Tool for Evaluating Subjective Answers using AI (TESA)

Shreya Singh
K. J. Somaiya Institute of Engineering and Information Technology, Sion, Mumbai
University of Mumbai, India
shreya.ks@somaiya.edu

Omkar Manchekar
K. J. Somaiya Institute of Engineering and Information Technology, Sion, Mumbai
University of Mumbai, India
omkar.manchekar@somaiya.edu

Ambar Patwardhan
K. J. Somaiya Institute of Engineering and Information Technology, Sion, Mumbai
University of Mumbai, India
ambar.p@somaiya.edu

Prof. Uday Rote
K. J. Somaiya Institute of Engineering and Information Technology, Sion, Mumbai
University of Mumbai, India
udayrote@somaiya.edu

Prof. Sheetal Jagtap
K. J. Somaiya Institute of Engineering and Information Technology, Sion, Mumbai
University of Mumbai, India
sheetaljagtap@somaiya.edu

Dr. Hariram Chavan
K. J. Somaiya Institute of Engineering and Information Technology, Sion, Mumbai
University of Mumbai, India
hariram@somaiya.edu

2021 International Conference on Communication information and Computing Technology (ICCICT) | 978-1-6654-0430-3/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICCICT50803.2021.9510119

Abstract*— Modes of evaluation can be bifurcated into two main types, namely, objective answer evaluation and subjective answer evaluation. The technique of objective answer evaluation is widely used for most entrance examinations. The reason behind doing so is that it tests the analytical as well as the reasoning ability of a student while also being extremely accurate and efficient. Examinations such as the GATE exam are known to have never repeated questions. Although this technique is liable, it is close to impossible to implement such a system at the institute level for evaluating the semester exams of engineering students. One of the major limitations of the objective examinations is that it fails to analyze how well a student has grasped a particular subject. There is hence a need for a system that automizes the task of repetitive answer sheet corrections and provides optimal accuracy. This proposed model has come up with a reliable system based on previous work that formulates the marks scored by the student based on the sentence similarity, Jaccard similarity, and grammar of the model answer and student answer.

Keywords—Natural Language Processing (NLP), Bidirectional Encoder Representations from Transformers (BERT), Sentence Similarity, Optical Character Recognition (OCR).

I. Introduction

Every year about 20 Lakh engineering students give their semester subjective written examinations, which accounts for more than 1.2 crores papers to be corrected every semester. These tests consist of multiple-choice questions that include explanations as responses. Subjective questions like these are the most effective way to assess a student's understanding and play an important role in determining how well a student has understood a subject.

However, the correction of these answers can be a tedious and tiresome task for the examiner. Additionally, the marks allotted to these answers may vary from examiner to examiner which leads to inconsistencies in the correction. This raises a concern to have an automated system that eliminates bias in correction while reducing the time and effort put into it. It shall also ensure greater accuracy by minimizing errors. The fundamental role of the system will be to take the answer sheets and model answers as inputs and return an unbiased and completely evaluated answer sheet as the output.

The goal of this proposed system is to develop a system which conducts the evaluation of subjective answer assessments and automatically generates the marks obtained by the student based on the model answer key provided. A system that optimizes and examines the different ways in which answers can be formed and structured and evaluates the result accordingly. The system will be able to assist the process of paper evaluation and hence decrease the efforts and time consumed in correction of papers. The key objectives are as follows:

- To find out a method for the conversion of scanned answer sheets to a readable text format.
- To come up with a system for the comparison of student answers to model answers.
- To devise a way to calculate the final grade of a student's answers.

Further in this paper, the related works, problem definition, proposed system, system evaluation and analysis and the conclusion are discussed.

II. Related Works

This segment will represent a literature survey of similar existing systems. The following different technologies were studied before the development of the proposed model.

978-1-6654-0430-3/21/$31.00 ©2021 IEEE

Authorized licensed use limited to: K J Somaiya College of Engineering - MUMBAI. Downloaded on September 25,2021 at 04:48:29 UTC from IEEE Xplore. Restrictions apply.

CAA provides prompt, helpful responses for each question based on the output provided [1]. NLP approaches used to evaluate answers are also explored. The proposed system was divided into three main phases which includes prediction of best answers for short questions using NLP, essay type long answer evaluation using tools and technologies such as sentence splitting, POS tagging, wordnet and tokenizing. In the final phase, an approach for qualitative evaluation of structured answers using keyword analysis and sentence analysis is explained [1].

