	Loss Functions:- The difference between the actual and the predicted value is the loss function. Backward propagation and cost(h) are two loss fucntions
Define neural ne	Neural networks is a interconnection of small units called nodes. It takes inputs assigned with some weights. Internally the weights are multiplied by the inputs to aggregate the sum and then pass it to next node. An activation functi
	Activation Function:- Activation function is used to predict the value of the data. It is applied to get the output of the network. It introduces the non-linearity of the model to free. 1) Sigmoid activation fuinction

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24OCT205

A	B	C	D
Questions	Answers	Grades	
Define neural networks?	Neural netowrk are the connections of millions or billions of nuron stacked together mimicing like a hum	1	
What is meant by activation function	activation function is used to determine the output of the neuron after summation of weights and inputs t Three activation functions are: -sigmoid -tanh -Relu	2	
Define loss functions. List any two lc	Loss function:-	0	
Define structure learning?	structure learning means to determine shape and structure of the network. It dosen't require weights anc	0	
Define neural networks?	neural networks is a network which consists of a set of neuron's connected together to transfer and proc	2	
What is meant by activation function	Activation function is a kind of function which is only used in hidden layer and output layer and we shoul There are many activation functions. some of them are -Sigmoid function -ReLu -softmax -reverse Relu	2	
Define structure learning?	Structured learning is a type of learning where the information is in a particular format and it follows a or	0	
Define loss functions. List any two lc	Loss function is a technique which is used to calculate the loss occurred during the execution. Which is t	0	
What is meant by activation function	Activation Function:- The function which is used to get the output after the inputs are given through the ir	2	
Define neural networks?	Neural Network:- The collection of neurons together to perform deep learning is called neural networks.	1	
Define structure learning?	Structure Learning:- This type of learning is a form of learning which changes the structure of the model	2	
Define loss functions. List any two lc	Loss Functions:- The difference between the actual and the predicted value is the loss function. Backwa	1	
Define neural networks?	Neural networks is a interconnection of small units called nodes. It takes inputs assigned with some wei	2	
	Activation Function:- Activation function is used to predict the value of the data. It is applied to get the output of the network. I 1) Sigmoid activation fuinction		

V. CONCLUSION

Table 2: The comparison of OCR, Similarity, BERT and GPT Models

Aspect	Similarity Models	BERT Models	GPT Models
Primary Purpose	Assessing text similarity.	Recognizing contextual information in text.	Generating human-like text responses.
Approach	Iterative improvements on text comparison.	Ensemble and optimization of language models.	Customization for single-answer scripts.
Challenges	Difficulty in handling subjective content.	Real-world data limitations, diverse datasets.	Adapting to specific research needs.
Performance	Varied success with refining techniques.	Hindered by data limitations.	Achieved 60% accuracy with room for improvement.
Key Findings	Limitations in handling subjective content.	Read-world data limitations impact success.	Potential for improvement with unique datasets.
Use Cases	Basic benchmarking, struggles with uniqueness.	Adaption to specific datasets.	Customization for single-answer scripts.
Conclusion	Not suitable for evaluating answer scripts.	Hindered by real-world data limitations.	Promising but room for improvement.

The comparison of OCR, Similarity, BERT, and GPT models as depicted in the provided summary offers an insightful overview of the strengths, limitations, and applicability of these diverse approaches to text analysis and generation.

A. Data Considerations

Our research indicated that the quality of the dataset is paramount across all models. For OCR, the granularity of data at the character level is critical, while for BERT and GPT, context and coherence of text play a significant role. The effectiveness of the Similarity Models hinges on the richness and subjectivity of the text data, underscoring the need for a diverse set of benchmarks.

B. OCR Models

The OCR models were primarily evaluated for their ability to recognize characters within images. Here, the success was largely dependent on the clarity and consistency of handwriting in the datasets provided. Limitations arose when models faced cursive or highly stylized handwriting, leading to inconsistencies in recognition. Despite trials with various models, including homemade, free, and paid services, the challenge of achieving consistent accuracy with diverse handwriting remained.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24OCT205

C. Similarity Models

These models' iterative approach to improving text comparison initially seemed promising for evaluating answer scripts. However, the subjective nature of content in educational settings posed difficulties, as the models struggled to interpret the nuances and context that a human grader naturally understands. Consequently, while these models served as a basic benchmark, they were ultimately found unsuitable for evaluating complex answer scripts, as they could not match the depth of human judgment.

D. BERT Models

The BERT models aimed to recognize contextual information within text, and their performance was often impeded by the limitations of real-world educational datasets, which are diverse and often sparse. While ensemble and optimization techniques were applied, the lack of a substantial and varied dataset meant that the models could not be trained to their full potential, leading to an adaptation challenge when applied to specific datasets.

E. GPT Models

The GPT models showed potential in generating human-like text responses, with a primary focus on customizing for single-answer scripts. The models were adaptable to specific research needs, achieving a 60% accuracy rate. Although this was a significant milestone, there was a consensus that with unique and larger datasets, further improvement in performance could be achieved.

F. Overall Conclusion

The comparative analysis of these models in our project underscores a recurrent theme: the success of machine learning models is intricately tied to the data they are trained on. Real-world educational data presents unique challenges due to its variability and complexity. While OCR models require clear and consistent data, Similarity Models need subjective understanding, and BERT and GPT models necessitate large datasets with varied contextual information to train effectively.

The promise shown by BERT and GPT models suggests a pathway forward—focusing on data enhancement and model fine-tuning could potentially bridge the gap between automated grading and human-like evaluation accuracy. Hence, our future efforts will be directed towards curating more comprehensive datasets, encompassing a wider array of handwriting and answer styles, to further improve the models' ability to evaluate student scripts with the same nuance and insight as experienced educators.

REFERENCES

[1]. AUTOMATION OF ANSWER SCRIPTS EVALUATION-A REVIEW Ravikumar M1, Sampath Kumar S1 and Shivakumar G1

[2]. Grade Aid: a framework for automatic short answers grading in educational contexts—design, implementation and evaluation Emiliano del Gobbo1 · Alfonso Guarino2 · Barbara Cafarelli1 · Luca Grilli1

[3]. NLP-based Automatic Answer Script Evaluation Md. Motiur Rahman1 and Fazlul Hasan Siddiqui2*

[4]. An Automatic Short Answer Correction System Based on the Course Material Zeinab Ezz Elarab Attia1* Waleed Arafa1 Mervat Gheith1

[5]. Automatic Evaluation of Descriptive Answers Using NLP and Machine learning. Prof. Sumedha P Raut1, Siddhesh D Chaudhari2, Varun B Waghole3, Pruthviraj U Jadhav4, Abhishek B Saste5