**A Review on Automated Explanatory Answer Evaluation Techniques.**

Sunny Shekhar Singh, Harisha Naik T

Department of Computer Applications, Presidency College (Autonomous)*,* Bengaluru, Karnataka, India

\***Address for Correspondence**

**Sunny Shekhar Singh**

Master of Computer Applications,

Presidency College (Autonomous)*,*

Bengaluru, Karnataka, India

Email: sunnyshekharsingh@gmail.com

**Harisha Naik T,**

Associate Professor

Department of Computer Applications,

Presidency College (Autonomous)*,*

Bengaluru, Karnataka, India

Email: harishtkola@gmail.com

|  |  |
| --- | --- |
| **Yashoda ABSTRACT** |  |

Institutions and all educational bodies have a process of evaluating answers for their tests and exams conducted by them. This process is time consuming and limits the test frequency as the man power needed for the evaluation is never enough. Since there are more people involved in checking multiple copies of different students, it can be said the way of different people(teachers) are different and hence the marking system will also change or vary. Factors like mindset, emotions, familiarity, etc add as bias and creates a biased decision, favouring individual(s).

Our model, on the other hand provides solution to many problems and helps increase the speed of the process while maintain standards which are same for everyone. These standards help remove bias and provides unbiased system. The model works for both online and offline exam process. The model verifies the answers by comparing with the stored and pre-defined answers/keys. Assessment tools help automate the evaluation system. this method reduces the workload from the people(teacher) and helps optimize the usage of resources.

**Keywords:** Text Mining, Machine Learning, The Automatic Assessment Tools, Lexical Analysis, Semantic Analysis, n-gram, Natural Language Processing, Pre-processing, Active Learning.

# INTRODUCTION

Conducting examination(s) is a necessity for every type of educational body. Exam or test is a method to evaluate the standards and limits of evaluate student.

It can help a understand the areas where improvement is required. The challenge for the smooth functioning of the examination system lies in the second half of the process. Evaluation of the answers written by the students is the main and final problem in the examination process. The descriptive nature of our prevailing examination system can't be overlooked. With the advancement of technology, we have achieved a good system for Objective based test. OMR (Optical Mark Recognition) is a technology used to detect and read marks made on paper.OMR is a reliable and efficient method for processing large quantities of data.

Evaluation of descriptive answers is different from Objective answers. Technologies are developing and we have a few approaches to this problem, but there is a need to advance and discover new efficient methods for descriptive answer evaluation.

# LITERATURE REVIEW

The Proposed system (Analysis of Descriptive Answer Evaluation Process using NLP Techniques) may help us understand the problem better as it divides the problem into different sections and works on individual problems to attain highly effective and efficient results. Keyword Matching, a process used here which is takes all the keywords as input for the answers and compare it with the keywords of the answers written by the student(s). The other technique employed is Semantic Analysis which considers word meanings and relationships within the context of the question and answer. Information Extraction and Entity Recognition helps us better understand and analysis the contents and reasoning. a Discourse Analysis and Text Structure: Machine Learning and Neural Networks: Addressing the bias is a strong aspect of the proposed model.

## The Auto Tutor

The tutoring research groups from the University of Memphis designed this tool [1]. It is complete automatic tool. It helps to students to learn subjects such as computer hardware, operating system as well as internet. It provides question as well as answers to the students and receives the answers from the standard input.

## The Apex

This is web-based application it will rate learner’s answers by referring the answer keys stored in the system [1]. The students can choose the topic and questions and they can answer the questions, the answers will then be compared with the stored answers. LSA is used to measure the semantic similarity.

## The Bilingual Evaluation Understudy (BLEU Algorithm)

This method is introduced by Papineni *et al* [1]. This is uses n-gram co-occurrence scoring method. The ideal of this method is measuring the closeness of translation as well the translation of reference by the use of numeric metric. Here n indicates sequences of words that can be applied to compare two different texts. The machine translates the input sentences. Among the machine translation and reference translation n-gram matching will be counted.

