A Novel AI-Based Paper Evaluation System Using Computer Vision and Natural Language Processing for Accurate Automated Grading

T. Daniya
Department of CSE (AI&ML)
GMR Institute of Technology
Rajam-532127, Andhra Pradesh, India

Abstract:

In educational institutions, paper evaluation is a significant and time-consuming process. Combining computer vision and machine learning approaches has showed promise in automating this process in recent years. However, many existing strategies concentrate just on studying the question or answer paper. In this paper, the proposed system consists of modules like the question paper analysis module the answer paper analysis module. The Answer and question papers are provided as an input in the form of image, pdf, word and excel format. Then those modules use an optical character recognition (OCR) technology like Tesseract to extract the text of the image or document. Then, it uses contemporary NLP algorithms and metrics like The Bag of words score (83%), ROUGE score (93%), BLEU scores (78%) are employed to assess papers and determine the similarity between modules and gave the final score with an accuracy of 94%.

Keywords:

Automatic paper evaluation, Natural Language Processing (NLP), Optical character recognition, BLEU score, ROUGE score.

Introduction

One method of evaluating a student's performance and skills is through the use of subjective questions and their responses. Of course, the answers can be anything the students need them to be, based on their personal comprehension of the material. Having said that, there are still a few crucial distinctions between subjective and objective solutions. They are longer than the objective questions, for starters. Second, writing them requires more time. In addition to that it’s defining feature both evaluator and student view characteristics. Context, and subjective tests and educators are more challenging and scarier. In the evaluator must pay close attention to grade a subjective response and go through each word. The overall rating is significantly impacted by the mental condition, amount of fatigue, and objectivity of the checker.

Consequently, allowing a system to handle this situation, a vital task of subjectivity analysis is a great deal more time and resource effective. Machine Assessment of objective reactions is generally straightforward and practical. A tool that immediately maps student learning Questions and one-word answers can be supplied into the responses. However, handling subjective answers is much complex. They have wide range of lengths and a huge vocabulary. They require a lot of focus and objectivity from the teacher evaluating them and carry a lot more context. It is difficult to evaluate these questions using computers task, primarily due to the ambiguity of natural language.

Additionally, people frequently utilize easy acronyms and synonyms, which complicates the procedure. Comparing the similarity of different texts, words, and documents, finding the text's context and mapping it to the solution's context, counting the documents' noun-phrases, contrasting phrases in the answers, and additional elements of subjective answer evaluation have all been the subject of extensive research. The requirement for better datasets, lack of hyper-parameter adjustment, and semantic context loss in Tf-Idf are having still issues.

Before diving into the data, cleaning, and tokenization, you must complete the essential preparatory processes. After that, you can compare the textual data with a number of tools, such as ontologies, conceptual graphs, document similarity, and latent semantic structures. According on similarity, keyword presence, structure, and language, the final score can be assessed. As this problem has been the subject of countless prior attempts, but there is still a need for improvement, some of which are discussed in the above are done in this work.

In this paper, we examine a method for evaluating subjective answers. Our research is based on natural language processing methods like Vectorization, and lemmatization, text representation methods like word2vec, Optical character recognition (Ocr) methods, word occurrence can be calculated using the tokenization and some other metrics and for calculating similarity between words like ROUGE score and Bag of words score, Bilingual understudy score. We compare the performance of model using various assessment metrics, such as F1-score, Accuracy. The above steps are clearly discussed in Methodology and results sections. We also go over numerous methods that have been employed in the past to assess subjective responses or, more generally, text similarity.

Related works

Weegar et al., (2023) explains about that by measuring the whole workload necessary to accurately grade an entire exam, with and without the use of machine learning. Over 400 students took part in the evaluation during an introductory computer science course. The exam had 64 questions and relatively short responses, and automated grading was done in two steps. A portion of the exam answers were first manually scored, and the remainder were then utilized as training data for machine learning models that classified the remaining answers. Various methods for choosing which responses to include in the training data were tested. The amount of time spent on various grading tasks as well as the effort saved by applying automated scoring and answer clustering were measured. The workload reduction overall was significant between 64% and 74% compared to totally manual grading.

Yue, X et al., (2023) focuses on the automated assessment of attribution by sophisticated language models (LLMs). The work explores two approaches for automated evaluation to address this problem: triggering linear regression models (LLMs) and refining smaller LMs. Using data from similar tasks such as question answering, fact-checking, natural language interpretation, and summarization, the investigators train these smaller language models. There are two evaluation sets: AttrEval-Simulation and AttrEvalGenSearch, which are annotated from a generative search engine and simulated based on QA datasets, respectively. For overall performance, metrics like the F1 score and Micro F1 score are utilized. They tested various models using zero-shot, few-shot, and fine-tuning parameters, and their results on the simulated test were 60% F1 and roughly 55% on the AttrEval-GenSearch test set restriction.

