Subjective Answer Evaluation using Machine Learning and Image Processing (NLP)

Sandip Ugile
Department of *Information Technology, BRACT's Vishwakarma Institute of Information Technology*
Pune,Maharashtra,India
sandip.22110679@viit.ac.in

Sankalp Tembhurne
Department of *Information Technology, BRACT's Vishwakarma Institute of Information Technology*
Pune,Maharashtra,India
sanklap.22111127@viit.ac.in

Pratik Wani
Department of *Information Technology, BRACT's Vishwakarma Institute of Information Technology*
Pune,Maharashtra,India
Pratik.22110799@viit.ac.in

Jayashri Bagade
Department of *Information Technology, BRACT's Vishwakarma Institute of Information Technology*
Pune,Maharashtra,India
jayashree.bagade@viit.ac.in

Abstract: This paper presents a new way to evaluate the quality of content in educational assessments by integrating image processing techniques, natural language processing (NLP), and cosine similarity. Image processing algorithms are used to extract relevant features from text, including layouts and graphics, to improve the response. Additionally, the text is processed using the Bidirectional Encoder Representation (BERT) model to capture ambiguous content. The fusion mechanism combines visual and textual data for easy analysis. Cosine similarity is used to measure the similarity between student and reference responses. Experimental results demonstrate the effectiveness of this method, achieving greater accuracy and robustness compared to traditional methods, reducing dependence on human observers, and minimizing errors. This research contributes to the evaluation of education by measuring the quality of responses with reliable, valid and objective methods.

Keyword:

Image processing, NLP, BERT model, cosine similarity, learning metrics, response content analysis, integration, pytesseract .

I. INTRODUCTION

Nowadays, the world is moving towards automation, so there needs to be automation in schools too. Manual evaluation of descriptive responses requires a lot of time and effort on the part of the analyst. When a teacher manually evaluates an assignment, the quality of the evaluation will vary depending on the evaluator's preferences. Therefore, the distribution of loans may not be appropriate. This system can be used to ease their burden and distribute points equally based on evaluation of answers.

The system will solve the evaluation problem by using machine learning and natural language processing as the basis for efficient operation. In machine learning, all results are based on the input data provided, while NLP will provide results that treat the terms and phrases as the original response. Answers will be evaluated based on NLP, just scan the test paper and the system will give points based on the available information.

The main purpose of our project is to ensure that the institution uses user-friendly and more interactive software. Assessment will ensure fairness and fairness in evaluating the answers because all students will use the same thinking strategies. Compared with other models, this article will demonstrate the improvement in performance using the BERT model architecture as the basis for good performance..

II. LITERATURE SURVEY

Bashir et al.[1] proposed an automated test based on BERT, which achieved a 91% accuracy in test results. Meanwhile, Kryvinska et al. ASAP-AES Collaborate first using BERT and XLNET on the Kaggle dataset, highlighting the importance of combining new data in new ways. Additionally, Javed et al. An NLP-based decision tree model was introduced, achieving an accuracy of 87% despite its performance-dependent nature. Introduced a semantic analysis method using concept mapping, fuzzy string matching, and syntax analysis. Their methods showed excellent precision, recall, and sensitivity of 95%, 94%, and 94.5%, respectively, and targeted most questions rather than answers. These differences reflect ongoing efforts to use NLP and ML techniques to improve the efficiency and accuracy of automated response assessment.

Kudale et al. proposed a mathematical model using NLP before using the technology and achieved 90.3% accuracy by calculating cosine similarity. In contrast, Mari et al. When using a machine learning model that includes data collection, processing, and deletion, the accuracy rate is up to 88%. However, they encountered

problems with long learning times, especially k-means integration. Experiments using WDM and cosine similarity models show that the accuracy of WDM increases from 15.6% to 13.4%, while the accuracy of cosine similarity increases to 87%. They emphasize the importance of relaxed WDM over traditional WDM. This trend reflects ongoing efforts to use NLP and ML technologies to improve the efficiency and accuracy of machine learning systems..

JETIR.ORG[2] employs NLP for subjective answer script evaluation, albeit facing challenges with lower accuracy initially, which improved to 80% after dataset enhancement. IJNLC incorporates ontology-based evaluation, achieving a high correlation of up to 90% but facing limitations due to slower multi-hash map usage. IJARIE integrates ML and NLP with Glove Word Embedding and cosine similarity, achieving an 85% accuracy at the expense of high training costs.