## The ERB (Evaluating Responses with BLEU)

Introduced by Perez, Alfonseca and Rodziguez in 2004. This uses an algorithm which is related to BLEU also uses a set of NLP technology. The central idea of this algorithm is similar to that of BLEU. The students answer and the reference answer are more alike and hence the score is high.

## ATENEA

This was developed by Perez et al. [1]. This method is related to BLEU algorithm. It can evaluate answers in English as well as Spanish. It can evaluate the answers at the same time it enables the students to personalize the user interface as per their need. It enables the students to select a question for that question they have choice to answer in any of the language, English and Spanish are preferable. During question answering the answers will be compared with the stored answer. It uses Natural Language Processing Techniques such as Stemming, Word Sense Disambiguation etc.

**The Auto Mark**

This model was developed by Mitchell et al [1]. It is used to access the Subjective Answers. Auto-marking, also referred to as automatic grading, utilizes NLP (Natural Language Processing) techniques to analyse and score student responses, particularly in open-ended assessments. It permits the candidate for setting reference making schemes. It provides freedom to the candidates for making schemes. It will correct spelling mistakes automatically when it is pre-processing the answers, that relives students to worry about spelling mistakes. Then the Sentence Analyser is used to identify the phrases as well as relationship among them. Then the Pattern Matching Module used to identify the similarity among the reference and the candidates answer. Finally, it offers a score to the candidates for their answer.

**Keyword Matching:**

Simple approaches rely on matching keywords or synonyms present in the answer key with the student's response [2]. While efficient, this can miss nuances and penalize alternative phrasings.

**Semantic Similarity**:

More modern procedures use semantic similarity measures, taking into account word implications and connections inside the context of the questions and also the answers [3]. This further develops precision however may experience hardships with metaphorical language or space explicit wording

**Information Extraction and Entity Recognition**:

Extracting key concepts and entities from both the answer key and student response allows for deeper analysis of content and reasoning [4]. This method can identify partially correct answers or misconceptions but requires substantial domain knowledge and training data.

**Discourse Analysis and Text Structure**:

Analyzing the overall structure and coherence of the descriptive answer can provide insights into the student's thought process and organization skills [5]. This requires advanced NLP techniques and may not be applicable to all answer types.

**Machine Learning and Neural Networks**:

Supervised learning approaches train models on large datasets of manually evaluated answers to automatically score new responses [6]. While promising, these models rely on quality training data and can be susceptible to bias.

## The C-rater

This is designed by the Education Testing Service (ETS). The input answer is pre-processed to correct spelling as well as grammar mistakes. Next it will be subject to POS (Part of Speech) tagged for removing ambiguity. The Feature Extraction technique is used extract relations among the predicates from the given answer. The NLP tools are used to process the model answer. Then the matching algorithm is applied on the resultant answer. The Gold map is a rule-oriented Pattern Matching Algorithm. This algorithm offers a score, and this is the feedback to the candidates.

## The Text Mining

The Text Mining extracts essential features from different sources of unstructured form of text. It is not easy to deal with un-structured form of text. The aim of Text Mining is to extract vital information from the text.