Xu, F et al., (2023) proposed the evaluation of long-form answers is the main aim of this paper, which also examines automated and human evaluation techniques. The human evaluation techniques are included like Crowd annotators are given a question, two candidate answers, and evidence documents. In Automatic Evaluation they used mainly two automatic metrics such as QAFactEval and RankGen for fine grading the LFQA. The paper also studies the relationship between automatic metrics and human judgements and analyses domain experts' preferences, arguing that LFQA requires a multifaceted evaluation strategy. Finally, the overall accuracy is about 62% and it is very less compared to the human evaluation.

Lu, Y et al., (2022) address the issues in evaluation, they suggested the ProtSi Network, a novel semi supervised design that initially applies few-shot learning to the subjective evaluation of answers. ProtSi Network, which combines Siamese Network, which is composed of BERT and encoder layers with Prototypical Network, simulates the natural process of evaluators assessing replies to evaluate students' responses by similarity prototypes. To avoid overfitting for efficient few-shot text classification, they used the unsupervised varied paraphrasing model Prot Augment. The ProtSi Network performs better than the most recent baseline models in experiments on the Kaggle Short Scoring Dataset in terms of accuracy and quadratic weighted kappa. Finally, the model shows the accuracy about 63% and the quadratic weighted kappa is about 74%.

Using deep neural networks (DNNs) to reliably evaluate subjective English queries is something that Zhao, S. (2022) looks into. Our primary goal is to overcome the problems associated with automated evaluation, such as its expensive price tag and ineffectiveness. In this research, the authors provide an evaluation strategy based on DNNs. Using the features of a Deep Neural Network, the model examines subjective questions in English. To make these subjective questions simpler and accurate, an assessment system that is based on DNNs strives. While the recall rate might approach 90%, the mean accuracy for this model is 85%. In terms of efficacy, recall rate, and accuracy, the proposed evaluation model—which is based on DNNs—displays outstanding results. Deep neural networks can efficiently evaluate subjective English questions, according to the research, eliminating the need for costly and time-consuming manual review.

The digitalization of invoice information is investigated by Kamisetty et al. (2022) by the utilization of Optical Character Recognition (OCR) technologies. The overarching objective is for machines to be able to read and evaluate invoices in any format and automatically extract relevant information. In this document, the basic steps of digitization are shown. We primarily use scanning or graphic capture technologies as our primary method for obtaining the invoices. After that, application programming interfaces (APIs) or specialist applications use optical character recognition (OCR) capabilities to find and read the data in the pictures. After extraction, the text is digitally altered, making storage, modification, and retrieval more easier. Results show that optical character recognition (OCR) has potential benefits for digital invoicing, with a precision rate of 60% or higher. One way in which automation saves time is by doing away with the need for humans to enter data. Reducing operating expenses linked to human data input is one sign that the previously outlined process is cost-effective.

Jain, P. H. et al. (2023) created a versatile optical character recognition (OCR) model framework by applying OCR methods on a separate dataset of English numerical digits. Using the MNIST dataset, the researchers trained their model using a convolutional neural network. Results from models trained with both MNIST and custom digits were comparable to those from their single model. Optical character recognition (OCR) for personalized handwriting is another area that their ideas try to improve. In order to measure the efficacy of classification algorithms, the researchers use commonly used measures like Accuracy and F-1 score to evaluate their classification models. With validation, the model achieves an accuracy of 99.18% and overall accuracy of 97.80%.

Using Python, a web programming language, and Natural Language Processing (NLP), Yan et al. (2022) investigated the creation and implementation of a system to automatically evaluate tests. To determine how similar phrases and keywords are in Chinese text content, the system employs crucial approaches such knowledge extraction, syntactic dependency analysis, and lexical assessment. This makes it possible for online exams to automatically score subjective answers. The TF-IDF keyword extraction technique is used to identify the word frequency in each phrase, which is then used to extract keywords from the paragraphs. phrase segmentation is the method that is used to apply this strategy. More than that, it uses a scoring mechanism to determine similarity, and that method was accurate to within 70%. The Python web framework is fundamental to the development of the product.