Nisarg Dave, Harsh Mistry and Jai Prakash Verma studied text mining and text comparison to develop a computerized checking mechanism to evaluate a subjective answer sheet [2]. They proposed to use OCR for pure text format and a JAVA open-source API (Jortho) for spell checking. They have given an elaborate explanation for text comparison.

Dharma Reddy Tetali , Dr. Kiran Kumar G and Lakshmi Ramana elaborated on a technique to evaluate descriptive answers by matching keywords in a answer key to the keywords and phrases in the answer base [3].

Prince Sinha along with his co-authors developed an application that would use the fundamentals of machine learning and apply keyword matching based on the datasets in order to evaluate the answers [4]. They state that the existing systems are only capable of evaluating MCQ type questions and there is a need for a system to evaluate subjective answers. In their paper, they propose to use OCR to scan the answer paper to split the answer keyword and based on these keywords, their application will allot marks in the range from 1 to 5.

Ms. Shweta M. Patil and Prof. Ms. Sonal Patil reviewed the techniques in Computer Assisted Assessment of free-text answers and then proposed a system to evaluate descriptive type answers using NLP [5]. They state that many evaluation are done considering specific concepts which if present in the answer, the marks were awarded. To overcome this, their new proposed system's technique is categorized into three main types: Statistical, Information Extraction and Full Natural Language Processing.

V. Lakshmi and Dr. V. Ramesh studied that computer based evaluation of answers plays a vital role in the world and that it is a faster method as compared to manual evaluation [6]. Their system proposes answer evaluation using Natural Language Processing and ANN called synsets to provide relations among short definitions and usage examples. They also use POS tagger and the wordnet tool in their system.

Piyush Patil, Sachin Patil, Vaibhav Miniyar and Amol Bandal also explain that manually correcting subjective answers is a time consuming task and there is a need for an automated system for this process [7]. Their suggested solution uses machine learning and natural language processing to solve this problem. Their algorithm performs tokenization, parts of speech marking, chinking, lemmatizing words, chunking, and word-netting to decide the subjective response. Their method is split into two modules, according to which scanned images will be translated to text, and the extracted text will be interpreted using NLP and Machine Learning to correct the answers and finally assign points.

Xinming Hu and Huosong Xia developed an automated assessment system which is based on Chinese automatic segmentation techniques [10]. To analyze term correlations, they use Latent Semantic Indexing (LSI).

Kittakorn Sriwanna explores a technique for evaluation using K-nearest neighbors [11]. The predicted scores will be done using the KNN algorithm which finds the target scores based on the nearest neighbor.

III. PROBLEM DEFINITION

Subjective questions are capable of examining the adopting ability of knowledge of the student, however, correctly evaluating the answer scripts is a challenging and complex job to perform and the assessment because it suffers from a number of questions such as trickiness, estimation of semantics, etc. The current way of checking subjective papers is adverse. Human errors are likely to occur since evaluators have to examine numerous answer papers at every term's end, this can result in inconsistencies in the correction of the student's answer sheet. However, artificial intelligence can help tackle this challenge by not only omitting human errors but also providing a quicker and faster output. This proposed model makes use of concepts of Artificial Intelligence, OCR, and NLP to solve this problem.

IV. PROPOSED SYSTEM

TESA comes with the opportunity of making a tedious and tiresome task in the field of education more efficient and less time consuming. The use of artificial intelligence to get optimized solutions in the form of marks obtained by the student is the core principle of the proposed system. The input answer sheets of the student will get compared to the model answer sheet by the evaluator and will then generate the final score based on multiple parameters. The score generated will be the final score which the student has obtained based on the answer given. The various parameters will be sentence splitting, Jaccard similarity, grammar checking and sentence similarity.

Few of the major challenges faced while developing a system for evaluation of student answers are stated below:

- Ensuring that every line of the model answer is compared to each line of the student answer i.e., even if the student structures the answer differently, it should be taken into consideration.
- A student can answer a question in multiple ways. Hence, it is essential to take into account the synonyms as well as similar meaning words of the words present in the model answer.
- The speed of the system must be fast so the algorithm chosen must generate quicker results.