## The QAL (Question Answer Language) and Indus Maker

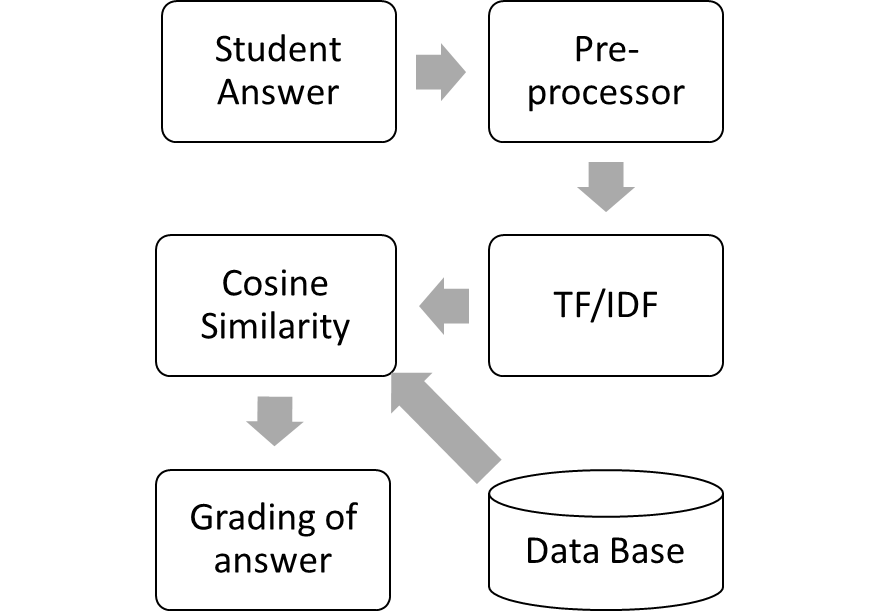
This method compares the answer text structure with the predefined structure. The structure is created the structure editor. The teachers state their required structure using a special language. This language was known as QAL. It was extended as language called Question Answer Markup Language (QAML].

## The Natural Language Processing (NLP) Techniques

NLP is used for understanding natural language of humans and process this input for various different applications using a variety of algorithms. Various Machine Learning training algorithms study millions of examples of text words, sentences, and paragraphs written by humans [10]. By studying the samples, the training algorithms gain an understanding of the “context” of human speech, writing, and other modes of communication. In NLP, there are two components NLG (Natural language generation) and NLU (Natural Language Understanding). In NLG, it generates text and speech synthesis to generate spoken output. In NLU, it extracts information from and speech recognition to get information from speech. Text classification refers to the classification of texts into several categories according to a certain classification basis. With the advent of the information age, a large amount of information is constantly emerging, followed by a large number of texts. Therefore, classifying texts to better process texts and obtain effective information has become one of the current problems. From another perspective, text classification also belongs to one of the application directions of natural language processing.

## Proposed methodology

The proposed methodology undergoes a series of phases to compute the best possible result(s). The processes like pre-processing or cleaning of the dataset, tf-idf, cosine similarity, answer grading, etc. is used in the proposed methodology. The proposed system is developed to counter the challenges faced by the institutes and educational bodies of grading the descriptive answers written by the students on large scale. The proposed system should also help maintain the unbiased nature of the examinations. And the system flow, in fig.1, is demonstrated.



**Fig. 1. Analysis on Subjective Answers**

**Descriptive Answers (Input)**

The answers written by the students are stored in the database after scanning. This process eliminates the physical dependency of the answer sheet Answer sheets are no longer required to continue the process.

## Preprocessor

This process is required for the normalization of the text that includes eliminating unwanted words, stemming, etc.

Pre-processing alludes to the changes applied to our dataset prior to taking care of it to the algorithm.

Data Preprocessing is a method that is used to convert the raw data into a data set which is organized and clean. In other words, whenever the data is collected from different sources it is collected in underdone format which is not reasonable for the analysis.

## Need of Data Preprocessing

For achieving better results from the applied model in Machine Learning projects the format of the data must be in a proper manner. Some specified Machine Learning model needs information in a specified format, for example, Random Forest algorithm does not support null values; therefore, to execute random forest algorithm null values must be managed from the original raw data set. Another aspect is that the data set should be formatted in such a way that more than one Machine Learning and Deep Learning algorithm are executed in one data set, and the best out of them is chosen.

Utilising the above process(s), we move forward to develop the methodology used in the proposed model.

## TF-IDF

Tf-idf stands for the term frequency-inverse document frequency, and the tf-idf is often used in information retrieval and text mining.TF-idf can be successfully used for stop-word filtering in various subject fields including text summarization and classification.