Kagliwal et al.[3] introduced Automatic Response Analysis (ASAG), which has the advantage of using time-consuming and contextual response processing with NLP technology such as BERT and XLNet. Their method involves collecting text and evaluating it using a variable model to extract important information. But they also acknowledged the limitations of sustaining a long-term response due to the long-term limitations of the change model. They suggest improvements, such as improving the content of the text to better capture important information, and explore ways to overcome the limitations of storing long answers in the model. It also shows that joint integration or the use of specialized knowledge can improve the performance of the system.

Waghmare et al.[4] Launch of Pariksha Software, an automated system designed to improve the process of analyzing content in exams using natural language processing (NLP) and machine learning (ML) technologies. The system solves problems with manual testing, especially in the context of the COVID-19 pandemic where online testing has become more common. It uses NLP techniques such as stemming, lemmatization, and stopped word removal, as well as ML algorithms such as Naive Bayes and Decision Trees, to evaluate responses based on criteria such as response length, ratio terms, syntax analysis, cosine, and more. similar terms model answers. Additionally, the system uses algorithms to detect inconsistencies in student responses. It provides huge benefits to schools by reducing the work of examiners and providing a more objective assessment process for online exams.

Bashir et al. [5] examined the use of natural language processing (NLP) in response evaluation and highlighted its importance in providing computer-like understanding of text and speech. They explored various NLP techniques such as syntax analysis, content extraction, logic and similarity analysis, dissimilarity analysis, and cosine similarity to evaluate the answered language. The system uses machine learning algorithms to complete the evaluation process, thus reducing workload and providing consistent scores. Previous studies have demonstrated the effectiveness of semi-automatic indexing techniques using query response libraries and hash indexes. The proposed system is effective, achieving 90.3% accuracy and providing information about students' language and thinking skills, ultimately streamlining the assessment process and ensuring that the feedback is useful to teachers.

Karthika et al. [6] proposed a method designed to extract variables such as word frequency, sentence length, and sentiment analysis from responses using NLP methods. Machine learning algorithms are then used to predict the scores of new answers using the extracted scores. The performance of the proposed system will be compared to existing systems using measurements such as accuracy, precision, and recall, and will be competed against to ensure its validity on new data.

Mangesh et al. [7] proposed a model that uses key attributes in descriptive responses, including content, Quantum Serge Therapy (QST), and syntax, to derive scores. Before and after collecting the data, the model extracts features such as points and QST, which are used to train the Gaussian Naive Bayes classifier. The accuracy of this method is approximately 80%. These methods include methods such as data collection and description, prioritization, similarity assessment, model training, estimation and forecasting. This implementation is made in Python; The accuracy and performance of measurements is improved by leveraging a rich library of functions for image processing and machine learning.

Patil et al. [8] proposed a combination of machine learning (ML) and natural language processing (NLP) to obtain evaluation of text responses. It tags, tags, and analyses text to assign tags based on semantic content using algorithms such as Naive Bayes. The system has two modules: one for extracting information from scanned images and the other for machine learning and natural language analysis. Experimental results show high agreement between human and system evaluation. Future improvements will include integrated feedback and expanded data to increase accuracy.

Devi and Mittal [9] stated in their study that rhyme, correct communication and helping strategy led to more evaluations. This study demonstrates the effectiveness of integrating

ontology into the machine learning process in evaluating content, especially in the field of computer graphics. Future studies could explore collaborative feedback and further refine the ontology review process to increase accuracy.

A cosine-based sentence similarity measure-based automated answer evaluation system is presented in the study by Madhumitha Ramamurthy and Ilango Krishnamurthi. It uses a synset-based word similarity measure for dimensionality reduction and presents 21 new cosine-based measures with various construction methods for document vector space. Using the MSR paraphrase corpus, Li's benchmark datasets, and the Kaggle short answer and essay datasets, the system is assessed. A significant correlation is found between the system-generated and human scores when they are compared using Pearson correlation. The study emphasizes the difficulties that come with evaluating text responses and the need for automation to lessen the labour-intensive process and any bias associated with human judgment. It also explores the application of similarity metrics in text processing, emphasizing techniques like TF/IDF and the bag-of-words approach.