To achieve the objectives stated previously, as well as to encounter the major challenges faced while developing a system for evaluation of student answer, this proposed system is divided into the following phases: Phase 1- OCR will be used to convert handwritten student answers into digital letters. Phase 2- Splitting the model answer and student answer into sentences and then applying Jaccard similarity, grammar checking and algorithm for sentence similarity using BERT on both the texts. Phase 3- Assigning marks based on

Authorized licensed use limited to: K J Somaiya College of Engineering - MUMBAI. Downloaded on September 25,2021 at 04:48:29 UTC from IEEE Xplore. Restrictions apply.

weighted average and displaying the score obtained by the student. A workflow diagram of phase 2 and phase 3 is depicted by Fig. 1.

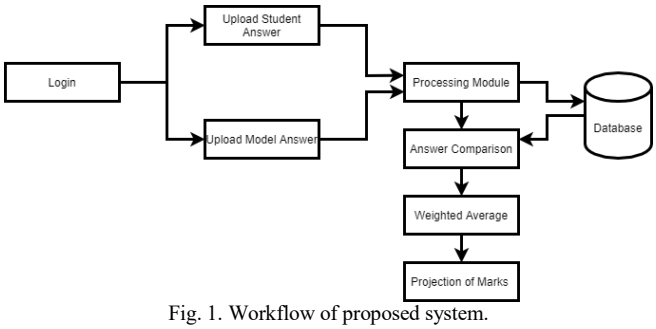

Fig. 1. Workflow of proposed system.

Existing systems evaluate answer on a text-base only i.e., they match identical keywords from two sets provided and award marks on the basis of the occurrence of those keywords [9]. This fails to take into account the different ways in which the same concept can be explained.

Several methods and algorithms were studied and implemented to find the best way for the evaluation of the subjective answers. For the implementation of the system, a method for estimation of the degree of text similarity was needed. Jaccard Similarity was hence tried. Jaccard similarity can be described as the total intersection upon the summation of union of both sets. For comparison, the following two sentences are given-
Example 1: Dogs are the most loyal to us and have always shown loyalty.
Example 2: Dogs and humans have always shown a bond of loyalty.

In the above examples, the words used are similar. Hence, a cosine score of 0.5 is allotted to it. Based on the percentage, similarity given by cosine similarity and Jaccard similarity as shown in Table 1, Jaccard similarity was chosen to be the better measure of similarity.

Table. 1. Comparison between measures of similarity.

Technique Used	Percentage Similarity
Cosine Similarity	50.33%
Jaccard Similarity	33.33%

The major setback of cosine similarity for the purpose of this system was that it takes into consideration even the repetition of the same words. As shown in Table 1, the measure of cosine similarity is higher primarily due to considering the repetitive similar words multiple times. This can generate a greater similarity level completely based on the number of times the word is repeated. Hence, Jaccard similarity is the better measure of similarity for this system.

POS tagging is the assignment of grammatical classes such as verb, noun etc. to every word in a natural language. This was done using HMM as shown in Fig. 2. A Hidden Markov Model (HMM) is a statistical construct that can be used to solve classification problems that have an inherent state sequence representation. The model can be visualized as an interlocking set of states. These states are connected by a set of transition probabilities, which indicate the probability of traveling between two given states [15]. The drawback of using this technique was that it required an extremely large dataset for model training.

```
Predicted labels:
-----------------
('DET', 'NOUN', 'VERB', 'PRON', 'VERB', 'VERB', 'DET', 'NOUN', 'CONJ', 'VERB', 'ADP', 'DET', 'NOUN', '.', 'ADV', 'NU
M', 'ADP', 'VERB', 'NOUN', '.', 'ADV', 'ADP', 'VERB', 'ADV', 'PRT', 'VERB', 'DET', 'NOUN', 'ADP', 'PRON', 'VERB', 'VE
RB', 'ADV', '.']

Actual labels:
--------------
('ADV', 'NOUN', 'VERB', 'PRON', 'VERB', 'VERB', 'DET', 'NOUN', 'CONJ', 'VERB', 'ADP', 'DET', 'NOUN', '.', 'ADV', 'NOU
N', 'PRON', 'VERB', 'NOUN', '.', 'ADV', 'ADP', 'VERB', 'ADV', 'PRT', 'VERB', 'DET', 'NOUN', 'ADP', 'PRON', 'VERB', 'V
ERB', 'ADV', '.')
```

Fig. 2. Implementation of HMM for POS tagging.