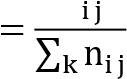
**TF-IDF** stands for “Term Frequency — Inverse Data Frequency”.

**Term Frequency (tf)**

Term Frequency (TF): This calculates how often a word appears in a particular document. The more a word shows up, the higher its TF score.[7] Intuitively, words that come up more frequently in a document are likely more relevant to that document's content.

Each document has its tf.

The formula can be computes as:



tf

**Inverse Data Frequency (idf)**

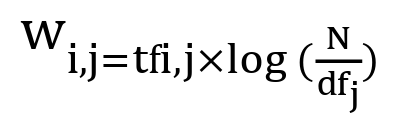
This part measures how common a word is across all the documents in the corpus. Words that appear very frequently across all documents (like "the" or "a") will have a low IDF score because they aren't very distinctive. Conversely, words that appear only in a few documents (like technical terms) will have a high IDF score since they are more specific. In the proposed model we have taken a total of 5 questions for analysis; each question carries 5 marks, and 3 model answers are prepared for each question. These 5 questions are given to each student. The total number of students is 10. Once the answers are submitted by students then it is evaluated by both system and the human assessor by comparing student answers with the 3 model answers.



**Average Inverse Data Frequency (A-idf)**

The proposed model introduces a new method Average-Inverse Data Frequency(A-idf) that will not only have one human assessor but three assessors to grade the answers written by the students. In this case, an average of the marks graded by the three assessors is considered. This helps the system to gain preciseness and avoid rigidity and biasness persisting in the previous models. The proposed model generates more accurate and reliable results.

Combining these two we come up with the TF-IDF score (w) for a word in a document in the corpus. It is the product of tf and idf:



tfij = number of occurrences of i in j

Dfij=number of documents containing i

N= total no of documents

## Cosine similarity

Cosine similarity is a metric used to measure how similar the documents are irrespective of their size. Mathematically, it measures the cosine of the angle between two vectors projected in a multi-dimensional space. The cosine similarity is advantageous because even if the two similar documents are far apart by the Euclidean distance (due to the size of the document), chances are they may still be oriented closer together. The smaller the angle, the higher the cosine similarity.

**Knowledge-Based**

The knowledge- base contains the information and the standard answer and the marks that is assigned for each question.

## Grading of Answer

After the completion of the above-described processes, grading of the answer is done according to the content of the students. Marks are assigned as per the grade of answer. The model evaluates the answers written by the students according to the three model answers previously stored in the system. After comparing both model answers and the student’s answer The Answer Evaluation comparison table is now maintained carefully by writing the marks scored by the students for the answers written. The table reflects the marks obtained by students for each question from both computer and the human assessors. An average of the marks graded by the three assessor is calculated. The marks allotted by the model is then compared to the average marks, we observe the evaluation is done almost similar.

# RESULTS AND DISCUSSION

To Analyse the performance of this proposed model, the system assign marks will be compared with the human assign marks for the same answers of each student. For that, we use NLP techniques for the experiment process. Here we have taken the subjective type of questions for the experiment process. Hence these answers are verified in detail, by checking the spellings and grammar of the sentence in student answer. Table 1 provides the detailed data of grades allotted by the human assessor and the proposed model. The results obtained by the proposed model outperforms all the previous models and methodologies. The difference (Average-Computer Analysis) gives a difference between the Human Analysis and the analysis made by the proposed model (Table 1). The difference is negligible and the model runs in sync with Human Analysis.