Eduardo C. Garrido-Merchan, Roberto Gozalo-Brizuela, and Santiago Gonzalez-Carvajal published a paper comparing Transformers' Bidirectional Encoder Representation (BERT) model with content frequency Inverse Document Frequency (TFIDF) contents. It evaluates the performance of BERT against traditional machine learning models in a variety of scenarios, demonstrating the superiority of BERT and its independence from the unique features of natural language processing (NLP). This study highlights the differences between conversational expert-generated rules and machine learning (ML) approaches based on data sources such as BERT and Transformer. BERT has been shown to outperform ML NLP methods and demonstrates its effectiveness on NLP problems. The article highlights the important role of adaptive learning (especially prior training) in improving the performance of BERT. However, he also acknowledges the limitations of BERT and recommends further research on hyperparameter autotuning and its application to robotics. The authors propose to explore the potential of BERT in new NLP tasks and its suitability for speech classification by character-remembering robots by modifying hyperparameter models using Bayesian optimization.

The article by Timothee Mickus, Denis Paperno, Mathieu Constant, and Kees van Deemter evaluates BERT as a distribution model focused on content placements. The case study explores the emergence of BERT as a state-of-the-art natural language processing (NLP) deep neural network and its role as a contextual embedding model. He talked about BERT's success in various NLP tasks and benchmarks, emphasizing its contribution to the understanding of natural language. In addition, by comparing BERT content embeddings with traditional distributed semantic models, this article highlights the necessity of a thought-based lexical semantics perspective for understanding BERT objects. This comprehensive review lays the groundwork for a later examination of the semantic equivalence of BERT and explores the impact of its design and training on its semantic properties.

The paper, written by Professor Sharayu Lokhande, Udit Chaudhary, Akash Singh, Pranay Gaikwad, Himanshu Guleria, and Shilpa Pawar, explores the field of automatic subjective assessment techniques with a special focus on leveraging LSTM (Long Short Term Memory) networks. The literature review highlights potential challenges associated with response evaluation, including cost, resource and time impacts, and safety concerns. It demonstrates the urgent need for automatic evaluation methods and demonstrates the purpose of this paper to advance machine learning, NLTK, Python, recurrent neural networks, and web technologies. The survey demonstrates the potential of NLP technology combined with OCR (optical character recognition) to reduce the burden on auditors and improve results. It also shows the importance of large and accurate data sets for training classifiers, especially as the learning process increases. Overall, the research paper lays the foundation for this paper's study of LSTM-based methods for automatic response measurement, pointing to future developments in word embedding methods and recording semantic similarity profiles.

III. Methodology

1. BERT-Embedding: Transformers' Bidirectional Encoder Representations is referred to as BERT. BERT is a machine learning framework that is free and open source, aimed at processing natural language. BERT uses the surrounding text to provide context for computers. Determine the definitions of any ambiguous terms in the text. Wikipedia information is used to pre-train the BERT system, and question-and-answer datasets can help it become even more efficient. For answer selection issues, the proposed approach is inspired to employ the BERT, a contextualized embedding, to simulate phrase similarity. It can record a statement more precisely by giving each word an embedding depending on the circumstances around it. Figure 1 depicts

the operation of BERT.fig 1 represents the working of bert model.

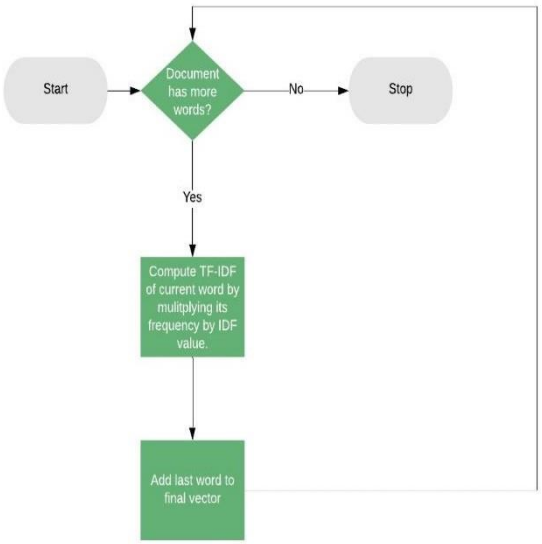

Fig.1. Working of BERT model.