Training accuracy basic HMM model: 97.49%
Testing accuracy basic HMM model: 96.09%

Knowledge Based methods are also used to calculate the semantic similarity between given sentences. These methods can be further classified into the following:

- *WordNet:* It is a library which comprises nouns, adjectives, adverbs etc. present in the english language, which are stored in bunches together in a set of synonyms called as synsets. Because of the unique structure of WordNet, it is used extensively in NLP and NLP related applications. WordNet uses it's lexical databse to calculate the similarity between the given sentences.
- *Translating Embeddings for Modeling Multi-relational Data (TransE):* It's a tool that's simple to learn, has less requirements, and can handle massive datasets. It interprets interactions as translations on the entities' low-dimensional embeddings in order to model them [16].

Keyword mapping is the process of extracting keywords from a given sentence and it can be done in two ways:

- Syntactic Approach: The keywords are considered by its position in the sentence or by the frequency of the word
- Semantic Approach uses the semantic relationship between words.

However, it cannot accurately deal with scenarios where we have synonyms in student answers of the words that are there in the model answer. It does not take into account the order in which words appear and the grammatical meaning behind sentences.

After all the analysis and research, the most efficient methods were found to be:

a) *OCR:* It stands for Optical character recognition and it is used for the conversion of handwritten text to editable and searchable data. An picture in either format is fed into the recognition system (jpeg, png, etc). This is done by scanning an image from one of the optical scanners or accessing it from internal storage. Image pre-processing is one of the most essential and primary steps in image processing. For Image pre-processing, a machine learning model can be built in a variety of environments. Python language is one such dependable environment, and r studio is another useful tool for efficiently predicting, plotting, and depicting data.

Authorized licensed use limited to: K J Somaiya College of Engineering - MUMBAI. Downloaded on September 25,2021 at 04:48:29 UTC from IEEE Xplore. Restrictions apply.

Obtaining the dataset, or the image to be worked on, is the first step of image pre-processing. The necessary libraries can then be imported, and these libraries come with a number of pre-built functions that assist in manipulating and visualising data effectively. Python has several powerful libraries for high-end arithmetic operations, such as numpy.

The AFORGE library is imported, which supports computer vision, image and video processing, ANN, Optical Recognition of Digital Characters using features of Machine Learning. The training set is then generated from the dataset. The training data is expected to train the computer, and the machine learns from this data. The test data is the information that is fed into the computer in order to obtain results. Function scaling is performed after the data set has been classified into test and train sets. It works with the image's most important features, such as the minimum bound box, segmentation etc.

Hence, it will be efficient for converting scanned student answer sheets to text format [13, 14].

b) Sentence splitting: As the name suggests, the function of this is to split multiple sentences in a paragraph into individual sentences accurately. This can be done using various NLP frameworks.

c) Jaccard Similarity: The Jaccard similarity is used for measuring the similarity between data sets, dissimilarity, and distance [12]. The Jaccard similarity coefficient between two data sets is calculated by dividing the total number of shared features by the total number of properties, as shown below.

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

Where,
J= Jaccard Distance, A= Set 1, B= Set 2

d) Bidirectional Encoder Representations from Transformers (BERT): It is a recently developed model which provides high accuracy and hence is reliable. The full form of BERT is Bidirectional Encoder Representations from Transformers and in our system it is used with Jaccard similarity to find out the sentence similarity. The BERT system is divided into two main parts: pre-training and fine-tuning. The model is conditioned on datasets through various pre-training activities. The BERT model is fine-tuned using labelled data from downstream tasks after it is initialised with the pre-trained parameters. Further, if they are all initialised with the same pre-trained parameters, each downstream activity has its own fine-tuned model. [8]. BERT embeddings are contextual i.e., BERT can differentiate between words with the same spelling used in different contexts. A cosine similarity is shown between the indicated word pairs. As stated by google, BERT achieves 93.2% F1 score (a measure of accuracy), surpassing the previous state-of-the-art score of 91.6% and human-level score of 91.2%.

The weighted average of Jaccard similarity, grammar checking and BERT will be then calculated to allot marks to a given student answer.