**Table 1: Comparison table between Human Analysis and Computer Analysis.**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Students** | **Answers** | **Human Analysis** | **Human Analysis** | **Human Analysis** | **Average** | **Computer Analysis** | **Difference** |
| Student 1 | Answer 1 | 4 | 3 | 4 | 4 | 4 | **0** |
| Answer 2 | 5 | 4 | 3 | 4 | 4 | **0** |
| Answer 3 | 3 | 4 | 4 | 4 | 4 | **0** |
| Answer 4 | 2 | 2 | 2 | 2 | 2 | **0** |
| Answer 5 | 5 | 4 | 4 | 5 | 4 | **-1** |
| Answer 6 | 1 | 2 | 1 | 2 | 1 | **-1** |
| Answer 7 | 2 | 2 | 2 | 2 | 2 | **0** |
| Answer 8 | 5 | 4 | 5 | 5 | 5 | **0** |
| Answer 9 | 4 | 3 | 1 | 3 | 3 | **0** |
| Answer 10 | 3 | 4 | 3 | 4 | 3 | **-1** |
| Total | 34 | 32 | 29 | 32 | 32 | **0** |
| Student 2 | Answer 1 | 5 | 3 | 4 | 4 | 4 | **0** |
| Answer 2 | 2 | 3 | 2 | 3 | 2 | **-1** |
| Answer 3 | 3 | 3 | 4 | 4 | 3 | **-1** |
| Answer 4 | 4 | 4 | 3 | 4 | 4 | **0** |
| Answer 5 | 5 | 5 | 5 | 5 | 5 | **0** |
| Answer 6 | 4 | 3 | 4 | 4 | 4 | **0** |
| Answer 7 | 5 | 4 | 5 | 5 | 5 | **0** |
| Answer 8 | 2 | 2 | 4 | 3 | 3 | **0** |
| Answer 9 | 4 | 3 | 4 | 4 | 4 | **0** |
| Answer 10 | 3 | 2 | 4 | 3 | 3 | **0** |
| Total | 37 | 32 | 39 | 36 | 37 | **1** |
| Student 3 | Answer 1 | 3 | 4 | 4 | 4 | 4 | **0** |
| Answer 2 | 4 | 3 | 3 | 4 | 3 | **-1** |
| Answer 3 | 5 | 4 | 5 | 5 | 5 | **0** |
| Answer 4 | 4 | 3 | 4 | 4 | 4 | **0** |
| Answer 5 | 3 | 2 | 3 | 3 | 3 | **0** |
| Answer 6 | 0 | 0 | 0 | 0 | 0 | **0** |
| Answer 7 | 4 | 3 | 2 | 3 | 3 | **0** |
| Answer 8 | 3 | 4 | 3 | 4 | 3 | **-1** |
| Answer 9 | 4 | 4 | 2 | 4 | 3 | **-1** |
| Answer 10 | 4 | 5 | 4 | 5 | 4 | **-1** |
| Total | 34 | 32 | 30 | 32 | 32 | **0** |
| Student 4 | Answer 1 | 4 | 3 | 4 | 4 | 4 | **0** |
| Answer 2 | 5 | 4 | 5 | 5 | 4 | **-1** |
| Answer 3 | 5 | 5 | 5 | 5 | 5 | **0** |
| Answer 4 | 2 | 1 | 2 | 2 | 2 | **0** |
| Answer 5 | 3 | 2 | 3 | 3 | 3 | **0** |
| Answer 6 | 4 | 3 | 3 | 4 | 3 | **-1** |
| Answer 7 | 1 | 2 | 2 | 2 | 2 | **0** |
| Answer 8 | 2 | 3 | 3 | 3 | 2 | **-1** |
| Answer 9 | 4 | 3 | 1 | 3 | 3 | **0** |
| Answer 10 | 4 | 4 | 4 | 4 | 4 | **0** |
| Total | 34 | 30 | 32 | 32 | 32 | **0** |
| Student 5 | Answer 1 | 5 | 5 | 4 | 5 | 5 | **0** |
| Answer 2 | 1 | 0 | 2 | 1 | 2 | **1** |
| Answer 3 | 1 | 1 | 2 | 2 | 1 | **-1** |
| Answer 4 | 2 | 2 | 3 | 3 | 2 | **-1** |
| Answer 5 | 4 | 3 | 4 | 4 | 4 | **0** |
| Answer 6 | 5 | 4 | 3 | 4 | 4 | **0** |
| Answer 7 | 3 | 3 | 2 | 3 | 3 | **0** |
| Answer 8 | 4 | 3 | 4 | 4 | 3 | **-1** |
| Answer 9 | 5 | 4 | 4 | 5 | 4 | **-1** |
| Answer 10 | 4 | 3 | 4 | 4 | 4 | **0** |
|  | Total | 34 | 28 | 32 | 32 | 32 | **0** |

