iJRASET
International Journal For Research in Applied Science and Engineering Technology

# INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

**Volume:** 9 **Issue:** XI **Month of publication:** November 2021

**DOI:** https://doi.org/10.22214/ijraset.2021.39090

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
*ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429*
*Volume 9 Issue XI Nov 2021- Available at www.ijraset.com*

# Subjective Answer Evaluator

Sarthak Kagliwal[1], Jagruti Agrawal[2], Tejas Dahad[3], Atharva Saraf [4], Karan Kangude[5]
[1, 2, 3, 4, 5]*Computer Engineering, Pune Institute of Computer Technology, Pune, Maharashtra, India*

***Abstract:*** ***The automatic assessment of subjective replies necessitates the use of Natural Language Processing and automated assessment. Ontology, semantic similarity matching, and statistical approaches are among the strategies employed. But most of the methods are based on an unsupervised approach. The proposed system uses an unsupervised method and is divided into two modules. The first one is extracting the essential data through text summarization and the second is applying various Natural Language models to the text retrieved from the above step and giving marks to them.***
***Keywords: Automatic Evaluation, NLP, Text Summarization, Similarity Measure, Marks Scoring.***

## I. INTRODUCTION

Descriptive examinations are used to assess students' comprehensive understanding of topics. This is in contrast to traditional types of answers, such as multiple-choice questions or fill-in-the-gap items, in which the student's understanding is limited to the options presented and thus not thoroughly examined. Subjective answer evaluation can be challenging. Subjective answers necessitate a significant amount of time and effort on the part of the evaluator. Uniformity is achieved in assessing by using computers to perform evaluations, as the same mechanism is used for all responses.

Automatic Response Grading is the process of grading essays without the use of humans, in which a student's answer and a model answer created for a certain prompt are used as input, and the answer is given a numeric score based on its semantic similarity.

Neural networks are used to train LMs in modern approaches. In the majority of natural language processing tasks, language models have proven to be effective. In a variety of ways, neural language models alleviate the shortcomings of traditional approaches. They broaden the context considered and demonstrate a high level of generalization across different contexts. The first neural models made use of recurrent neural networks, such as long short-term memory networks. The transformer architecture underwrites the most recently introduced neural models for language modeling, such as BERT and XLNET. The performance of several NLP tasks has been robustly increased with the implementation of these models.

In this paper, a reference-based technique is used to train an NLP model on one of Hewlett Foundation's most difficult datasets. On this dataset, two models based on BERT and XLNET are finetuned. The key benefit of the methodology given here is that, unlike most other approaches, grading is done without manually extracting features.

## II. MOTIVATION

One of the most essential techniques for evaluating students' academic success is subjective answers. Because the student-to-evaluator ratio is so high, manually grading them becomes an extremely time-consuming operation for the evaluators. Consequently, manual grading becomes arduous and cost-ineffective. Furthermore, manual grading is a contentious matter due to divergent perspectives among evaluators, each of whom may favor a distinct voice or writing style. As a result, there is a need for an Automated Answer Scoring Methodology to facilitate and simplify the traditional system while reducing subjectivity.

## III.BACKGROUND WORK

Most of the systems proposed before are suitable for short-length responses and do not work properly on responses having lengths greater than a certain limit. [5] In the Intelligent Essay Assessor, the LSA model is used. LSA creates a high-dimensional semantic representation of the information in the domain based on this training. Words from the text are represented as vectors in this semantic space, with the semantic similarity between words defined by the cosine of the angle between the vectors for those words. [6] In ETS, task-specific features like n-grams are combined with more general features like the semantic similarity of the response to model answers to create the proposed system. [3] In Fast and Easy Short Answer Grading with High Accuracy, their model uses features like alignment, semantic vector similarity, question demoting, term weighting, and length ratio to improve the similarity. But the drawback is, it works only on short-length responses. [4] Propose a method that combines token and sentence level characteristics for grading short responses.[1] They developed a broad framework for both abstractive and extractive summarization, as well as a novel document-level encoder based on pretrained BERT. Experiment findings across three datasets revealed that their model produces state-of-the-art outcomes under both automatic and human-based evaluation processes.

However, they do not have language generating skills. [14] They provided a Bayesian method to essay grading based on the well-developed literature on text categorization. Their preliminary assessment of the technique based on one item, a sparse dataset, and just two classifications look promising. According to their research, scoring based on arguments outperforms scoring based on keywords or key phrases. The computer used brute force to identify arguments in their study.[18] It's an essay scoring model and restricted to a limited area of study based on dataset.