For the frontend implementation of the proposed system, Django framework has been used. Since the language used to build the system is python, Django is considered to build the evaluation portal of the system.

V. System Performance and Evaluation

This segment elaborates on the utilization and integration of the processes mentioned above into this system.

- Sentence splitting is used to split the model and student answers into sentences accurately. The system makes sure that the sentences are split taking abbreviations into account. For example, the sentence should not be split after the 'dot (.)' in "Dr." or "Prof."

- Jaccard similarity is then implemented to calculate the measure of similarity between the model answer and the student answer.

- BERT: It is a technique developed by Google. Using BERT, the semantic similarity between the model answer and student answer is obtained. Fig. 3 represents BERT implementation used for comparing one model answer to multiple student answers.

```
======================
Model Answer: Today is a great day.


Today is a great day. (Score: 1.0000)
Today is a great day. (Score: 1.0000)
It is a bright, lovely and sunny. (Score: 0.8626)
It is bright, lovely and sunny. (Score: 0.8602)
It is shining, beautiful and sunlit. (Score: 0.8293)
It is a greay day today. (Score: 0.8117)
The fox and the cow got into a big fight. (Score: -0.0557)
======================
```

Fig. 3. BERT implementation.

- In the final process, the weighted average of the Jaccard similarity and BERT is quantified to assign and allocate the total marks obtained.

The detailed flowchart for the proposed system discussed in the previous sections is shown in Fig. 4.:

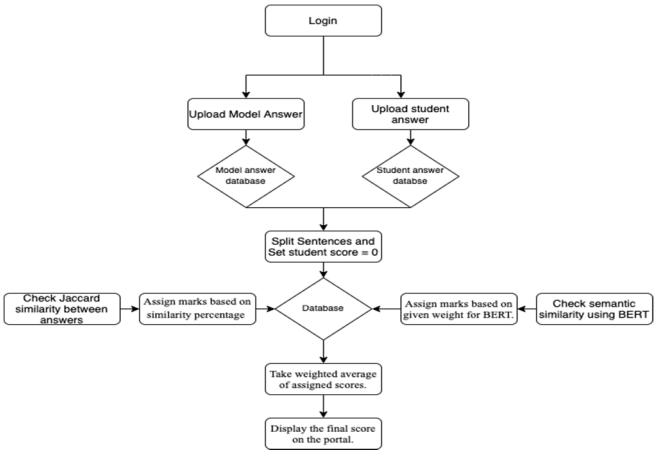

Fig. 4. Detailed Flowchart of the Proposed System.

Authorized licensed use limited to: K J Somaiya College of Engineering - MUMBAI. Downloaded on September 25,2021 at 04:48:29 UTC from IEEE Xplore. Restrictions apply.

The final implementation results are displayed in Fig. 5, Fig. 6, Fig. 7, Fig. 8 and Fig. 9.

Fig. 5. Depicts a login page of the model which can be accessed by the exam-cell members as well as by the teachers. New users can register using the "Register" button and for the returning users, their username and password is taken as the input and the concerned member can login using their respective credentials.

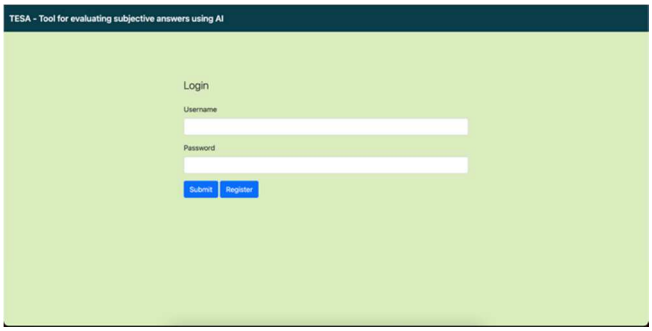

Fig. 5. Login page for evaluators.

Fig. 6 and Fig. 7 represent the pages where the exam-cell members can upload the model answer and student answer respectively. These pages are exclusively accessible to the exam-cell members. For every answer, the subject ID, subject name and out of marks are to be filled for error free evaluation. The answer file can also be uploaded by choosing a file from a device.