The software we are using to get pre-trained models is called Sentence Transformers. Considering that our use case requires English. Functionality. We make use of the pre-trained bert-base-nli-mean-tokens prototype as part of the BERT English working philosophy. To compare the degree of similarity between words, we should utilize a tool such as cosine similarity. To measure Recall and Precision, each token in x is compared to the token xˆ that is the closest match, and vice versa. It's an avaricious and lonely relationship. The F1 score is calculated by summing up precision and recall. Precision, recall, and F1 score are metrics that may be found using equations 1, 2, and 3.

equation 1:

$$R_{Bert} = \frac{1}{|x|} \sum_{x_i \in x} \max_{\hat{y} \in \hat{x}} x_i^T \hat{x}_j$$

equation 2:

$$P_{Bert} = \frac{1}{|\hat{x}|} \sum_{\hat{y} \in x} \max_{x_i \in x} x_i^T \hat{x}_j$$

equation 3:

$$F_{BERT} = 2\frac{P_{BERT.R_{BERT}}}{P_{BERT} + R_{BERT}}$$

2. Cosine similarity: It is important to evaluate similar texts in various applications such as approval, evaluation, and plagiarism evaluation. The principle of measuring the angle between two vectors, essentially measuring their distance. The following equation describes how to calculate the cosine similarity between two nonzero vectors using the Euclidean point object model. This mathematical concept plays an important role in determining the coherence of text, thus allowing accurate measurements in various areas of text analysis.

3. Pytesseract:

Extracting text from images is an important task in many scientific fields; It converts information written in images into machine-

$$Sim(A,B) = \frac{A \cdot B}{\|A\| X \|B\|} = \sum_{1}^{n} \frac{A_i X B_i}{\sqrt{\sum_1^n A_i^2 X} \sqrt{\sum_1^n B_i^2 X}}$$

readable data. In machine learning, Pytesseract (a Python wrapper for the Google Tesseract-OCR engine) is important for this purpose. The process involves first processing the image to enhance the text appearance and then using Pytesseract to extract the text using the "image-to-string" function. Post-processing methods can be used to restore deleted files. The extracted data can then be fed into a mechanical pipeline for further analysis, classification, or other related processing. Overall, using Pytesseract to extract text from images helps convert visual data into readable text, thus facilitating research in many fields. As fig.2 represents the working of pytesseract and hoe the text is extracted from the image it is used in this project for extraction the answers of students for given questions. Figure 2 represnets the working of pytesseract.

As fig.3 shows the system first uses pytesseract for optical character recognition (OCR) to extract text from scanned images of responses. The extracted text is then processed and converted into a high vector representation to embed the content using BERT (Bidirectional Encoder Representation with Transformers). The cosine similarity between the answers is then calculated, which is used to evaluate the similarity of students' answers with their placements. The system helps in comprehensive and comprehensive evaluation of student performance by providing feedback or responses based on predefined criteria or criteria.