## IV. PROPOSED METHOD

### A. *Dataset*

The William and Flora Hewlett Foundation used data from a competition held on kaggle.com. The data is made up of eight sets of essays written by students in grades 7 through 10. Each of the eight sets of essays has its own set of characteristics that are used to grade them. This ensures that the automated grader is well-versed in various types of essays. Each essay has a resolved score as well as one or more human scores. Each essay set has its own grading rubric, which is usually holistic but in one case is trait-based. Each essay ranges in length from 150 to 1200 words. Some essays are more reliant on secondary sources than others.

### B. *Text Extraction*

Students' answers can be in the written format on some piece of paper. We want this answer to be in string format in order to provide it as input for our model. A python class called pytesseract was used to extract text from a picture. The extracted text has been used for further processing.

### C. *Pre-processing*

The Summarized text contains some words which carry less important information and can be ignored to facilitate further text processing tasks. After the extraction of each answer, all the non-useful contents such as links, symbols, extra spaces are removed. For this, a python library called regex is used. Removal of stopword and lemmatization is not used here, as it leads to loss of information. After this, the dataset is split into train and test datasets using the scikit learns train test split module in the ratio of 9: 1 respectively. The dataset contains 8 sets of essays on eight different sets of topics making a total of 12978 essays. Each essay is graded by at least two graders. Some set of essays is graded by more than two raters, so for these essays scores from only the first two graders is considered for consistency. The average of both the scores is taken and converted such that the score falls in the range from $0 - 5$. In this dataset, no model or reference answer is provided. So the set of answers with the max score from both the raters is considered as the reference answers. Later for a particular set, instead of keeping the same model answers for all the other answers, random answers are taken from the created pool of model answers. Most of the transformer-based models have limitations for the length of the input sequence, which is around 512 tokens. As we are dealing with long answers rather than short ones, the length of the answer will be mostly greater than 512 words. For this problem, the input sequence is stirred into sequences of smaller lengths. For a particular instance, both the student's answer and the model answer are stirred into equal no of parts.

### D. *Generating the Score*

ASAG employing transformer-based architectures may be thought of as learning the textual entailment between a student response and a model answer. In what follows, the model architecture and input representation for both BERT and XLNET is briefly present. For this experiment both BERT Base and XLNET Base are used. There are 12 layers, 12 attention heads and 768 neurons for these two models. The number of total parameters is 110 million. For BERT, cased and uncased versions are trained and released. However, only the cased version is released with XLNET

The hyperparameters for all the grading were experimentally selected and are as follows both BERT and XLNET:

*1)* Epochs = 4
*2)* Batch Size = 6
*3)* Dropout probability for all the layers = 0.1
*4)* Learning rate = 3e-5

To avoid overfitting/underfitting, one of the most common regularization techniques is employed, which is to stop the training process on a set of initial epochs before full convergence. This is done in combination with the dropout technique. We set the number of epochs to 4 experimentally and trained our model for the full number of epochs. Then, we analysed the flow of loss change per epoch to control the overfitting and underfitting. In the end, models are evaluated on the validation sets and stopped the training process before the occurrence of any of these two problems. The training at is stopped the 2-4 epoch for all the mode