The supplemental function of optical character recognition (OCR) within the framework of efficient supply chain management has been evaluated by Shahin, M et al. (2023). This study aims to provide an analysis of how MBID could enhance a Lean 4.0 competitive manufacturing process. In this work, we admit that the overall low resolution of each image is an intrinsic restriction of the experimental sample, which comprised of only 200 side view photographs. Overall detection accuracy is 95% and the Character Error Rate (CER) is 0.0041, demonstrating the validity and efficiency of the suggested method. All things considered, the proposed system achieved a 95% detection accuracy and a Character Error Rate (CER) of 0.0041. One way to make the proposed method even better is to use deep learning models like CNN to figure out how to identify the characters and get feature representations on their own.

The use of a Convolutional Neural Network (CNN) model for the successful detection of numerical values and handwritten typography in photos is demonstrated by Mishra, A. et al., (2023). The approach was trained using a dataset that contained more than 100,000 images captured by the National Institute of Science and Technology (NIST). If the model is applied to handwritten text that differs significantly from the NIST dataset, its performance can be compromised. Key input image properties, like blurriness or low resolution, can substantially impact the algorithm's efficacy. Training computational neural networks (CNNs) allows for the identification of typographic styles. Increasing the model's accuracy by using more complex CNN architectures or additional datasets. The development of multilingual handwriting recognition technologies capable of reading several scripts and languages, such as Chinese and Arabic. With a margin of error of 2.53%, the authors achieved a precision of 90.54%.

Johri, E et al., (2021) develop a system called ASSESS that makes it simpler and easier to evaluate subjective answers on exams. The suggested system automatically evaluates students' subjective responses and provides feedback so they can get better. It does this using natural language processing and semantic learning. Additionally, it offers text-to-speech and speech-to-text accessibility options, visual statistics for teachers and students to review after each test, and a fully functional hands-free mode for students with disabilities like slow typing, impaired vision, and amputated hands. The system provides a solitary platform for testing, evaluation, and feedback for both teachers and students. It offers a dashboard with charts to analyze the performance of each student individually or collectively. This system has the benefit of being almost finished, performing better, and serving a sizable audience.

Bashir, M. F et al., (2021) discusses the difficulties of utilizing artificial intelligence to evaluate subjective papers and suggests for a remedy that employs solution statements and keywords to assess responses. The accuracy of the machine learning model that the authors developed to predict answer grades was 88% without the MNB model, and it was increased by 1.3% with MNB. The purpose of the paper is to offer a more effective method of addressing the issue of subjective answer evaluation. Existing research typically have sentences that are randomly arranged. Extensive training of the algorithm on an adequate quantity of data provides the benefit of serving as a confidence booster for the Result Prediction Module. Utilising large data sets allows for a significant augmentation in the number of classes or grades included in the model. The word2vec model can be tailored for the selective assessment of subjective responses in a specified domain. Based on this small dataset, a machine learning model achieves an accuracy rate of 86%.

Amur, Z. H et al., (2022) proposed the effectiveness of the learner is assessed using these systems based on their intellectual and cognitive abilities. The study incorrectly assumed the response to How-many questions and 100 cases were mistakenly forecasted. The proposed BERT model is assessed and shown to be useful for automatically grading brief answers. It was unable to recognize the factoids' response. Students can quickly remedy their mistakes and perform better thanks to the system's prompt feedback. Determining how effectively students can comprehend and meet the learning objectives on a cognitive level depends on the evaluation method. By using this data which is logistic regression, the study added 1636 features, more than it trained for. The matrix multiplication method utilized in the study, softmax functions, did not result in improved performance. Precision, Recall, Specificity, and F1 are all 0.95%, 0.78%, and 0.96% respectively.

Singh, S et al., (2021) focuses on enhancing the correction of handwritten subjective answers. To achieve this, the authors propose the usage of the RAKE technique, combined with a Word2vec model, to extract keywords and phrases from the answers. The process involves converting PDF files to text files using Optical Character Recognition (OCR) and then using a Python class called tesseract to extract the answers in text format. The next step in the evaluation process involves Natural Language Processing (NLP) tasks. The answers are preprocessed using techniques such as spell check, tokenization, POS tagging, and lemmatization. After preprocessing, the answers are compared and marks are calculated using RAKE method. The accuracy of the model is reported to be 95% without text extraction and 93% with text extraction.

Sinha, S. K. et al. (2022) introduce a computational framework for evaluating response scripts by harnessing natural language processing techniques. In the initial phase of the procedure, algorithms are used to categorize data according to keywords in order to generate a succinct summary of the collected data. The determination of the final grade is based on four similarity criteria, namely Cosine. Using answer scripts, Natural Language Processing

(NLP) can be used to extract text. In order to enhance the efficiency of text processing, specific terms are eliminated from the condensed text.