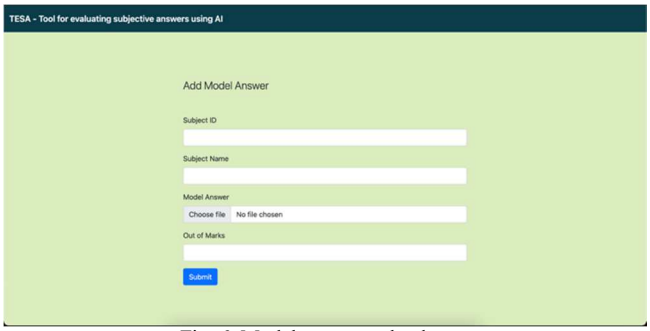

Fig. 6. Model answer upload page.

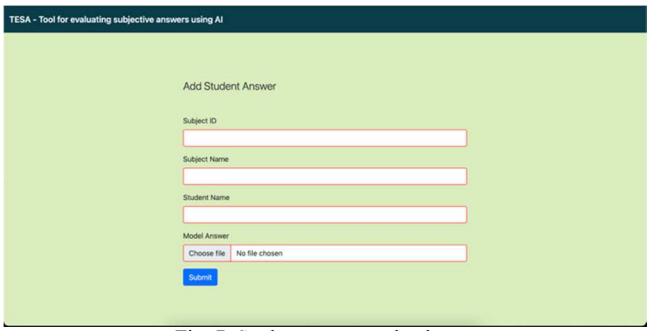

Fig. 7. Student answer upload page.

Fig. 8. Displays the evaluation page where a teacher can evaluate the answers uploaded previously. This page can be accessed by the teachers or evaluators only. The "Evaluate" button will run the backend code to evaluate the uploaded answers and upon evaluation, the scores obtained by the students upon evaluation can also be checked.

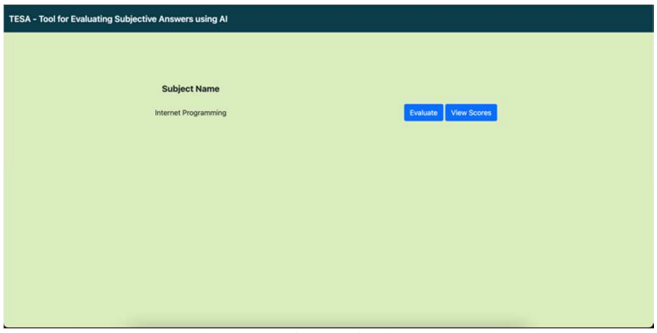

Fig. 8. Teacher evaluation page.

Fig. 9 presents the final evaluated scores which can be seen after the evaluation is complete.

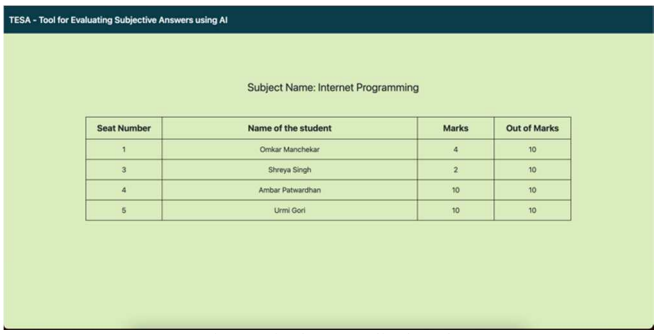

Fig. 9. Final scores displayed.

VI. Conclusion

All the studies which have been reviewed show that there are various different techniques for the evaluation of subjective answer sheets. The advantage of the system lies in the fact that it uses a weighted average of the closest to accurate techniques to provide the most optimized result. TESA is a systematic and reliable system which eases the role of evaluators and provides faster and more efficient outputs. This system offers a reliable, robust, and obvious short response time result. In the future, a system can be developed to evaluate diagrams as well as tables, an inbuilt system can also be made to type and make diagrams to shift examinations from handwritten paper based to completely online.