*E. System Design*

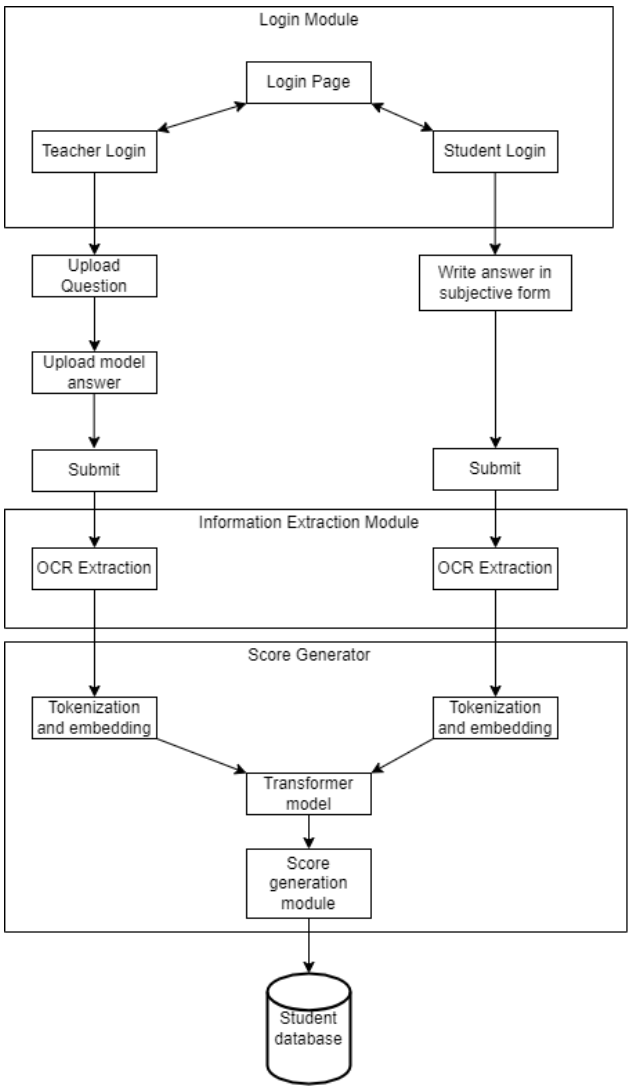



Fig. 1 Architecture Diagram

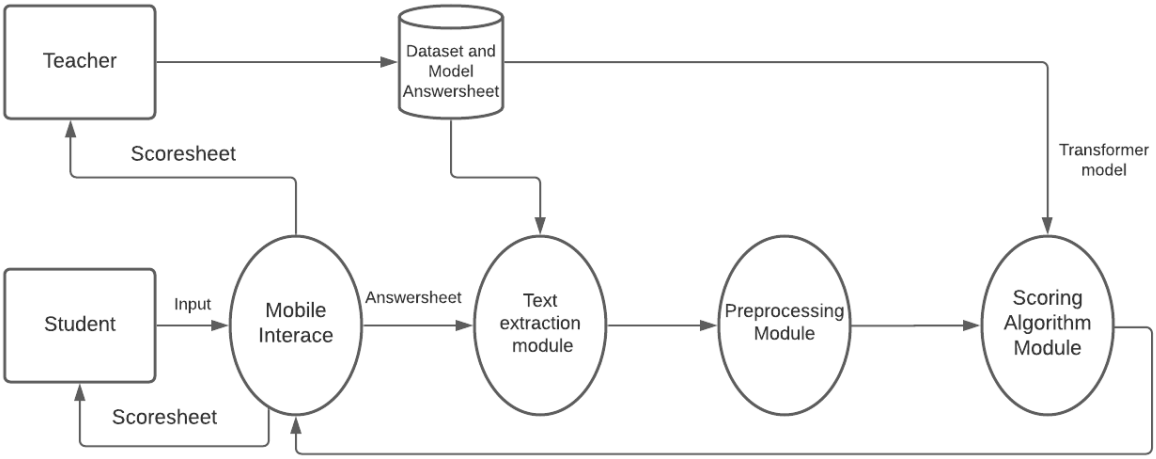



Fig. 2 Activity Diagram

## V. RESULTS

### A. Evaluation Metrics

1) *Accuracy:* Proportion of correctly graded answers.
2) *Macro Average F1 score:* Precision, recall and F1 scores are calculated independently for each grade label and then averaged over all grade labels

Results

| Models | Dataset | Acc | M-F1 |
|---|---|---|---|
| ETS (Heilman and Madnani, 2013) | Beetle And SciEntsBank | 0.643 | 0.478 |
| Sultan et, al. (Sultan et al., 2016 | Mohler Dataset | 0.489 | 0.3298 |
| MEAD(Ramachandra n and Foltz, 2015) | ASAP-SAS, 2012 | - | 0.379 |
| Mavarniya (Mavarniya et al., 2018) | Beetle and SciEntsBank | - | 0.579 |
| BERT Base uncased | ASAP-SAS, 2012 | 0.66 | 0.484 |
| XLNET Base cased | ASAP-SAS, 2012 | 0.658 | 0.47 |

BERT Base uncased achieved the top values for all the evaluation measures in terms of accuracy. Close to BERT Base uncased, it is XLNET and BERT Base cased. Most of the other dataset consist of instances of relatively small when compared to the instance length in ASAP-AES dataset. Although both the transformer-based models performs well and gives promising results.
However, one possible reason for overall lower classification results is the limitation of the input length of the transformer models. Because of low input length, the dataset was needed to be stirred and by doing so the model may interpret the stirred sequences as individual answers, consequently giving overall less accuracy.