TK's is used to create bigrams, which are collections of two adjacent objects from a string of tokens. The frequency distribution is used to determine text structural similarity. This model was tested on various datasets, and the accuracy ranged from 0.71 to 0.85. Due to the unavailability of a large dataset, an unsupervised approach was chosen for this paper.

Proposed Methodology

This Paper methodology entails employing Natural Language Processing (NLP) to build an automated grading system for subjective exam papers. To turn handwritten responses into digital text, optical character recognition (OCR) will be used to process them. The text will next be analyzed and understood using NLP techniques in order to extract pertinent information. The system will evaluate the responses using predetermined assessment criteria, and Natural language processing techniques will help for text preprocessing with calibration. To make sure that the grading is correct and consistent, we measure the different similarity scores and provide the best score for the given answer sheet.

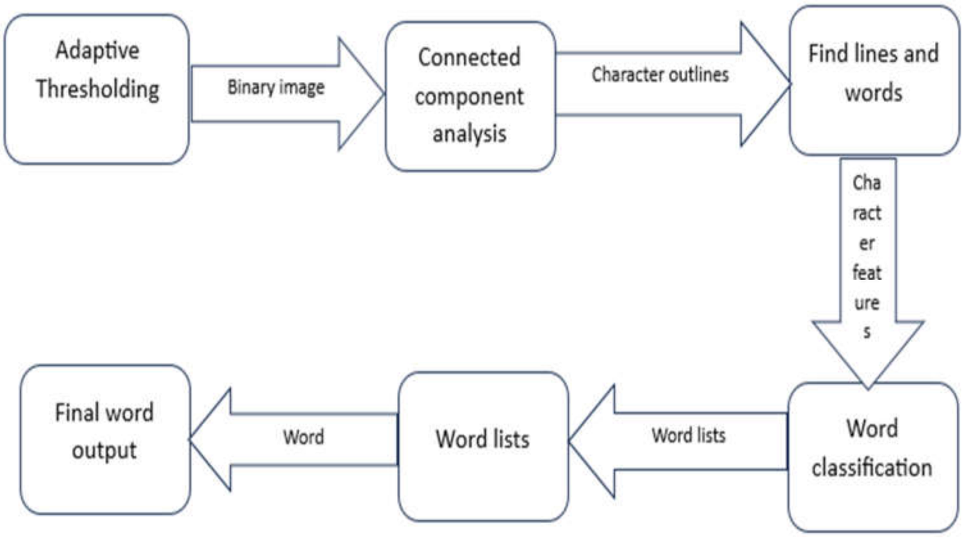


Figure 1. Tesseract Architecture

The following steps are followed for the Digital Grading:

- Uploading the Answer Sheet and the Question Answer Key Sheet
- Extracting the Answer sheet text and Question Answer Key text using OCR (Tesseract model)
- Performing Text Preprocessing of both the question key and answer key text using NLP tecniques.
- Calculating the score using different Similarity Metrics

Step-1: In this Step the answer sheet and Question Answer Key must be uploaded for the evaluation process. It can be offered in a variety of formats, including documents, PDFs, text files, Excel spreadsheets, and even image files. This adaptability makes it simple to integrate and evaluate responses in a variety of formats.

Step-2:

After uploading both the answer sheet and question answer key sheet the text can be extracted using the different libraries and packages like PyPdf2 and openpyxl. Now we discuss in brief about how the text is extracting

PyPDF2: Python's PyPDF2 package is used to extract and work with data from PDF files. With PyPDF2, you may divide, combine, and extract text and metadata from PDF files, among other PDF-related operations. A PDF file must normally be opened, read through its contents, and then the needed data can be extracted.

openpyxl: Open and modify Excel files, particularly those with the XLSX extension, with the help of the Python function openpyxl. In other words, it's a tool for working with Excel spreadsheets. With "openpyxl," you can access the formulas, cell values, and formatting components of a particular Excel sheet, enabling you to retrieve data from the workbook.

DOCX: Word documents (DOCX) can be read and edited using the python-docx module, which is called the "extract_text_from_docx" Python function. To begin, the process gets the provided DOCX file, sets the text variable to null values, and then adds the written content to each paragraph one after the other. An easier way to automate text extraction from DOCX files using Python is to concatenate the gathered textual information into a string.