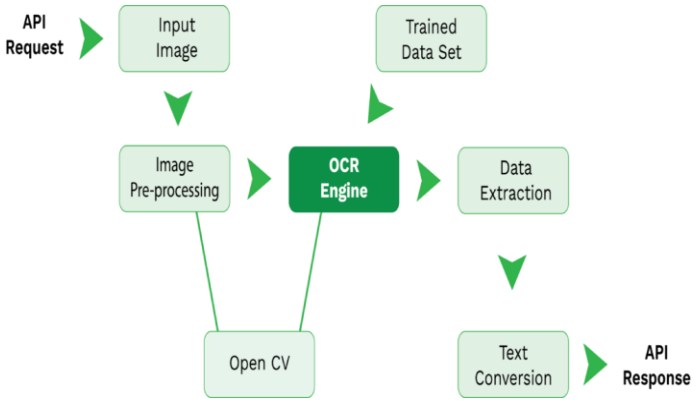

Fig.2. Working of Pytesseract

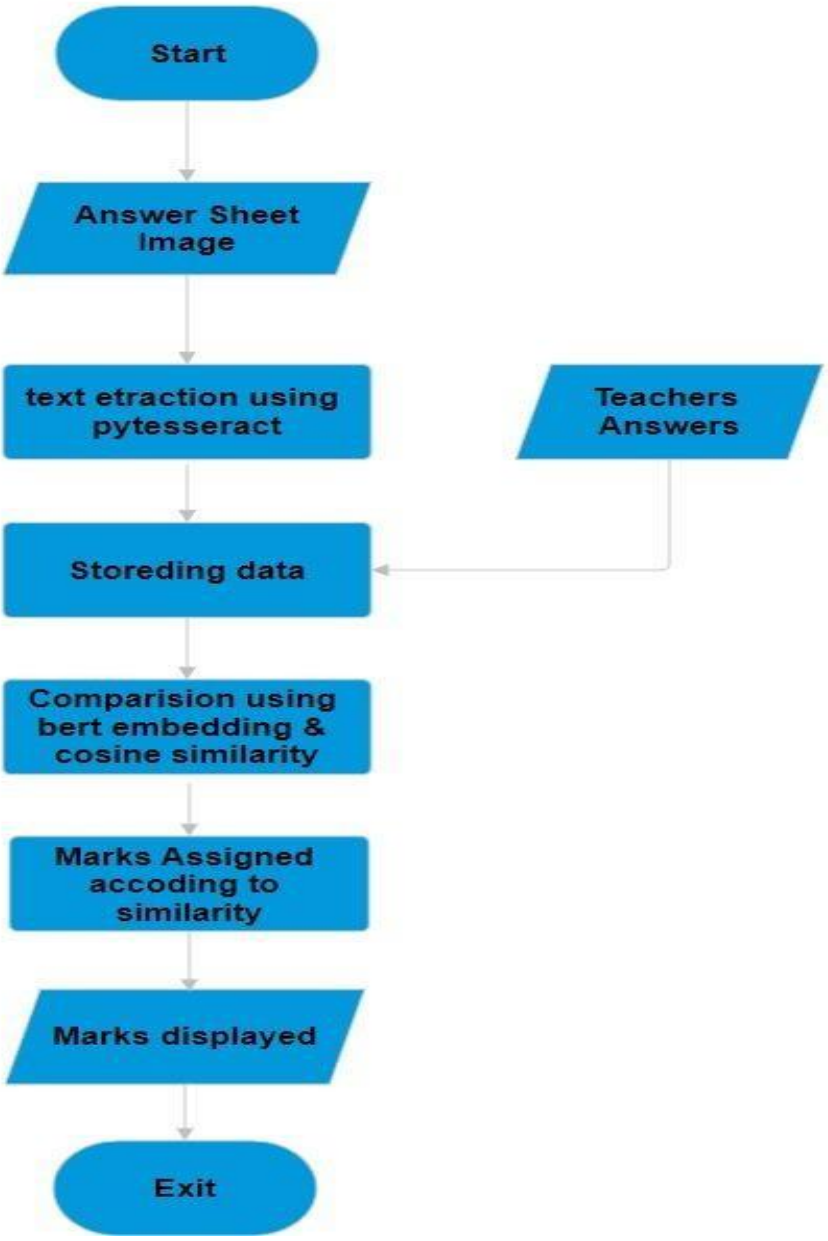

Fig.3. Workflow of model

IV. Data Collection

The data collection process involves teachers and students. Initially, there are 4 students in the test and there are 4 questions on the test paper, each worth 5 points. The teacher's answers to 4 questions were recorded as the answers of 4 students to the same question. Thus, the data includes questions, teacher responses, and student responses.