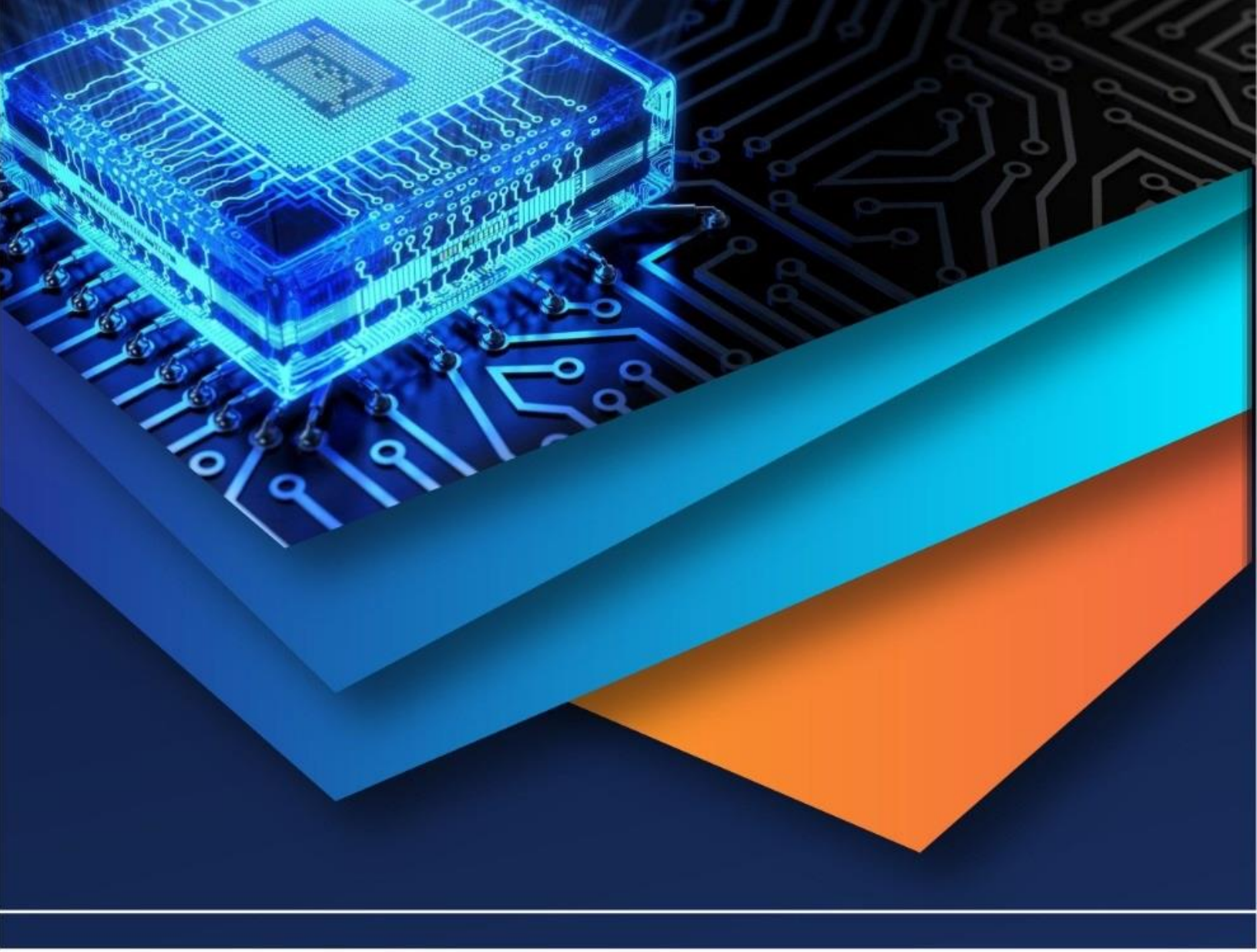
## VI. CONCLUSION

An Automatic Subjective Answer Grading system using modern language models such as BERT and XLNet was proposed. Overall, the approaches can be considered good enough for the long answers. These language model results appear to indicate that modern language models are successful in creating a semantic representation of student and model answers and correctly classifying them. In this approach, for handling a large size of answers extractive text summarization using a text rank algorithm was used. Further for generating the score, transformer-based models BERT and XLNET were implemented on the ASAP-AES Kaggle dataset. BERT and XLNet seem to either equal or outperform the results obtained with human-engineered features.