Text file: To get text from a given text file (or "txt_path" in Python), you can use the "extract_text_from_txt" method. A read-only file is accessed, its contents are extracted, and stored in the 'text' variable. Appropriate file management and prompt application termination are guaranteed by this method upon implementation. To efficiently and conveniently extract text from TXT files, this technique is provided to Python developers. The function's end result is the extracted material.

Tesseract Model (OCR): A free software tool called Tesseract uses optical character recognition (OCR) to extract text from photographs. One method of automating the process of text extraction and conversion from images is known as optical character recognition (OCR). Use this method on scanned documents, digital document copies, scene images, or photographs with text superimposed on top of them.

Step-3:

This phase involves applying several preprocessing techniques to the collected text. Allow me to provide a brief overview of each strategy now.

Tokenization: "Tokenization" captures the essence of this approach as a method of rendering huge texts more manageable. A token could be as little as a single letter, as large as an entire sentence, or as little as a group of words. One of the most important steps in NLP and text analysis is tokenization, which helps computers understand and interpret user-supplied text.

Stop Words removal: Stop Exclusion of words: Stop words are successfully decreased when unnecessary and unneeded terms like and, the, and in are removed from a work. It is feasible to remove terms that don't add anything to make text analysis more precise and less noisy. The initial stage of processing natural language is usually to remove "stop words" so that the computer can focus on the important parts of the text and work less.

Punctuation Removal: Erasing all punctuation from a text, like commas, periods, and exclamation points, is called punctuation removal. This process of preprocessing text data makes it easier to study by ignoring punctuation, which is often less important for meaning than other parts of speech.

Stemming: Stemming is a way to make writing more consistent by breaking words down to their root, or most basic form. Even if the term isn't a word by itself, it gets rid of endings or beginnings that make words different to make a single language stem.

Lemmatization: Lemmatization is a text normalization approach that breaks down words into their lexical or root form, or lemma. Lemmatization guarantees that the final word is a recognized word in the language, as opposed to stemming.

Spellchecker: A spell checker is a piece of software that finds and fixes misspelt words in text documents. It compares words to a dictionary of perfectly spelt terms that has already been defined and offers corrections for words that don't match any entries.

One should perform these techniques on both answer text and Question answer key text.