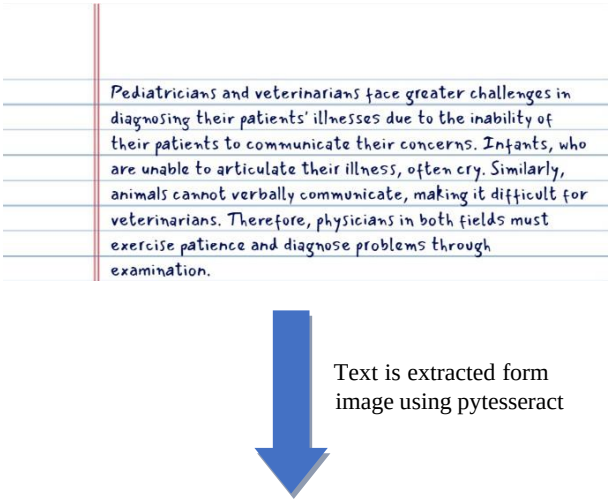

Fig.4.Text extraction form image

A method of extracting ideas from a database containing questions, teacher responses, and student responses. Initially, the BERT model was pre-trained to obtain similar scores on teacher and student responses. Additionally, cosine is used to measure similarity. The scores of these three parameters are combined in the ratio7:2 to calculate the score of the final machine produced by the system. Then, during the test, the student's answers are compared with the score generated by the system. The final label is determined by a neural network with a thick layer and a sigmoid activation function. The release process is started to keep the nervous system regular and prevent irritation. The prediction results in a label of 0 or 1, corresponding to correct or incorrect answers, respectively. Parameters were not specified when training the needs of the neural network, including the use of the Adam optimizer with a learning rate of 0.001. To improve performance, the model is trained for a certain period.

V. Results and Discussion

The performance of the suggested model for automated answer evaluation is covered in this section. Five pupils' responses are taken into consideration for the evaluation together with the human equivalents. The similarity score and difference between student and human responses are computed by the algorithm. Every response is assessed by a human before being assessed by our system. Table II displays the combined score that the computer and human evaluators assigned because BERT is bidirectional that is, able to process text from both the left and the right it is utilised in this work. The transformer model's encoder segment is used by BERT. First, the Machine score is computed after each BERT score is obtained separately. Individual similarity measures, however, did not produce higher rankings.

In this model we have given input of answer of five students written by them as a input to our machine and as a result the machin eshowns the text extracted and also give the appropriate maks according to the the quality of answer. The more the similarity in between the teachers answer and the students answer more marks will be given . we also given the exact answer as the teacher’s answer the machine successfully give 4.9 marks for it. For another scenario if student do not have written any answer the marks given to it is zero. In this way tested different cases. The evaluation's dataset, which consisted of 100 distinct answer scripts, was manually put together. Our method produced the best results, with a recall of 79%, accuracy of 91%, and precision of 83%.

VI. Output Screenshots:

	A	B
1	Answers	Marks
2	The kite festival of Ahmedabad is so special because there is International kite festival at Ahmedabad organised by Gujarat state. Internation Kite Festival is held either at the Sardar Patel Stadium or at the P	5
3	The poet calls the vet 'omniscient' because he is expected to know everything about animals. The vet knows about the illnesses and treatments of all animals, so he is called omniscient	5
4	Pediatricians and veterinarians have a more difficult job diagnosing their patients illnesses because their patients cannot express their concerns. The infants are unable to express their illness and often cry	5
5	When animals are not well, they become quiet an inactive. Some may even groan. They also stop eating food.Animals will use many parts of their bodies to convey various feelings. Being aware of how your p	5
6		

Fig.5 The extracted text is stored in database

```
PS C:\Users\sanka\OneDrive\Desktop\Subjective> python -u "c:\Users\sanka\OneDrive\Desktop\Subjective\app3.py"
Processing folder: student1
Text Extracted from ans1.png:
The kite festival of Ahmedabad is so special because there is
International kite festival at Ahmedabad organised by Gujarat
state. Internation Kite Festival is held either at the Sardar Patel
Stadium or at the Police Station. I+ attracts international kite
flyers as well as local champions. Little Indian paper kites
compete with those from Japan, the USA, the UK, and other

countries.


Marks Obtained: 4.9583190910458335

Text Extracted from ans2.png:
The vet is called ‘omniscient’ by the poet because he is presumed
to have all-encompassing knowledge about animals. This includes
knowing about the illnesses and treatments for all animals.

Marks Obtained: 3.413641903145373

Text Extracted from ans3.png:
Pediatricians and veterinarians face greater challenges in
diagnosing their patients’ illnesses due to the inability of their
patients te communicate their concerns. Infants, whe are unable
te articulate their illness, often cry. Similarly, animals cannot
verbally communicate, making it difficult for veterinarians.
Therefore, physicians in both fields must exercise patience and
diagnose problems through examination.
```