References

[1] P.A.A. Dumal, W.K.D Shanika, S.A.D Pathinayake and T.C. Sandanayake.: Adaptive and Automated Online Assessment Evaluation System. In: 2017 11th International Conference (SKIMA)

[2] Nisarg Dave, Harsh Mistry and Jai Prakash Vera, Assist. Prof.: Text Data Analysis: Computer Aided Automated Assessment System. In: IEEE-CICT 2017, 978-1-5090-6218-8/17/$31.00 ©2017 IEEE

[3] Dharma Reddy Tetali, Dr. Kiran Kumar G and Lakshmi Ramana.: A Python Tool for Evaluation of Subjective Answers (APTESA). IJMET Volume 8, Issue 7, July 2017, pp. 247–255, Article ID: IJMET_08_07_029

[4] Prince Sinha, Sharad Bharadia, Dr. Sheetal Rathi and Ayush Kaul.: Answer Evaluation Using Machine Learning. In: https://www.researchgate.net/publication/333856264, March 2018

[5] Ms. Shweta M. Patil and Prof. Ms. Sonal Patil.: Evaluating Student Descriptive Answers Using Natural Language Processing. (IJERT) ISSN: 2278-0181 Vol. 3 Issue 3, March - 2014

[6] V. Lakshmi and Dr. V. Ramesh.: Evaluating Students' Descriptive Answers Using Natural Language Processing and

Authorized licensed use limited to: K J Somaiya College of Engineering - MUMBAI. Downloaded on September 25,2021 at 04:48:29 UTC from IEEE Xplore. Restrictions apply.

Artificial Neural Networks. 2017 IJCRT | Volume 5, Issue 4 December 2017 | ISSN: 2320-2882

[7] Piyush Patil, Sachin Patil, Vaibhav Miniyar and Amol Bandal.: Subjective Answer Evaluation Using Machine Learning. International Journal of Pure and Applied Mathematics Volume 118 No. 24 2018, ISSN: 1314-3395 (on-line version)

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee and Kristina Toutanova.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: Proceedings of NAACL-HLT 2019, pages 4171–4186 Minneapolis, Minnesota, June 2 - June 7, 2019. 2019 Association for Computational Linguistics

[9] Shun LONG, Qunhao FENG and Wenwei CHEN.: A Novel Approach to Automatic Rating of Subjective Answers based on Semantic Matching of Keywords. In: 2016 12th International Conference on Computational Intelligence and Security

[10] Xinming Hu and Huosong Xia.: Automated Assessment System for Subjective Questions Based on LSI. 978-0-7695-4020-7/10 $26.00 © 2010 IEEE

[11] Kittakorn Sriwanna.: Text Classification for Subjective Scoring Using K-Nearest Neighbors. In: The 3rd International Conference on Digital Arts, Media and Technology (ICDAMT2018) 978-1-5386-0572-1/18/$31.00 ©2018 IEEE

[12] Suphakit Niwattanakul, Jatsada Singthongchai, Ekkachai Naenudorn and Supachanun Wanapu.: Using of Jaccard Coefficient for Keywords Similarity. In: Proceedings of the International MultiConference of Engineers and Computer Scientists 2013 Vol I, IMECS 2013, March 13 - 15, 2013, Hong Kong

[13] Dr Sunanda Dixit, Bharath, Amith Y, Goutham M L, Ayappa K and Harshitha D.: Optical Recognition of Digital Characters Using Machine Learning. International Journal of Research Studies in Computer Science and Engineering (IJRSCSE), Volume 5, Issue 1, 2018, PP 9-16

[14] Polaiah Bojja , Naga Sai Satya Teja Velpuri, Gautham Kumar Pandala, S D Lalitha Rao Sharma, Polavarapu and Pamula Raja Kumari.: Handwritten Text Recognition using Machine Learning Techniques in Application of NLP. IJITEE, ISSN: 2278-3075, Volume-9 Issue-2, December 2019

[15] Sanjeev Kumar Sharma and Dr. Gurpreet Singh Lehal.: Using Hidden Markov Model to Improve the Accuracy of Punjabi POS Tagger. In: 2011 IEEE International Conference on Computer Science and Automation Engineering, CSAE 2011. 2. 10.1109/CSAE.2011.5952600.

[16] Antoine Bordes, Nicolas. Usunier, Alberto Garcia-Duran, Jason Weston and Oksan Yakhnenko.: Translating Embeddings for Modeling Multi-relational Data. In:https://papers.nips.cc/paper/2013/file/1cecc7a77928ca8133fa24680a88d2f9-Paper.pd

Authorized licensed use limited to: K J Somaiya College of Engineering - MUMBAI. Downloaded on September 25,2021 at 04:48:29 UTC from IEEE Xplore. Restrictions apply.