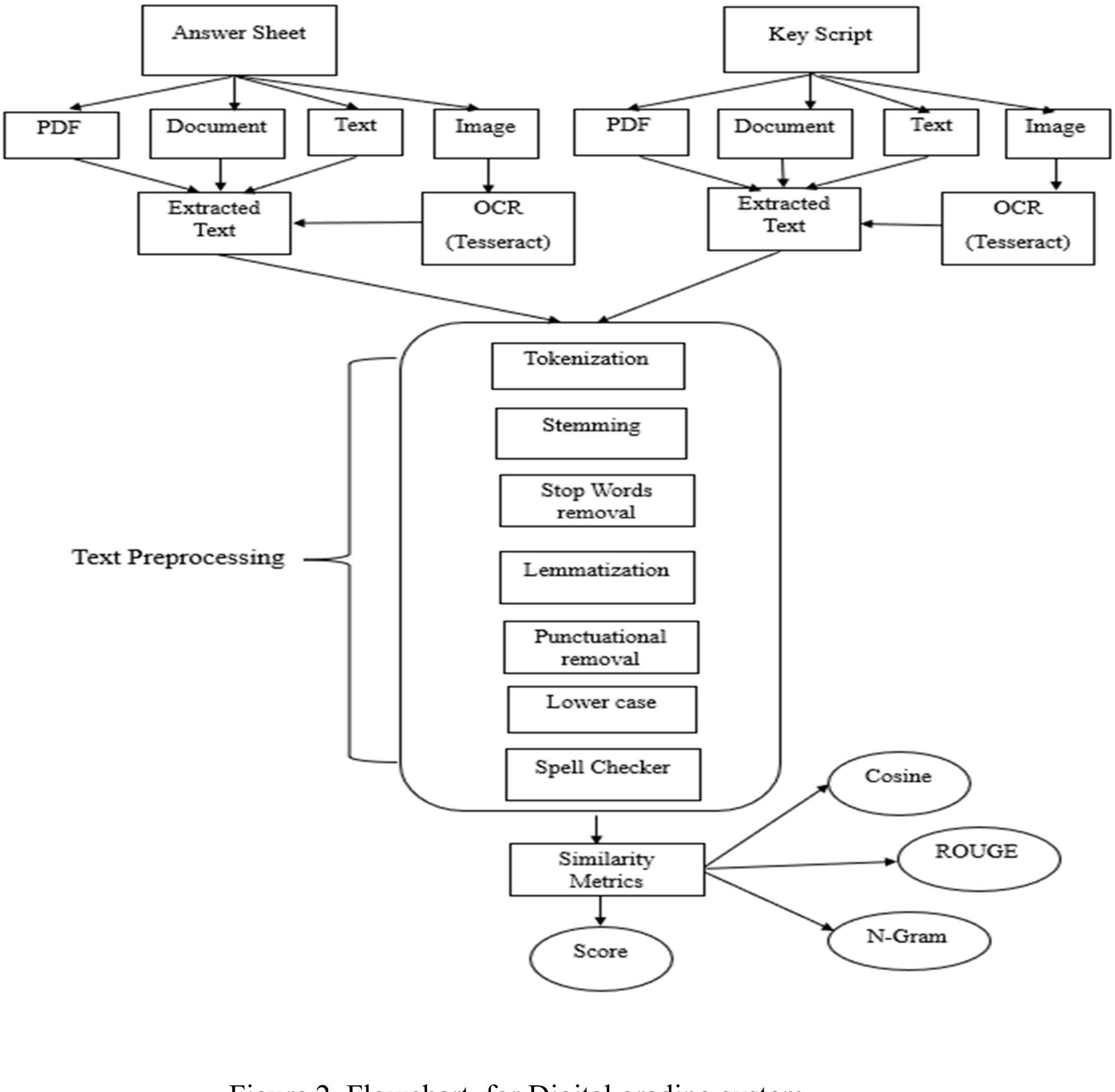


Figure 2. Flowchart for Digital grading system.

Step-4:

BLEU score: The quality of machine-generated text is measured using the BLEU (Bilingual Evaluation Understudy) score, especially in translation jobs. It compares the generated text's precision of N-grams (contiguous word sequences) to the reference text, which is often a human translation. The range of a BLEU score is 0 to 1, with 1 denoting a perfect match with the reference text.

$$\log BLUE = \log BP.exp\left(\sum_{n=1}^{N} w_n\, logp_n\right)$$

$$= \log BP + \log exp\left(\sum_{n=1}^{N} w_n\, logp_n\right)$$

$$= \min\left(1 - \frac{r}{c}, 0\right) + \left(\sum_{n=1}^{N} w_n\, logp_n\right)$$

ROUGE score: It is a set of metrics and measures used to assess how well machine-generated text or summaries compare to reference or human-generated texts in terms of quality and resemblance. Machine learning and natural language processing (NLP) applications like text summarization and translation frequently use ROUGE. ROUGE-N measures the overlap of n-grams.
The below formula used

ROUGE-N (candidate, references) = m a x k {ROUGE-N single (candidate, references k)}

BERT score: The quality of resemblance between two texts is measured using the BERT Score. The BERT (Bidirectional Encoder Representations from Transformers) model's calculation measures how similar the two texts' embeddings are to one another.

Bag of words Score: It is a quick way to gauge how similar two texts are. In order to compare these word counts, it is necessary to tally the frequency of each word in each manuscript. The BoW Score increases with word frequency overlap, indicating more similarity between the texts.

Accuracy: Measurements, predictions, and outcomes are considered accurate to the extent that they match the true or expected value. It is a gauge of how accurate or faultless a specific procedure or system is.

Accuracy = (TP+ TN) / (TP+ TN + FP + FN)

Results and Discussion

The proposed model was implemented using the Tesseract Model which is used for optical character recognition. This model is better for text recognition compare to other models like CNN, DNN and some other large language models. Now, the study uses Rouge, a unique measurement rarely used in other research papers, to investigate a wider range of assessment criteria and learn more about the functionality of natural language generation models.