Challenges and Future Directions:

Addressing Bias and Fairness: NLP models trained on biased data can perpetuate unfairness in scoring, particularly for students from diverse backgrounds or writing styles [8]. Mitigating bias requires careful data selection, model interpretability, and human oversight.

Domain-Specific Adaptation: General-purpose NLP tools may not capture the nuances of specific subject areas. Domain adaptation techniques or specialized models are crucial for accurate assessment in different disciplines [9].

Integrating with Educational Practices: Seamless integration of NLP-based evaluation tools into existing learning management systems and grading workflows is essential for adoption by educators [9].

Interpretability and Explainability: Making automated feedback transparent and understandable for both students and instructors can enhance learning and foster trust in the evaluation process [10].

# CONCLUSION

In this study we have examined different techniques used for evaluating subjective answers. These approaches compare the candidates answers with standard descriptive answers. It verifies the candidates answer by matching it with the stored answers. The evaluation is performed automatically. It saves time and resources to a great extent. It relives the examiners all the difficulties which are encountered in the traditional valuation system. The semantic similarity-oriented approaches produce accurate results. The similarity measure techniques provide required similarities between the answers. Hence the Universities and other Educational Institutions can adopt these systems for evaluation purpose.

# REFERENCES

1. P. Rinchen, “Comparative Study of Techniques used for Automatic Evaluation of Free Text Answer,” *Int. J. Recent Innov. Trends Comput. Commun.*, vol. 2, no. 12, pp. 3839–3842, 2014.
2. Ms.  Shweta M.  Patil and Prof.  Ms.  Sonal Patil, “Evaluating student descriptive answers using Natural Language Processing,” International Journal of Engineering Research & Technology, https://www.ijert.org/evaluating-student-descriptive-answers-using-natural-language-processing
3. International Journal of Applied Engineering Research ISSN 0973-4562 Volume 9, Number 24 (2014) Paper code: 29420 – IJAER “Semantic Similarity- A Review of Approaches and Metrics” - Ms.D.Akila1 and Dr. C.Jayakuma
4. H. Ahmed, S. Hina and R. Asif, "Evaluation of descriptive answers of open ended questions using NLP techniques," 2021, pp. 1-7, doi: 10.1109/ICCIS54243.2021.9676405.
5. A Review of Machine Learning Algorithms for Text-Documents Classification Aurangzeb Khan, Baharum Baharudin, Lam Hong Lee\*, Khairullah khan - JOURNAL OF ADVANCES IN INFORMATION TECHNOLOGY, VOL. 1, NO. 1, FEBRUARY 2010
6. Five sources of bias in natural language processing- Dirk Hovy, Shrimai Prabhumoye - First published: 20 August 2021 https://doi.org/10.1111/lnc3.12432
7. Analysis of Descriptive Answer Evaluation Process using NLP Techniques - Harisha Naik T1\*, N Kartik1 and Hemanth K S - ISSN: 0976 – 0997 - Vol.13 / Issue 75 / December / 2022
8. Domain Adaptation in Natural Language Processing: Inaugural dissertation at the Ludwig Maximilian University of Munich- Marina Sedinkina from E Mangelinsk, Russian Federation Munich 2021
9. Learning Management Systems in Higher Education- Aabha Chaubey & Bani Bhattacharya, 2015
10. Yang, W., Wei, Y., Wei, H. et al. Survey on Explainable AI: From Approaches, Limitations and Applications Aspects. Hum-Cent Intell Syst 3, 161–188 (2023). https://doi.org/10.1007/s44230-023-00038-y)