Industrial Engineering Journal
ISSN: 0970-2555
Volume : 53, Issue 6, June : 2024

AI BASED PAPER AUTOMATIC EXAMINATION PAPER EVALUATION SYSTEM

Dr. A.LAXMIKANTH, Associate Professor,

M. RAMYA, R. JAHNAVI, N. SAI KIRAN, T. RAKESH

SRI INDU COLLEGE OF ENGINEERING AND TECHNOLOGY

Sheriguda (V), Ibrahimpatnam (M), RangareddyDist – 501 510

ABSTRACT

Subjective paper evaluation is a tricky and tiresome task to do by manual labor. Insufficientunderstanding and acceptance of data are crucial challenges while analyzing subjective papersusing Artificial Intelligence (AI). Several attempts have been made to score students' answersusing computer science. However, most of the work uses traditional counts or specific wordsto achieve this task. Furthermore, there is a lack of curated data sets as well. This paper proposesa novel approach that utilizes various machine learning, natural language processingtechniques, and tools such as Wordnet, Word2vec, word mover's distance (WMD), cosinesimilarity, multinomial naive bayes (MNB), and term frequency-inverse document frequency(TF-IDF) to evaluate descriptive answers automatically. Solution statements and keywords areused to evaluate answers, and a machine learning model is trained to predict the grades ofanswers. Results show that WMD performs better than cosine similarity overall. With enoughtraining, the machine learning model could be used as a standalone as well. Experimentationproduces an accuracy of 88% without the MNB model. The error rate is further reduced by1.3% using MNB

INDEX : manual labor, si, wordnet, word2vec, tf-idf

1.INTRODUCTION

1.1 Introduction:

Subjective questions and answers can assess the performance and ability of a student in anopen-ended manner. The answers, naturally, are not bound to any constraint, and students arefree to write them according to their mindset and understanding of the concept. With that said,several other vital differences separate subjective answers from their objective counterpart. Forone, they are much longer than the objective questions. Secondly, they take more time to write.Moreover, they carry much more context and take a lot of concentration and objectivity fromthe teacher evaluating them.Evaluation of such questions using computers is a tricky task, mainly because natural languageis ambiguous. Several preprocessing steps must be performed, such as cleaning the data andtokenization before working on it. Then the textual data can be compared using varioustechniques such as document similarity, latent semantic structures, concept graphs, ontologies.The final score can be evaluated based on Similarity, keywords presence, structure, language.Several attempts have been made in the past to solve this problem, but there is still room forimprovements, some of which is discussed in this paper. Subjective exams are considered morecomplex and scarier by both students and teachers due to their one fundamental feature, context.A subjective answer demands the checker check every word of the answer for scoring actively,and the checker's mental health, fatigue, and objectivity play a massive role in the overall result.Therefore, it is much more time and resource-efficient to let a system handle this tedious andsomewhat critical task of evaluating subjective answers. Evaluating objective answers withmachines is very easy and feasible. A program can be fed with questions and one-word answersthat can quickly map students' responses. Nevertheless, subjective answers are much morechallenging to tackle. They are varied in length and contain a vast amount of vocabulary.Furthermore, people tend to use synonyms and convenient abbreviations, which makes theprocess that much tricky

2. LITERATURE SURVEY

TITLE: ''Measurement of text similarity: A survey,'' Information

ABSTRACT: Text similarity measurement is the basis of natural language processing tasks,which play an important role in information retrieval, automatic question answering, machinetranslation, dialogue systems, and document matching. This paper systematically combs theresearch status of similarity measurement, analyzes the advantages and disadvantages of currentmethods, develops a more comprehensive classification description system of text similaritymeasurement algorithms, and summarizes the future development direction. With the aim ofproviding reference for related research and application, the text similarity measurementmethod is described by two aspects: text distance and text representation. The text distance canbe divided into length distance, distribution distance, and semantic distance; text representationis divided into string-based, corpus-based, single-semantic text, multi-semantic text, and graph-structure-based representation. Finally, the development of text similarity is also summarizedin the discussion section.

TITLE: ''A survey on the techniques, applications, and performance of short text semanticsimilarity,''

ABSTRACT: Short text similarity plays an important role in natural language processing(NLP). It has been applied in many fields. Due to the lack of sufficient context in the short text,it is difficult to measure the similarity. The use of semantics similarity to calculate textualsimilarity has attracted the attention of academia and industry and achieved better results. Inthis survey, we have conducted a comprehensive and systematic analysis of semantic similarity.We first propose three categories of semantic similarity: corpus-based, knowledge-based, anddeep learning (DL)-based. We analyze the pros and cons of representative and novel algorithmsin each category. Our analysis also includes the applications of these similarity measurementmethods in other areas of NLP. We then evaluate state-of-the-art DL methods on four commondatasets, which proved that DL-based can better solve the challenges of the short text similarity,such as sparsity and complexity. Especially, bidirectional encoder representations fromtransformer model can fully employ scarce information of short texts and semantic informationand obtain higher accuracy and F1 value. We finally put forward some future directions.

TITLE: ‘‘Subjective answer evaluation using machine learning,’

ABSTRACT: This project proposes a novel approach that utilizes various machine learning,natural language processing techniques, to evaluate descriptive answers automatically.Solution statements and keywords are used to evaluate answers, and a machine learningmodel is trained to predict the grades of answers.With enough training, the machine learning model could be used as a standalone as well.Experimentation produces an accuracy of 97% with the Proposed model.Interestingly, artificial intelligence is utilized extensively as an efficient tool for predictingsuch a problem.The proposed work utilizes the deep learning technique along with some preprocessingsteps to improve the prediction of Answer Evaluation.

TITLE: ‘‘Automated assessment system for subjective questions based on LSI,’’

ABSTRACT: Subjective question is capable of examining the adopting ability of knowledgeof the student, but the assessment for it suffers from a number of questions such as trickiness,synonymy and polysemy. This reduces the advantage of subjective question for online exercise.In this paper we explore an approach to automated assessment system for subjective questionbased on latent semantic indexing. Chinese automatic segmentation techniques and subjectontology are used for transferring the reference answers to a term-document matrix, which isthen projected to a k-dimensional LSI space by the statistical technique Singular ValueDecomposition to solve the problem of synonymy and polysemy. A reference unit vector isintroduced to alleviate the problem of trickiness. The system then concludes the quality of thesolution according to the similarity between the projected vectors. The experimental resultsprove the feasibility of our theoretical architecture and flow for automated assessment ofsubjective question.

TITLE: ‘‘From word embeddings to document distances,’’

ABSTRACT: We present the Word Mover’s Distance (WMD), a novel distance functionbetween text documents. Our work is based on recent results in word embeddings that learnsemantically meaningful representations for words from local co-occurrences in sentences. TheWMD distance measures the dissimilarity between two text documents as the minimum amountof distance that the embedded words of one document need to "travel" to reach the embeddedwords of another document. We show that this distance metric can be cast as an instance of theEarth Mover’s Distance, a well studied transportation problem for which several highlyefficient solvers have been developed. Our metric has no hyperparameters and is straight-forward to implement. Further, we demonstrate on eight real world document classification datasets, in comparison with seven state-of-the-art baselines, that the WMD metric leads tounprecedented low k-nearest neighbor document classification error rates.

TITLE: ‘‘Similarity analysis of law documents based on Word2vec,’’

ABSTRACT: With the increasing demand for computer-assisted wisdom in justice, deeplearning has gradually become an effective means of helping intelligent justice. The similarityanalysis of law documents is the basis of intelligent justice, while law documents based onseveral types of cases are quite different in terms of format and length, which causes trouble inanalyzing similarities. For that we propose a more specific approach to dealing with lawdocuments, combining Word2vec with legal documents corpus. To measure the efficiency ofthe proposed method, we designed two sets of controls. The experimental results show that theWord2vec model can improve the accuracy by 0.20 compared with the bag of words (BOW)model, and the equipped law documents corpus can increase by 0.05-0.10 based on theWord2vec model. Thus, the combination of Word2vec and the law documents corpus is morecompatible with the simple and efficient application of similarity analysis of law documents.