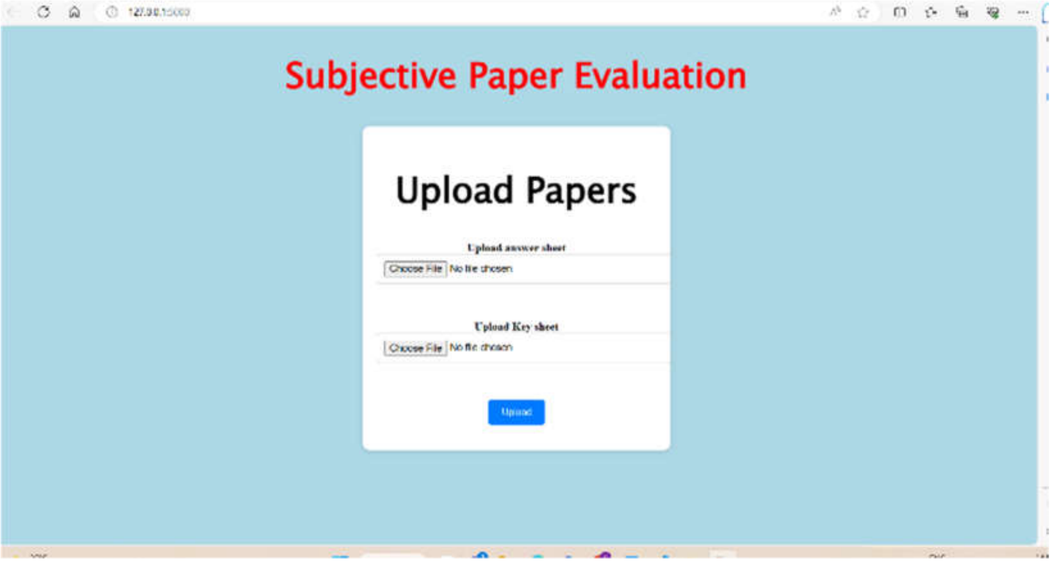
We can observe as the below table shows that ROUGE (93%) is the best performing metric for evaluating the quality of machine-generated text, followed by BLEU (78%), BERT (75%), and BOW (83%). This is probably because ROUGE is a contextual language model, which considers the meaning of the full phrase when assessing the effectiveness of a response. Although they are context-aware metrics, BERT and BLEU are not as advanced as ROUGE. BOW is a straightforward measure that counts the words in a response but ignores the sentence's context.

Table.1. Various Proposed metrics scores

Question Number	**Proposed Metrics**				
	ROUGE Score	**BOW Score**	**BERT Score**	**BLEU Score**	**Accuracy**
Q1	93.43	90.72	85.613	73.09	95.71
Q2	94.91	83.84	70.371	82.36	96.91
Q3	94.612	90.66	70.725	79.43	95.61
Q4	88.312	76.02	75.833	73.29	89.42
Q5	92.861	80.92	79.23	78.69	94.32
Q6	92.80	77.98	64.89	78.27	93.81
Q7	94.64	87.81	78.37	82.25	94.23
Q8	93.64	86.81	77.32	83.21	95.35
Q9	93.80	71.60	82.55	78.80	94.62
Q10	93.48	88.81	70.28	79.01	95.43
Avg	93.24	83.52	75.51	78.83	94.54

The accuracy was calculated by comparing the model score to the manual correction score. The model finally achieved an overall performance of 94%.

For uploading the papers and to measure the score we had deployed our model using the flask (Python) which can be used as handson by the faculty to automatucally socre the marks.



Subjective Paper Evaluation

Upload Papers

Upload answer sheet

Choose File No file chosen

Upload Key sheet

Choose File No file chosen

Figure 3. Deployment for the model

The accuracy of my model in comparison to other models is now clearly illustrated in the graph below.

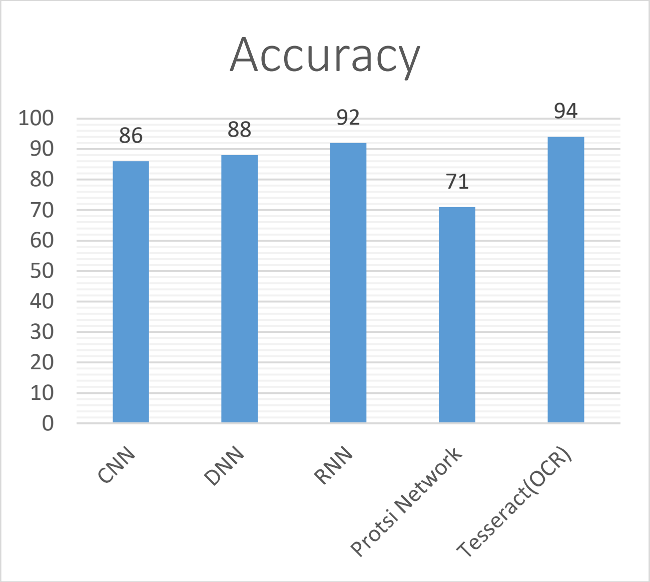


Figure 4. Comparative analysis of models

Conclusion

In this paper, we created a Digital Grading System that uses automation and Natural Language Processing (NLP) to transform paper-based assessment. Our main contributions include developing an optical character recognition system using Tesseract for extracting text from image files, providing flexibility with PDF, Word, and Excel sheet formats for answer evaluation, and making use of cutting-edge metrics like ROUGE, Bleu, Bert, and Bow scores for scoring accuracy. We are dedicated to improving the system's efficiency in order to achieve even better outcomes in the future, even as we solve issues with multiple-choice questions and handwriting styles.

References

Amur, Z. H., Hooi, Y. K., & Soomro, G. M. (2022). Automatic Short Answer Grading (ASAG) using Attention-Based Deep Learning MODEL. In 2022 International Conference on Digital Transformation and Intelligence (ICDI) (pp. 1-7). IEEE. https://doi.org/10.1109/ICDI57181.2022.10007187

Bashir, M. F., Arshad, H., Javed, A. R., Kryvinska, N., & Band, S. S. (2021). Subjective answers evaluation using machine learning and natural language processing. IEEE Access, 9, 158972-158983. https://doi.org/10.1109/ACCESS.2021.3130902

Jain, P. H., Kumar, V., Samuel, J., Singh, S., Mannepalli, A., & Anderson, R. (2023). Artificially Intelligent Readers: An Adaptive Framework for Original Handwritten Numerical Digits Recognition with OCR Methods. Information, 14(6), 305. https://doi.org/10.3390/info14060305

Johri, E., Dedhia, N., Bohra, K., Chandak, P., & Adhikari, H. (2021, May). Assess-automated subjective answer evaluation using semantic learning. In Proceedings of the 4th International Conference on Advances in Science & Technology (ICAST2021). http://dx.doi.org/10.2139/ssrn.3861851

Kamisetty, V. N. S. R., Chidvilas, B. S., Revathy, S., Jeyanthi, P., Anu, V. M., & Gladence, L. M.

(2022, March). Digitization of Data from Invoice using OCR. In 2022 6th International Conference on Computing Methodologies and Communication (ICCMC) (pp. 1-10). IEEE. https://doi.org/10.1109/ICCMC53470.2022.9754117

Lu, Y., Qiu, J., & Gupta, G. (2022). ProtSi: Prototypical Siamese Network with Data Augmentation for Few-Shot Subjective Answer Evaluation. arXiv preprint arXiv:2211.09855. https://doi.org/10.48550/arXiv.2211.09855

Mishra, A., & Ram, A. S. (2023). Handwritten Text Recognition Using Convolutional Neural Network. arXiv preprint arXiv:2307.05396. https://doi.org/10.48550/arXiv.2307.05396

Shahin, M., Chen, F. F., & Hosseinzadeh, A. (2023). Machine-based identification system via optical character recognition. Flexible Services and Manufacturing Journal, 1-28. https://doi.org/10.1007/s10696-023-09497-8

Singh, S., Shah, Y., Vajani, Y., & Dholay, S. (2021). Automated Paper Evaluation System for Subjective Handwritten Answers. In 2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT) (pp. 1-6). IEEE. https://doi.org/ 10.1109/ICCCNT51525.2021.9579912

Sinha, S. K., Yadav, S., & Verma, B. (2022, March). NLP-based automatic answer evaluation. In 2022 6th International Conference on Computing Methodologies and Communication (ICCMC) IEEE. https://doi.org/ 10.1109/ICCMC53470.2022.9754052

Weegar, R., & Idestam-Almquist, P. (2023). Reducing Workload in Short Answer Grading Using Machine Learning. International Journal of Artificial Intelligence in Education, 1-27. https://doi.org/10.7910/DVN/NMEM7D

Xu, F., Song, Y., Iyyer, M., & Choi, E. (2023). A critical evaluation of evaluations for long-form question answering. arXiv preprint arXiv:2305.18201. https://doi.org/10.18653/v1/2023.acl-long.181

Yang, W. (2022, December). Design and Implementation of Automatic Examination Scoring System Based on Natural Language Processing. In 2022 3rd International Conference on Big Data and Informatization Education (ICBDIE 2022) (pp. 1048-1058). Atlantis Press . https://doi.org/10.2991/978-94-6463-034-3_107

Yue, X., Wang, B., Zhang, K., Chen, Z., Su, Y., & Sun, H. (2023). Automatic evaluation of attribution by large language models. arXiv preprint arXiv:2305.06311. https://doi.org/ 10.48550/arXiv.2305.06311

Zhao, S. (2022). Evaluation of English Subjective Questions Based on Deep Neural Networks. Scientific Programming, 2022. https://doi.org/10.1155/2022/1225634