Fig.6 machine output as extracted text and respective marks


```
PS C:\Users\sanka\OneDrive\Desktop\Subjective> python -u "c:\Users\sanka\OneDrive\Desktop\Subjective\app3.py"
Processing folder: student1
Text Extracted from ans1.png:
The kite festival of Ahmedabad is so special because there is
International kite festival at Ahmedabad organised by Gujarat
state. Internation Kite Festival is held either at the Sardar Patel
Stadium or at the Police Station. I+ attracts international kite
flyers as well as local champions. Little Indian paper kites
compete with those from Japan, the USA, the UK, and other

countries.


Marks Obtained: 4.9583190910458335

Text Extracted from ans2.png:
The vet is called ‘omniscient’ by the poet because he is presumed
to have all-encompassing knowledge about animals. This includes
knowing about the illnesses and treatments for all animals.

Marks Obtained: 3.413641903145373

Text Extracted from ans3.png:
Pediatricians and veterinarians face greater challenges in
diagnosing their patients’ illnesses due to the inability of their
patients te communicate their concerns. Infants, whe are unable
te articulate their illness, often cry. Similarly, animals cannot
verbally communicate, making it difficult for veterinarians.
Therefore, physicians in both fields must exercise patience and
diagnose problems through examination.
```

VII. Conclusion

The current manual assessment method makes it more difficult to score student responses. Further issues are brought up by the valuation scheme, which requires significant financial, time, and people resources. An assessment tool that runs automatically is recommended for assessing descriptive response types in order to overcome these obstacles. The suggested system uses BERT to automatically check and score descriptive answers. The evaluation's dataset, which consisted of 100 distinct answer scripts, was manually put together. Our method produced the best results, with a recall of 79%, accuracy of 91%, and precision of 83%.

VIII. References

[1] Farrukh Bashir, Hamza Arshad, Abdul Rehman Javed, Natalia Kryvinska, Shahab S. Band, (2017) Subjective Answer Evaluation Using Machine Learning and Natural Language Processing.

[2] Gaurang Kudale, Nishant Mali, Nachiket Suryawanshi, Mukesh Bansode, Prof. Richa Agarwal (2023) Automated Subjective Answer Evaluation Using NLP.

[3] Sarthak Kagliwal , Jagruti Agrawal, Tejas Dahad , Atharva Saraf , Karan Kangude (2021) Subjective Answer Evaluator.

[4] Ajay Waghmare, Supriya Chaudhary, Mohit Kambayat, Abhishek Girkar, and Natural Language Processing (2021) Subjective Answer Evaluation.

[5] Muhammad Farrukh Bashir, Hamza Arshad, Abdul Rehman Javed, N. Kryvinska, "Subjective Answers Evaluation Using Machine Learning and Natural Language Processing"

[6] Karthika. K ,Akshaya. G, Manoshree.T., "International Journal of Multidisciplinary Research Transactions(IJMRT), Subjective Answer Evaluation Using Machine Learning And Natural Language Processing"

[7] Sangeeta Mangesh , Prateek Maheshwari, Aditi Upadhyaya, "Subjective Answer Script Evaluation Using Natural Language Processing", journal of emerging technologies and innovative research(JETIR).

[8] Piyush Patil, Sachin Patil, Vaibhav Miniyar, Amol Bandal, "Subjective Answer Evaluation Using Machine Learning", International Journal of Pure and Applied Mathematic, May 23, 2018

[9] M. Syaamala Devi and Himani Mittal , "Machine Learning Techniques With Ontology For Subjective Answer Evaluation" , International Journal on Language Computing , 2 April 2016

[10] Madhumitha Ramamurthy ,Ilango Krishnamurthi, "Design and Development of a Framework for an Automatic Answer Evaluation System Based on Similarity Measures", ResearchGate

[11] Eduardo C. Garrido-Merchan, Roberto Gozalo-Brizuela ,Santiago Gonzalez-Carvajal ," Comparing BERT Against Traditional Machine Learning Models in Text Classification",Jcce

[12] Timothee Mickus, Denis Paperno, Mathieu Constan, Kees van Deemter, "Assessing BERT as a Distributional Semantics Model" ,2020

[13] Prof. Sharayu Lokhande, Udit Chaudhary, Akash Singh, Pranay Gaikwad, Himanshu Guleria, Prof. Shilpa Pawar, "Automated Subjective Answer Evaluation System",2019