3. PROBLEM STATEMENT

It appears there might be some confusion. The information you provided in your previousmessage seems to be more focused on the proposed system and its objectives rather thandescribing the existing system. The existing system typically refers to the state of affairs ormethodologies in place before the implementation of the proposed system. If you haveinformation about the existing system, you could provide details on how the assessment ofdescriptive answers is currently handled, whether it's manual evaluation by teachers or anyexisting tools or methods in use.

LIMITATIONS OF SYSTEM:

Subjectivity and Bias:Issue: Manual evaluation can be subjective, leading to variations in grading amongdifferentevaluators.Impact: Inconsistencies in grading may result in unfair assessments and disparities instudents' grades.Time-Consuming:Issue: Manual grading of descriptive answers is a time-consuming process, especiallyin scenarios with a large number of students or complex questions.Impact: Teachers may face challenges in providing timely feedback to students, andthe overall assessment process may be delayed.Scalability Challenges:Issue: As the number of students and assessments increases, scalability becomes asignificant challenge for manual evaluation.Impact: Educational institutions may struggle to efficiently manage and scale theassessment process, particularly during peak times.

4. PROPOSED SYSTEM & IT'S ADVANTAGES:

The proposed system, "A Descriptive Answer Evaluation System Using CosineSimilarity Technique," offers a transformative approach to address the limitations of traditionalmanual evaluation methods for descriptive answers. The primary objective is to leveragecomputer-assisted assessment tools, particularly in the context of the evolving challenges posedby the COVID-19 pandemic. The system aims to alleviate the subjectivity and bias inherent inmanual grading by introducing an automated evaluation process based on the cosine similaritytechnique. This method allows for a more objective assessment of descriptive answers,irrespective of their length, enabling a fairer and more consistent grading system.One of the key features of the proposed system is its ability to significantly reduce thetime required for assessment. By automating the evaluation process, teachers can allocate moretime to providing detailed and timely feedback to students. The scalability of the systemaddresses the challenges associated with handling a large number of assessments efficiently.This shift towards a computer-assisted solution not only streamlines the evaluation process butalso minimizes the resource intensive nature of manual grading, potentially leading to costsavings for educational institutions.Moreover, the proposed system enhances the feedback loop for students by providingpictorial representations of the results using the cosine similarity technique. This visualrepresentation not only facilitates a quick understanding of the assessment outcome but alsoserves as a valuable learning aid. The web-based application aspect further modernizes theassessment approach, making it more adaptable to the digital learning environment. In essence,the proposed system strives to revolutionize the assessment of descriptive answers, making itmore objective, efficient, and accessible in the contemporary educational landscape.

4.1 ADVANTAGES:

Objective Assessment: The system employs the cosine similarity technique, aquantitative measure that provides an objective evaluation of descriptive answers. Thishelps eliminate subjective biases often associated with manual grading, ensuringfairness and consistency in assessments.

Time Efficiency: Automated evaluation significantly reduces the time required forgrading descriptive answers. This efficiency benefits both teachers and students byexpediting the feedback process, allowing for quicker identification of areas ofimprovement and enhancing the overall learning experience.

Scalability: The proposed system is designed to handle a large volume of assessmentsefficiently. As the number of students and evaluations increases, the automatedapproach ensures scalability, addressing the challenges posed by manual grading interms of time and resource constraints.

Enhanced Feedback: The system provides visual representations of assessment resultsusing the cosine similarity technique. These pictorial representations offer a clear andconcise overview of performance, aiding students in understanding their strengths andweaknesses. This enhanced feedback supports a more targeted approach to learning andimprovement.

5. IMPLEMENATION

1 KEYWORDS:

Keywords are question-specific things that are essential for answering that question. Thesekeywords play a significant role in penalizing or promoting the score evaluated by the similaritymeasurement module and must only contain the essential words in lower case.

2 SOLUTION:

The solution is a subjective answer that is being used to map students' responses. This solutionmust contain all the keywords and contexts discussed in the answers in separatelines/paragraphs. The teacher/evaluator typically prepares the solution to the question.

3 ANSWER

The answer is a subjective response from the student that is to be evaluated. It usually containssome or all of the keywords and spans 1 to a few sentences depending on the type of questionand the student's writing style. It almost always contains synonym words compared to thesolution and, therefore, requires much more semantic care when processing.

4 DATA COLLECTION

To train and test the proposed model, there is a need for a massive amount of corpus containingsubjective question answers, but there is no publicly available labeled subjective questionanswers corpus to the best of our knowledge. In this work, we create subjective answers labeledcorpus. For generating corpus, the important thing is to target those websites and blogs wheresubjective questions and answers exist. We crawl various websites and collect a subjectivequestion answers corpus, and the crawl data belong to various domains such as computerscience and general knowledge.

5 DATA ANNOTATION

After getting crawled data, there is a further need to annotate data because that crawled data isunlabeled. To annotate data, a group of different volunteers is selected, which belong to thedomain of our subjective question answers corpus. We hire 30 different annotators fromdifferent colleges and universities and reside in Pakistan's different cities. Most of them arestudents and teachers. The average age of annotators is in the 21-25 range, whereas someannotators are in the age range of 27-51. We task annotators to best score the subjective questionanswers according to the answers given by students.

6 PREPROCESSING MODULE

After taking inputs from the user, both the solution and the answer go through somepreprocessing steps, which involve tokenization, stemming, lemmatization, stop wordsremoval, case folding, finding, and attaching synonyms to the text. Note that stop words are notremoved when passing the data to word2vec because word2vec contains a vast vocabulary andcan utilize those stop words to make better semantic sense of the text. However, stop words areremoved before passing to a machine learning model such as Multinomial Naive Bayes becausethey hinder the machine's ability to learn the patterns.

7 SIMILARITY MEASUREMENT MODULE

This module consists of WDM and Cosine Similarity functions which take two sentences orword vectors and return their Similarity. WDM tells us the dissimilarity while Cosine Similaritymeasures Similarity. Our approach uses both of these similarity measures one at a time andcompares the results at the end. Various similarity (or dissimilarity) thresholds . 1)THRESHOLDS ANALYSIS Various thresholds used in this paper have been experimentallydeduced to produce the optimal result. WDM thresholds of WDM_LOWER andWDM_UPPER represent the dissimilarity between two sentences, where more dissimilarityrepresents high similarity. 0.7 threshold for WDM_LOWER was experimentally observed torepresent semantically very similar sentences, and 1.6 thresholds for WDM_UPPER wereobserved to represent semantically less similar sentences. Anything beyond 1.6 is assumed tobe too dissimilar to consider viable for comparison. Similarly, Cosine similarity thresholdsCOS_LOWER and COS_UPPER represent the similarity between two sentences. It should benoted that cosine similarity does not take the context of two sentences into account whenmeasuring similarity as opposed to WDM, hence the usage of both of these similarity (ordissimilarity) measuring approaches.

8 RESULT PREDICTING MODULE

Result Predicting Module is the core of this work. shows the working of this module.It operates on the following Algorithm 1: We now have the overall score calculated by ourmodule using either WDM or Cosine Similarity while considering the maximum matchedsolution/answer sentence pairs. This result can be compared to an actual score or fed into amachine learning model to be trained.

9 MACHINE LEARNING MODEL MODULE

This model consists of Machine learning models trained on the data obtained from the resultprediction module.Its working is as follows:

• Input data from Result Prediction Module.

• Preprocess the solution and answer, removing stop words, and use Countvectorizer torepresent them in either Bag of Words or TF-IDF form.

• Convert the overall score obtained from Result Prediction Module into some category. Foucategories A, B, C, and D, are used in the paper, representing 1st, 2nd, 3rd, and 4th quarter of

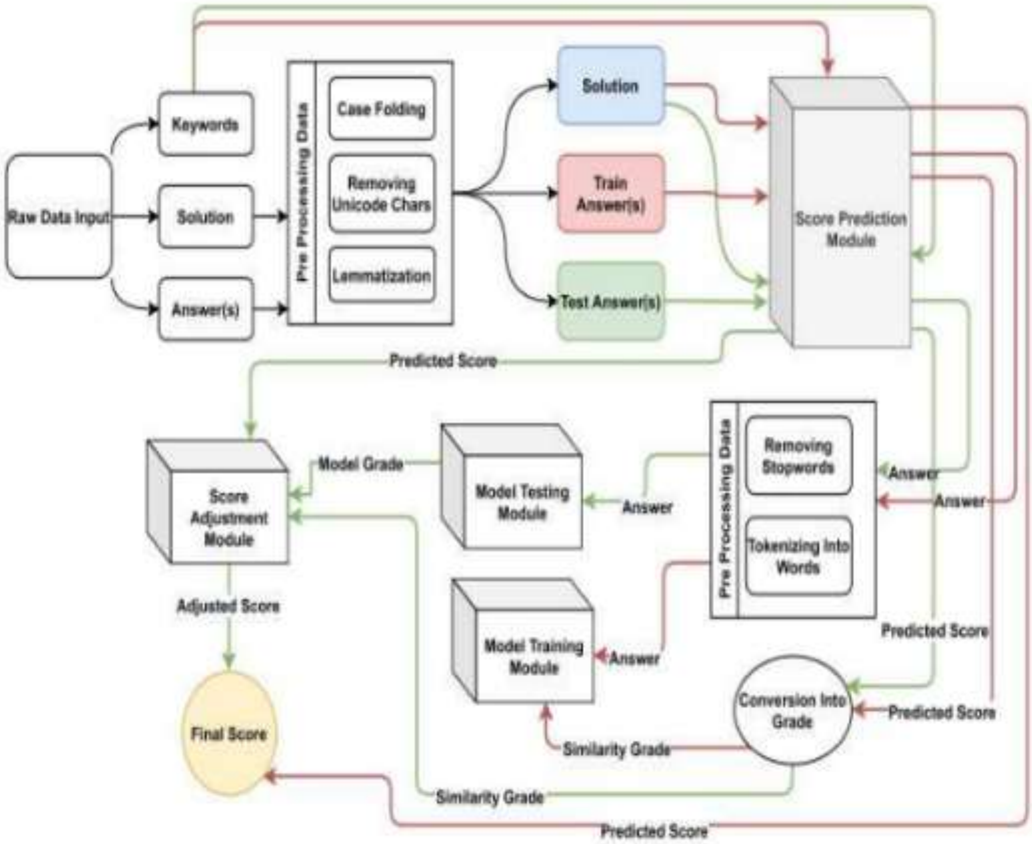
a 100. For example, A represents marks from 0 to 25, and B represents 26 to 50.

• The number of categories is kept to a minimum because of the unavailability of the actualdataset. Practically, these categories can be extended to cover smaller score ranges.

• A machine learning model such as Multinomial Naive Bayes, which performs well for multi-class classification, is chosen.

• The preprocessed answer is used as testing data with the machine learning model to predictits class/category, and that category is checked with the result obtained from Result PredictionModule. This gives us confidence in the predicted result from the model.

6. SYSTEM ARCHITECTURE



7.EXPECTED RESULTS

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 53, Issue 6, June : 2024

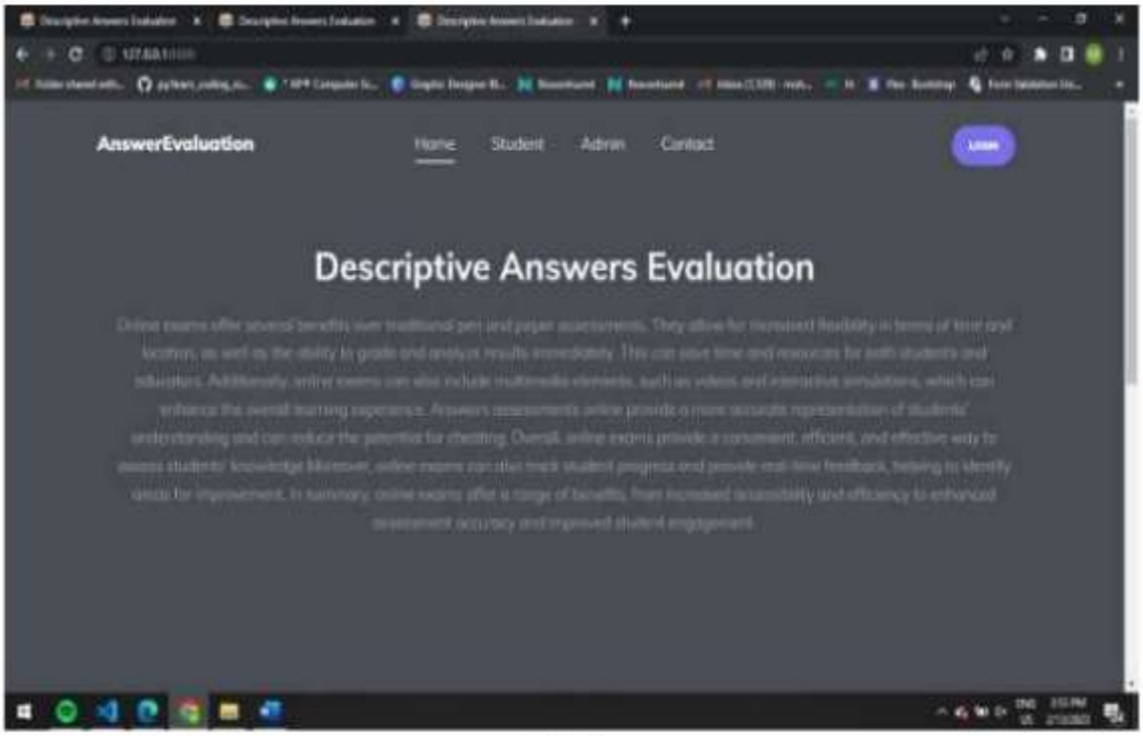


Fig.1

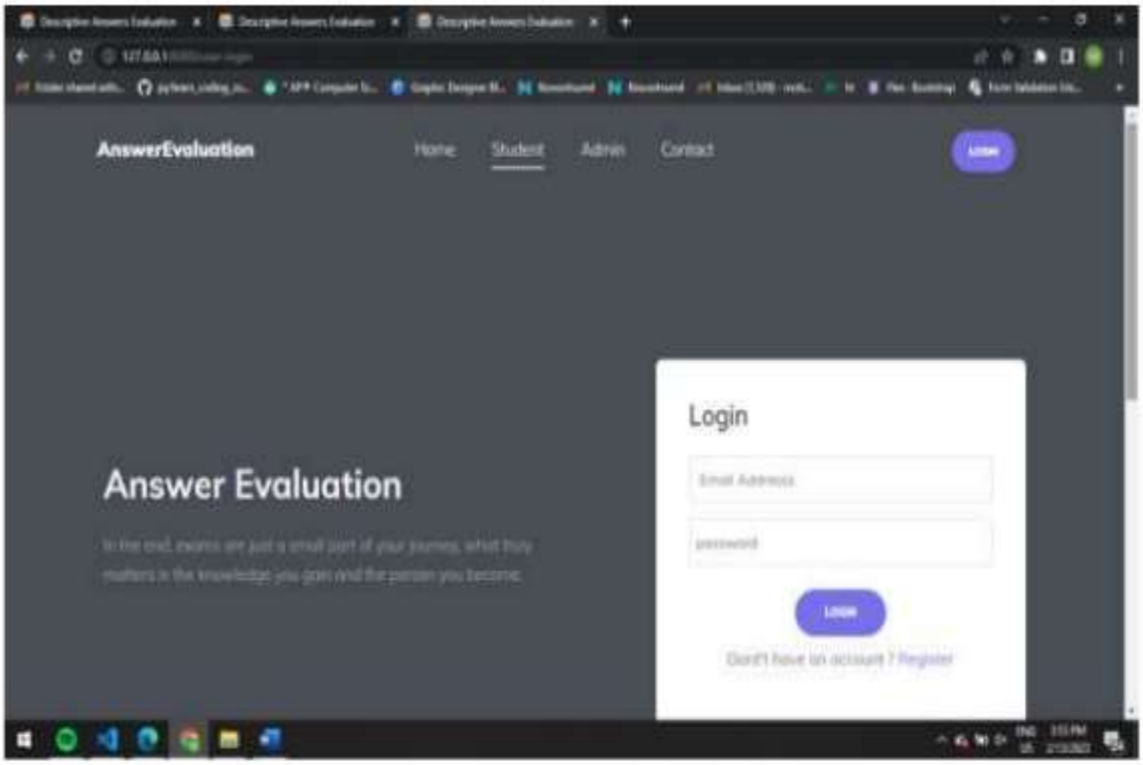


Fig.2

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 53, Issue 6, June : 2024

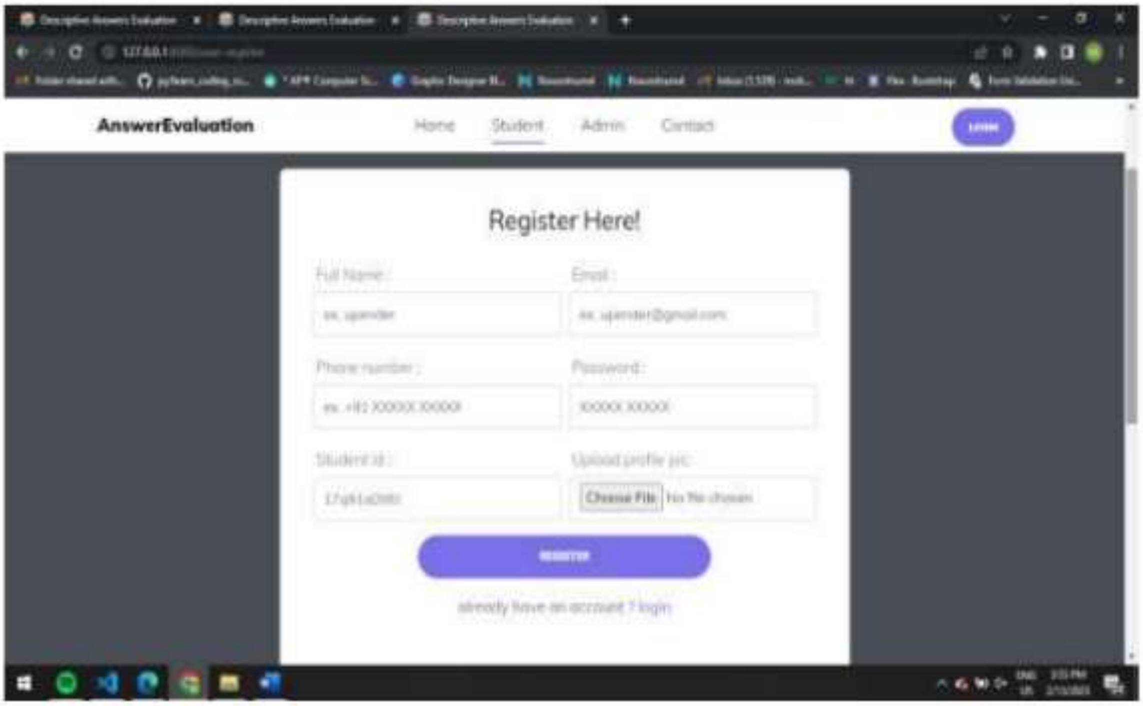


Fig.3

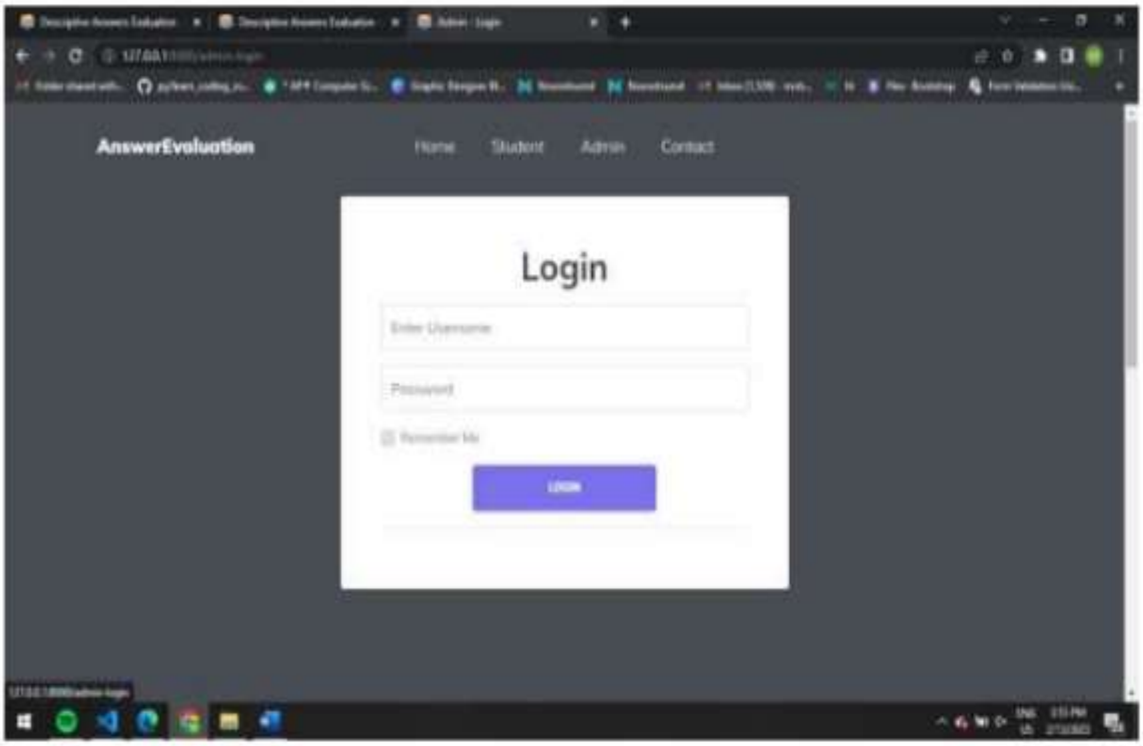


Fig.4

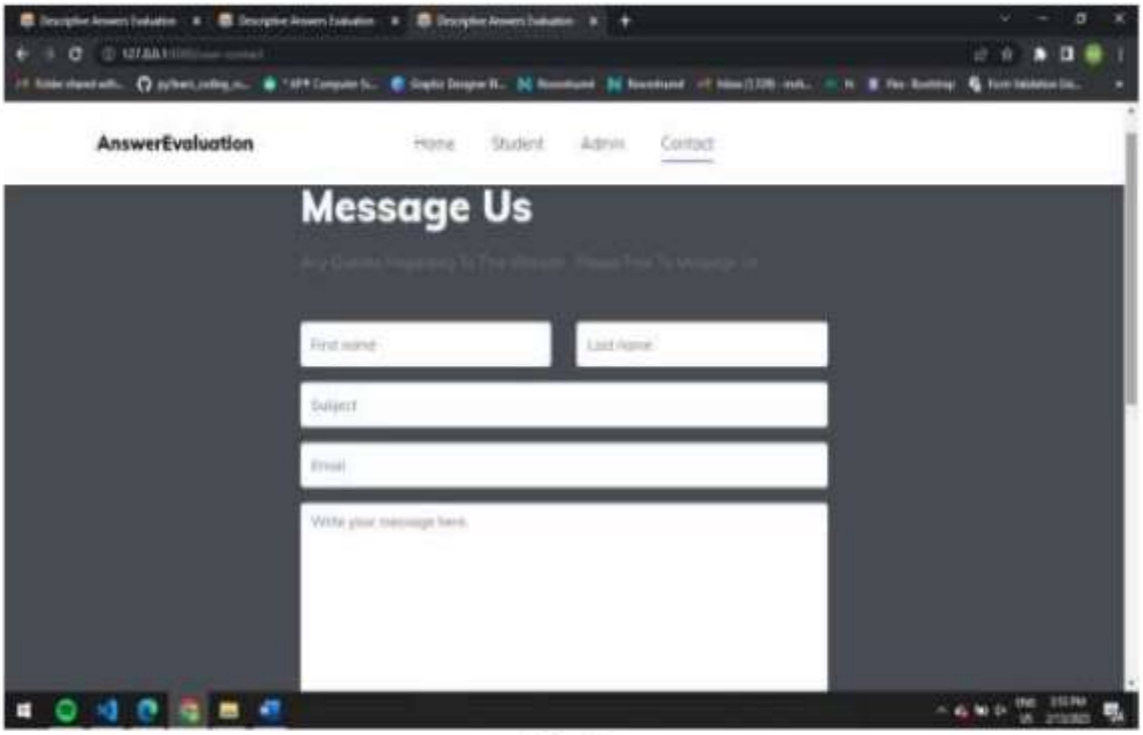


Fig.5

User:

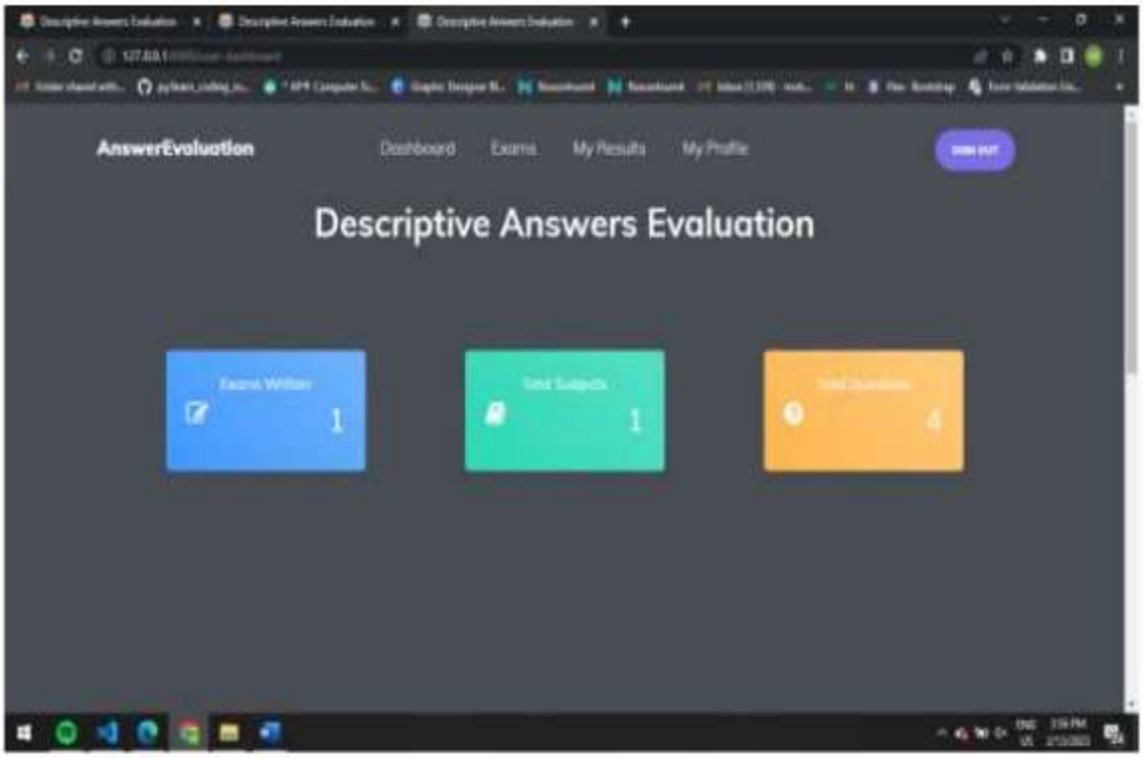


Fig.6

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 53, Issue 6, June : 2024

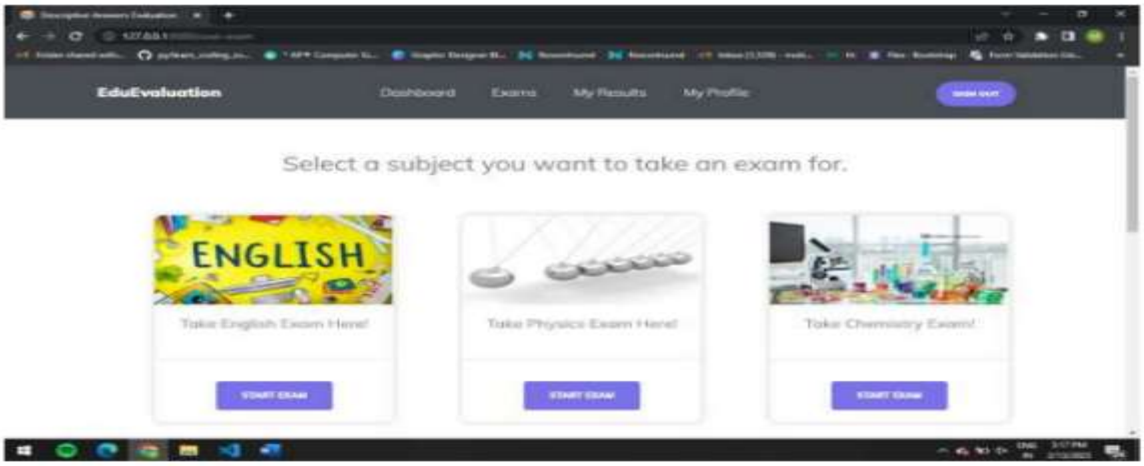


Fig.7

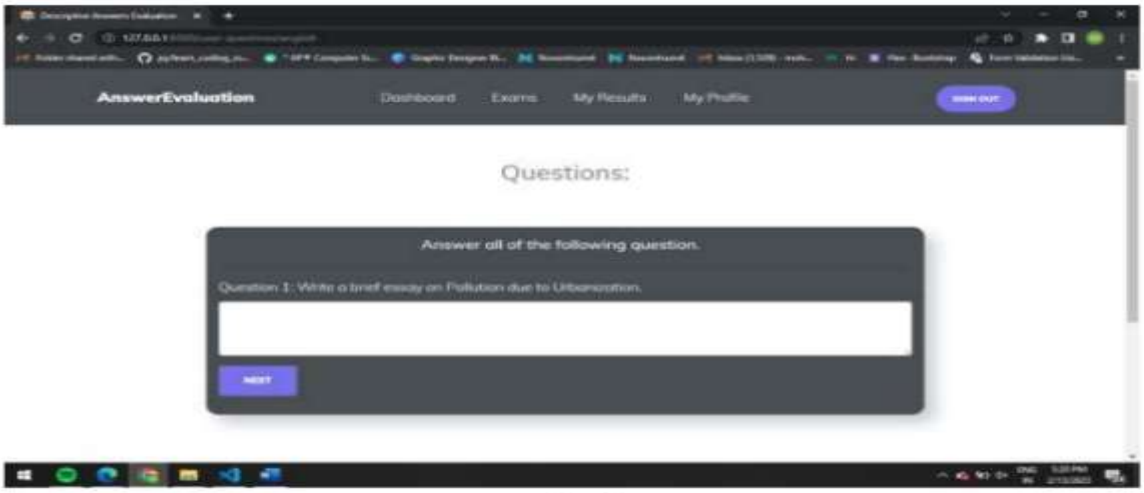


Fig.8

Fig.9

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 53, Issue 6, June : 2024

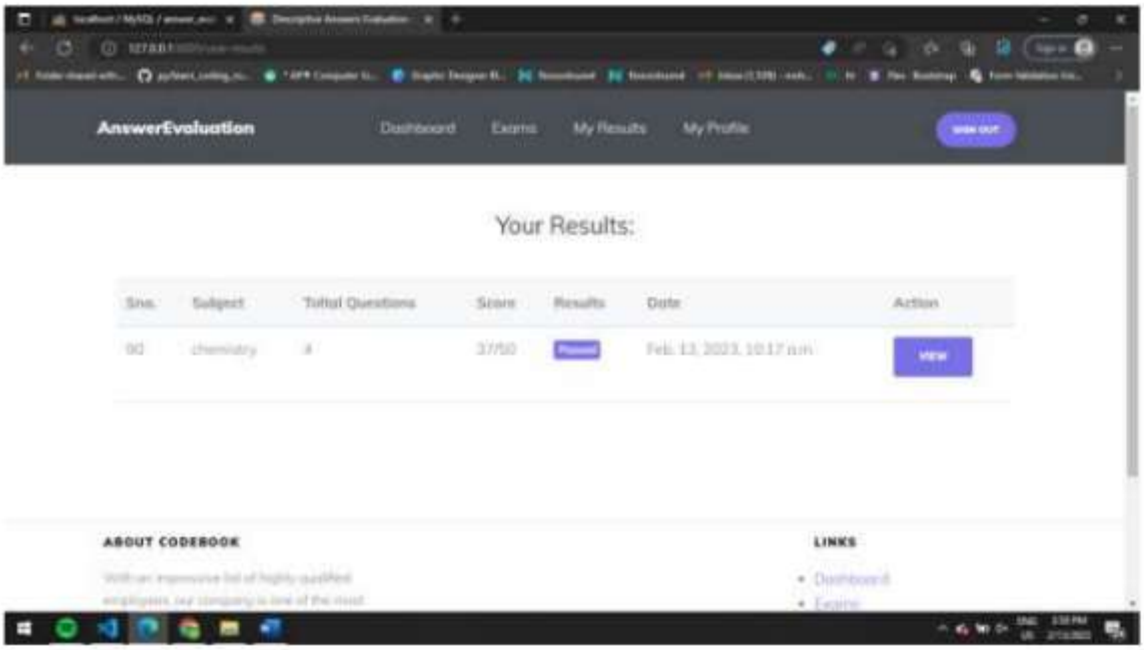


Fig.10

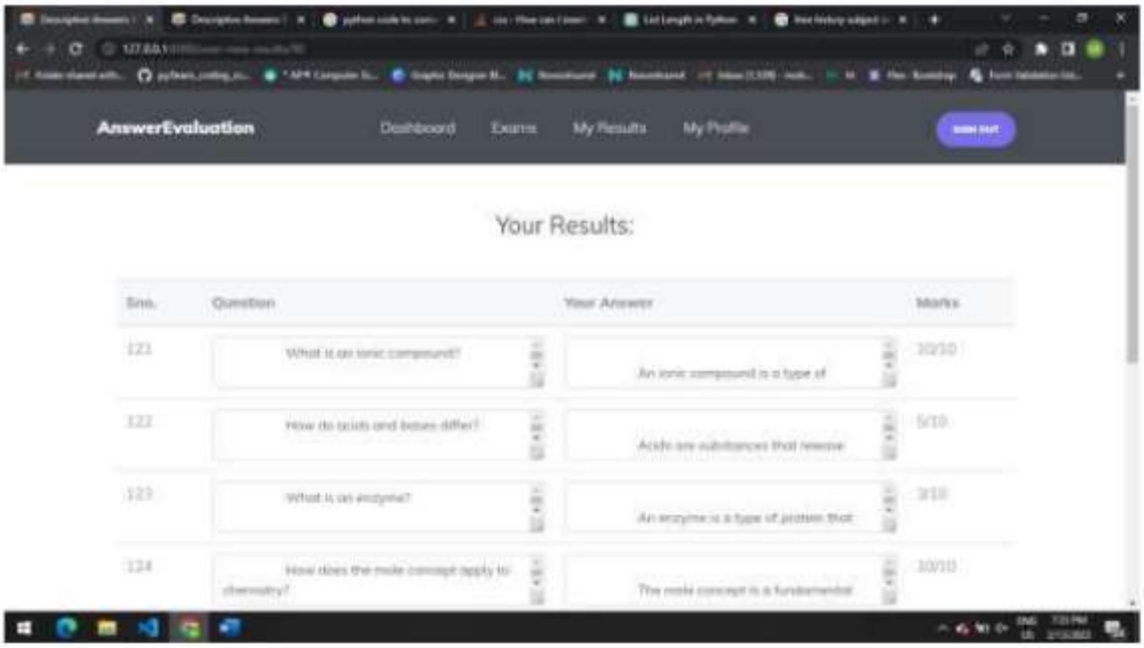


Fig.11

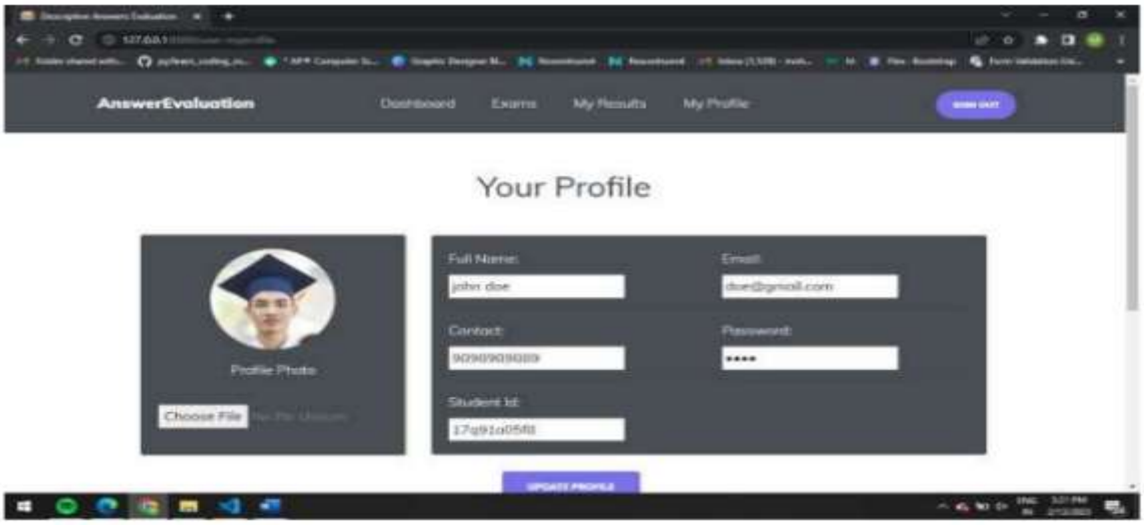


Fig.12

Admin:

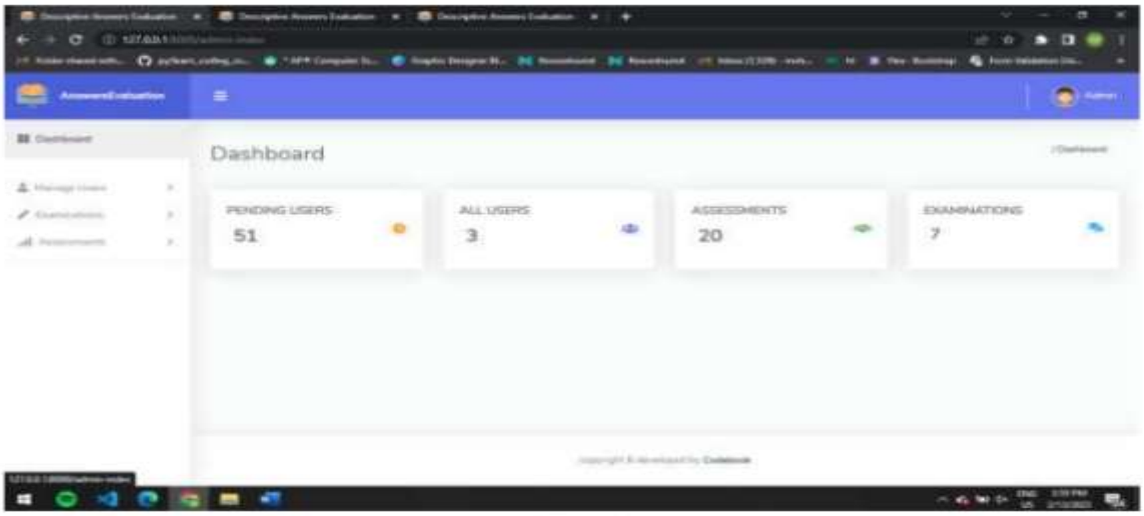


Fig.13

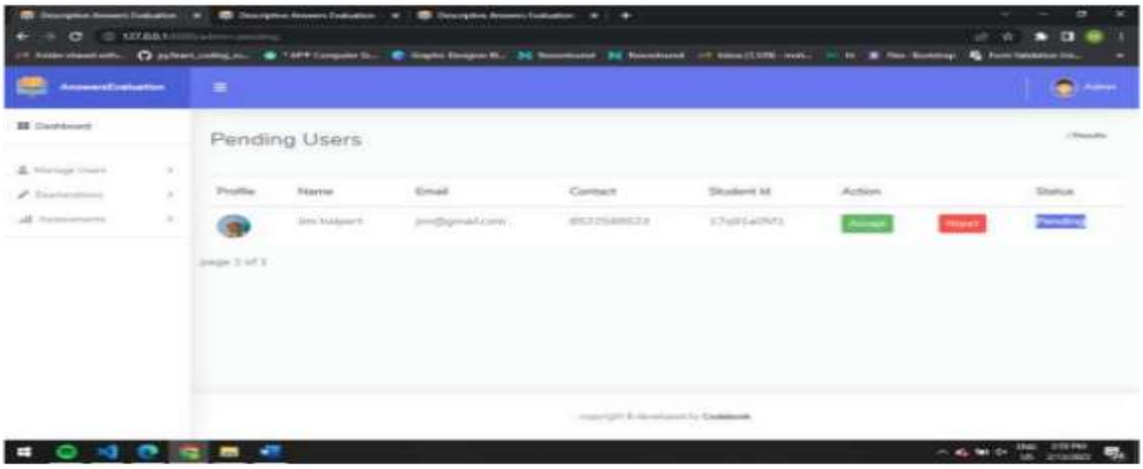


Fig.14

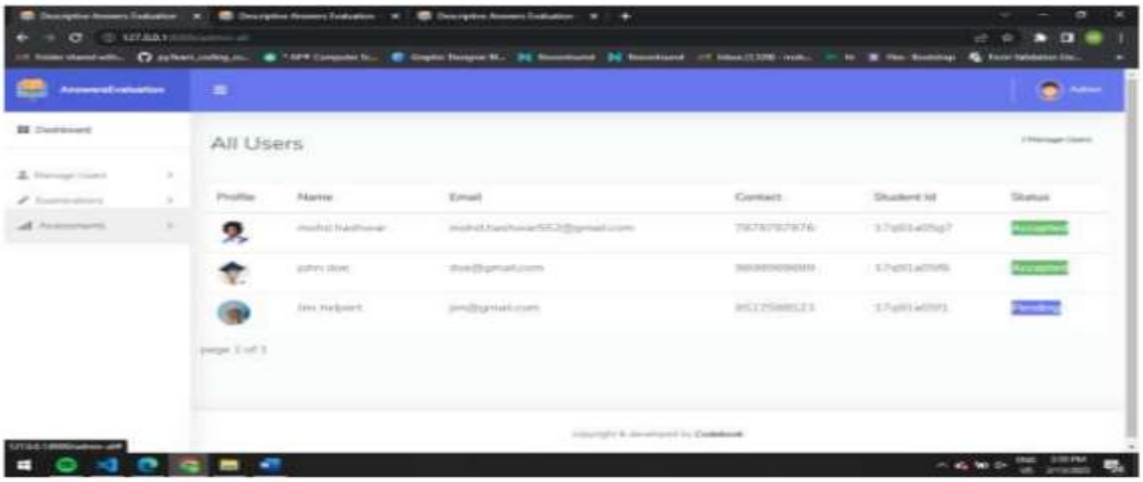

All Users
Profile
Name
Email
Contact
Student Id
Status

Fig.15

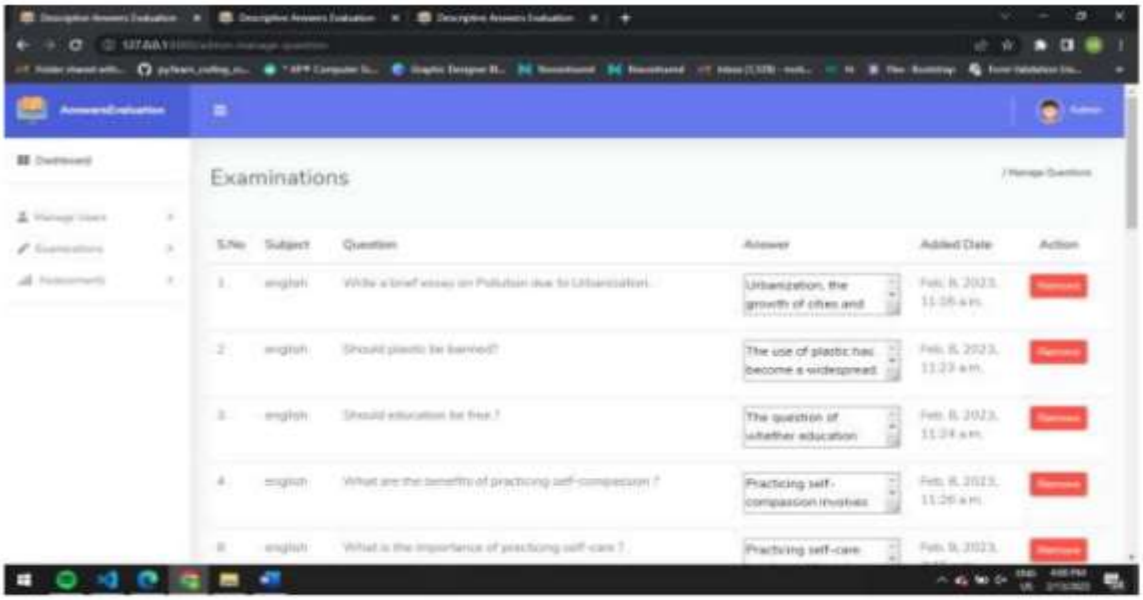

Examinations
S.No
Subject
Question
Answer
Added Date
Action
Urbanization, the growth of cities and
The use of plastic has become a widespread
The question of whether education
Practicing self-compassion involves
Practicing self-care

Fig.18

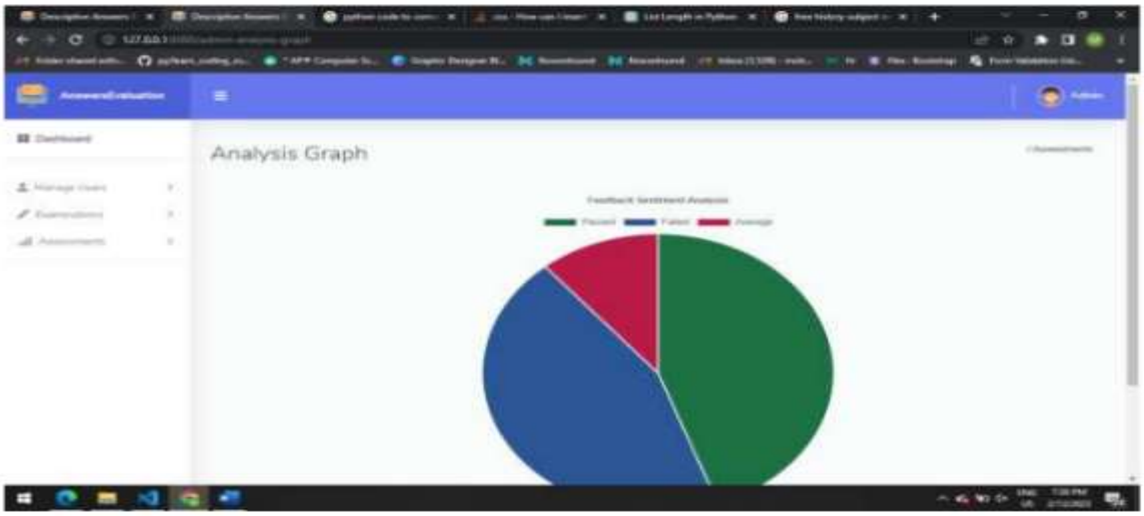


Fig.20

8. CONCLUSIONS

This paper proposed a novel approach to subjective answers evaluation based on machinelearning and natural language processing techniques. Two score prediction algorithms areproposed, which produce up to 88% accurate scores. Various similarity and dissimilaritythresholds are studied, and various other measures such as the keyword's presence andpercentage mapping of sentences are utilized to overcome the abnormal cases of semanticallyloose answers. The experimentation results show that on average word2vec approach performsbetter than traditional word embedding techniques as it keeps the semantics intact. Furthermore,Word Mover's Distance performs better than Cosine Similarity in most cases and helps train themachine learning model faster. With enough training, the model can stand on its own and predictscores without the need for any semantics checking. In terms of future improvements, theword2vec model can be trained especially for subjective answers evaluation of a particulardomain, and with large data sets, the number of classes or grades in the model can be significantlyincreased. Subjective answers evaluation remains an interesting problem to tackle, and in thefuture, we hope to find more efficient ways to solve this problem.

6. REFERENCES

[1] J. Wang and Y. Dong, ''Measurement of text similarity: A survey,'' Information, vol. 11, no. 9, p. 421, Aug. 2020.

[2] M. Han, X. Zhang, X. Yuan, J. Jiang, W. Yun, and C. Gao, ''A survey on the techniques, applications, and performance of short text semantic similarity,'' Concurrency Comput., Pract. Exper., vol. 33, no. 5, Mar. 2021.

[3] M. S. M. Patil and M. S. Patil, ''Evaluating Student descriptive answers using natural

language processing,'' Int. J. Eng. Res. Technol., vol. 3, no. 3, pp. 1716–1718, 2014.

[4] P. Patil, S. Patil, V. Miniyar, and A. Bandal, ''Subjective answer evaluation using machine learning,'' Int. J. Pure Appl. Math., vol. 118, no. 24, pp. 1–13, 2018.

[5] J. Muangprathub, S. Kajornkasirat, and A. Wanichsombat, ''Document plagiarism detection using a new concept similarity in formal concept analysis,'' J. Appl. Math., vol. 2021, pp. 1–10, Mar. 2021.

[6] V. Suma and Shavige Malleshwara Hills, "Data Mining based Prediction of Demand in Indian Market for Refurbished Electronics," Journal of Soft Computing Paradigm (JSCP) 2, pp. 101-110, 2020.

[7] M. Kusner, Y. Sun, N. Kolkin, and K. Weinberger, ''From word embeddings to document distances,'' in Proc. Int. Conf. Mach. Learn., 2015, pp. 957–966.

[8] C. Xia, T. He, W. Li, Z. Qin, and Z. Zou, ''Similarity analysis of law documents based on Word2vec,'' in Proc. IEEE 19th Int. Conf. Softw.Qual., Rel. Secur. Companion (QRS-C), Jul. 2019, pp. 354–357.

[9] H. Mittal and M. S. Devi, ''Subjective evaluation: A comparison of several statistical techniques,'' Appl. Artif. Intell., vol. 32, no. 1, pp. 85–95, Jan. 2018.

[10] L. A. Cutrone and M. Chang, ''Automarking: Automatic assessment of open questions,'' in Proc. 10th IEEE Int. Conf. Adv. Learn. Technol., Sousse, Tunisia, Jul. 2010, pp. 143–147.

[11] G. Srivastava, P. K. R. Maddikunta, and T. R. Gadekallu, ''A two-stage text feature selection algorithm for improving text classification,'' Tech. Rep., 2021.

[12] H. Mangassarian and H. Artail, ''A general framework for subjective information extraction 68from unstructured English text,'' Data Knowl. Eng., vol. 62, no. 2, pp. 352–367, Aug. 2007.

[13] B. Oral, E. Emekligil, S. Arslan, and G. Eryigit, ''Information extraction ˇ from text intensive and visually rich banking documents,'' Inf. Process. Manage., vol. 57, no. 6, Nov. 2020, Art. no. 102361.

[14] H. Khan, M. U. Asghar, M. Z. Asghar, G. Srivastava, P. K. R. Maddikunta, and T. R. Gadekallu, ''Fake review classification using supervised machine learning,'' in Proc. Pattern

Recognit. Int. Workshops Challenges (ICPR). New York, NY, USA: Springer, 2021, pp. 269–288.

[15] W. K. Michael and A. G. Thomas, "A Framework for the Evaluation of Statistical Prediction Models," CHEST, vol. 158, no. 1, pp. S29-S38, 2020.

[16] N. Madnani and A. Cahill, ‘‘Automated scoring: Beyond natural language processing,’’ in Proc. 27th Int. Conf. Comput.Linguistics (COLING), E. M. Bender, L. Derczynski, and P. Isabelle, Eds. Santa Fe, NM, USA: Association for Computational Linguistics, Aug. 2018, pp. 1099–1109.

[17] S. Deepa, A. Alli, Sheetac and S. Gokila, "Machine learning regression model for material synthesis prices prediction in agriculture," in materialstoday, 2021.

[18] G. Grefenstette, ‘‘Tokenization,’’ in Syntactic Wordclass Tagging. Springer, 1999, pp. 117–133.

[19] K. Sirts and K. Peekman, ‘‘Evaluating sentence segmentation and word Tokenization systems on Estonian web texts,’’ in Proc. 9th Int. Conf. Baltic (HLT) (Frontiers in Artificial Intelligence and Applications) vol. 328, U. Andrius, V. Jurgita, K. Jolantai, and K. Danguole, Eds. Kaunas, Lithuania: IOS Press, Sep. 2020, pp. 174–181.

[20] S. Matthew and Lewis, "Identifying airline price discrimination and the effect of competition," International Journal of Industrial Organization, vol. 78, 2021.