## REFERENCES

[1] Y. Liu and M. Lapata, "Text Summarization with Pretrained Encoders", Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), doi: 2019.10.18653/v1/D19-1387
[2] K. Yao, L. Zhang, D. Du, T. Luo, L. Tao and Y. Wu, "Dual Encoding for Abstractive Text Summarization," in IEEE Transactions on Cybernetics, vol. 50, no. 3, pp. 985-996, March 2020, doi: 10.1109/TCYB.2018.2876317.
[3] M. Sultan, C. Salazar and T. Sumner, "Fast and Easy Short Answer Grading with High Accuracy", Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2016. doi: 10.18653/v1/n16-1123 .
[4] S. Saha, T. Dhamecha, S. Marvaniya, R. Sindhgatta and B. Sengupta, "Sentence Level or Token Level Features for Automatic Short Answer Grading?:Use Both", Lecture Notes in Computer Science, pp. 503 517, 2018 . doi:10.1007/978 3 319 93843 137
[5] Foltz, Peter & Laham, Darrell & Landauer, T.. (1999). The intelligent essay assessor: Applications to educational technology. Interactive Multimedia Electronic Journal of Computer-Enhanced Learning.
[6] Heilman, Michael and Nitin Madnani. "ETS: Domain Adaptation and Stacking for Short Answer Scoring." *SEMEVAL (2013).
[7] L. Ramachandran, J. Cheng and P. Foltz, "Identifying Patterns For Short Answer Scoring Using Graph-based Lexico-Semantic Text Matching", Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications, 2015. doi: 10.3115/v1/w15-0612
[8] Das, B. Sharma, S. Rautaray and M. Pandey, "An Examination System Automation Using Natural Language Processing", 2019 International Conference on Communication and Electronics Systems (ICCES), 2019. doi: 10.1109/icces45898.2019.9002048

[9] N. Reimers and I. Gurevych, ”Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks”, Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 2019. doi: 10.18653/v1/d19-1410
[10] M. Ahmed, C. Dixit, R. E. Mercer, A. Khan, M. R. Samee and F. Urra, ”Multilingual Semantic Textual Similarity using Multilingual Word Representations,” 2020 IEEE 14th International Conference on Seman tic Computing (ICSC), San Diego, CA, USA, 2020, pp. 194- 198, doi: 10.1109/ICSC.2020.00040
[11] Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. 2019. Unified language model pre-training for natural language understanding and generation. arXiv preprint arXiv:1905.03197.
[12] P. D. Turney and P. Pantel, “From frequency to meaning: Vector space models of semantics,” J. Artif. Intell. Res., vol. 37, pp. 141–188, 2010.
[13] T. Kakkonen, N. Myller, E. Sutinen, and J. Timonen, “Comparison of dimension reduction methods for automated essay grading,” Educ. Technol. Soc., vol. 11, no. 3, pp. 275–288, 2008
[14] L. Rudner and T. Liang, “Automated essay scoring using Bayes’ theorem,” J. Technol. Learn. …, vol. 1, no. 2, 2002.
[15] P. Diana, A. Gliozzo, C. Strapparava, E. Alfonseca, P. Rodr, and B. Magnini, “Automatic Assessment of Students ’ free-text Answers underpinned by the Combination of a B LEU -inspired algorithm and Latent Semantic Analysis,” Mach. Transl., 2005
[16] W. Wang and B. Yu, Text categorization based on combination of modified back propagation neural network and latent semantic analysis, Neural Comput. Appl., vol. 18, no. 8, pp. 875881, 2009.
[17] F. Noorbehbahani and a. a. Kardan, The automatic assessment of free text answers using a modified BLEU algorithm, Comput. Educ., vol. 56, no. 2, pp. 337345, 2011.
[18] J. Burstein, K. Kukich, S. Wolff, C. Lu, M. Chodorow, L. Braden-Harder, and M. D. Harris, Automated scoring using a hybrid feature identification technique, Proc. 17th Int. Conf. Comput.Linguist. -, vol. 1, p. 206, 1998.
[19] L. Bin, L. Jun, Y. Jian-Min, and Z. Qiao-Ming, Automated essay scoring using the KNN algorithm,Proc. - Int. Conf. Comput. Sci. Softw. Eng. CSSE 2008, vol. 1, pp. 735738, 2008.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

# INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY