DESCRIPTIVE ANSWER EVALUATION SYSTEM BASED ON COSINE SIMILARITY

K.VijayKumar[1],N. Mounika[2],T. Amulya[3],Pavani[4],Keerthi[5]
Assistant Professor, Department of Computer Science and Engineering[1]
Student, Department of Computer Science and Engineering[2,3,4,5]
Sree Dattha Institute of Engineering and Science, Sheriguda, Telangana. [1,2,3,4,5]

ABSTRACT:
Manually reviewing subjective articles is a laborious process fraught with challenges related to data understanding and acceptance, which impedes the effective use of AI for evaluation. Many have explored using computer technology to assess student answers, often relying on traditional methods or specific terminology, yet validated datasets remain scarce. This paper proposes a novel approach to automatically evaluate descriptive responses by integrating machine learning, natural language processing, and toolkits such as Wordnet, Word2vec, WMD, cosine similarity, MNB, and TF-IDF. Solution statements and keywords are leveraged for evaluation, and a machine learning model is trained to predict grades. Results indicate that WMD outperforms cosine similarity in effectiveness. With sufficient training, the machine learning model can operate independently. Experimental findings demonstrate an 88% accuracy without MNB, which improves by 1.3% when MNB is included.

Index Terms: Subjective article evaluation, automatic assessment, machine learning, natural language processing, Wordnet, Word2vec, WMD, cosine similarity, MNB, TF-IDF.
DOI Number: 10.48047/nq.2024.22.4.nq24011 **NeuroQuantology 2024; 22(4):97-105**

I.INTRODUCTION

One way to evaluate a student's progress and competence is using subjective questions and answers, which are known for their open-ended character. Students are encouraged to compose their solutions from their own experiences and expertise on the given subject, since there are no restrictions on this. But there are many more important differences between subjective and objective answers. In contrast to the objective questions, this one is noticeably longer. Secondly, it takes more time to write them. Because of the extra background they convey, they also need the instructor's full focus and objectivity.

The ambiguity present in natural language is the most significant of several reasons why computer assessment of such questions is difficult. Data must be cleaned and tokenized, among other preparatory steps, before it can be used. Document similarity, ontologies, idea networks, and latent semantic structures are some of the methods that may be used to compare textual data afterwards. To get at the final score, several factors might be considered, including language, structure, keyword presence, and similarity. Prior attempts to resolve this problem have been explored in this article, along with potential avenues for improvement. Because of their one constant quality—context—both students and teachers regard subjective assessments with a reasonable amount of fear and difficulty. The

scorer must carefully examine each word in order for an answer to be deemed subjective; yet, the final score is greatly influenced by the checker's emotional condition, degree of fatigue, and degree of impartiality.

It is more economical to have a computer assess subjective replies due to the necessity and complexity of this task. Objective responses may be quickly and realistically evaluated by machines. It is possible to quickly map students' responses by entering questions and one-word answers into a computer. On the other hand, it's much more challenging to address personal ideas. Their vocabulary is extensive, and they discuss many different subjects. To add insult to injury, people often resort to convenient acronyms and synonyms, thus adding to the complexity.

II.LITERATURE SURVEY

Jiapeng Wang and Yihong Dong proposed that text similarity measurement is the basis of natural language processing tasks, which play an important role in information retrieval, automatic question answering, machine translation, dialogue systems, and document matching. This paper systematically reviews the research status of similarity measurement, analyzes the advantages and disadvantages of current methods, develops a more comprehensive classification description system of text similarity measurement algorithms, and summarizes the future development directions. With the aim of providing references for related research and application, the text similarity measurement method is described in two aspects: text distance and text representation. Text distance can be divided into length distance, distribution distance, and semantic distance; text representation is 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 summarized in the discussion section.

- Wei Yun and Chen Gao proposed that short text similarity plays an important role in natural language processing (NLP) and has been applied in many fields. Due to the lack of sufficient context in short texts, it is difficult to measure the similarity. The use of semantic similarity to calculate textual similarity has attracted the attention of academia and industry and achieved better results. In this survey, we have conducted a comprehensive and systematic analysis of semantic similarity. We first propose three categories of semantic similarity: corpus-based, knowledge-based, and deep learning (DL)-based. We analyze the pros and cons of representative and novel algorithms in each category. Our analysis also includes the applications of these similarity measurement methods in other areas of NLP. We then evaluate state-of-the-art DL methods on four common datasets, which proved that DL-based methods can better solve the challenges of short text similarity, such as sparsity and complexity. Especially, the bidirectional encoder representations from transformers model can fully employ scarce information of short texts and semantic information, obtaining higher accuracy and F1 values. We finally put forward some future directions.

- S. Patil and Sonal Patil proposed that computer-assisted assessment of free-text answers has seen significant developments in recent years due to the need to evaluate a deep understanding of lesson concepts, which most educators and researchers agree cannot be done by simple MCQ testing. In this paper, we have reviewed the techniques underpinning this system, described currently available systems for marking short free-text responses, and finally proposed a system that evaluates descriptive type answers using Natural Language Processing.

- Jirapond Muangprathub, Siriwan Kajornkasirat, and Apirat Wanichsombat propose an algorithm for document plagiarism detection using

incremental knowledge construction with formal concept analysis (FCA). The incremental knowledge construction supports document matching between the source document in storage and the suspect document. A new concept similarity measure is also proposed for retrieving formal concepts in the knowledge construction. The proposed concept similarity measure employs appearance frequencies in the obtained knowledge construction. Our approach can be applied to retrieve relevant information because the obtained structure uses FCA in concept form, which is definable by a conjunction of properties. This measure is mathematically proven to be a formal similarity metric. The performance of the proposed similarity measure is demonstrated in document plagiarism detection. Moreover, this paper provides an algorithm to build the information structure for document plagiarism detection. Thai text test collections are used for performance evaluation of the implemented web application.

- Muhammad Farrukh Bashir and Hamza Arshad introduce a pioneering method that harnesses a range of machine learning and natural language processing techniques to automate the evaluation of descriptive answers. Their approach integrates solution statements and keywords to assess responses, employing a machine learning model trained specifically to predict answer grades. With sufficient training, their model has the potential to operate independently as a standalone tool. In their experiments, they achieved an impressive accuracy rate of 97%, highlighting the efficacy of artificial intelligence in addressing such evaluative challenges. They enhance their methodology by incorporating deep learning techniques and preprocessing steps, aiming to further refine the accuracy and reliability of answer evaluations. This innovative framework represents a significant advancement in the field of automated assessment, leveraging cutting-edge technologies to streamline and improve the grading process for descriptive answers.

III.PREVIOUS WORK

Existing systems for evaluating subjective papers predominantly rely on manual assessment, which is time-consuming and susceptible to subjective biases. Efforts to automate this process using AI and ML techniques, such as NLP for sentiment analysis and keyword extraction, aim to enhance efficiency and reduce bias. However, current automated systems often struggle with nuanced evaluations and depend heavily on simplistic metrics like word counts and keyword frequencies. Challenges include the scarcity of diverse and validated datasets, limiting the robustness and generalizability of AI models for comprehensive subjective paper assessment.

IV.PROPOSED MODEL

The proposed model in this paper integrates a variety of advanced machine learning and natural language processing techniques alongside key tools such as Wordnet, Word2vec, Word Mover's Distance (WMD), cosine similarity, Multinomial Naive Bayes (MNB), and Term Frequency-Inverse Document Frequency (TF-IDF) to automate the evaluation of descriptive answers. This approach utilizes solution statements and keywords to assess responses and trains a machine learning model to predict answer grades. Experimental findings demonstrate that WMD consistently outperforms cosine similarity in effectiveness. Furthermore, with sufficient training, the model has the potential to operate independently as a standalone tool. Initial experimentation achieves an impressive accuracy rate of 88%, which further improves by 1.3% when incorporating the MNB model. This innovative framework aims to streamline and enhance the efficiency of subjective answer evaluation processes, leveraging cutting-edge technologies to achieve more precise and reliable grading outcomes.

V.SYSTEM ARCHITECTURE

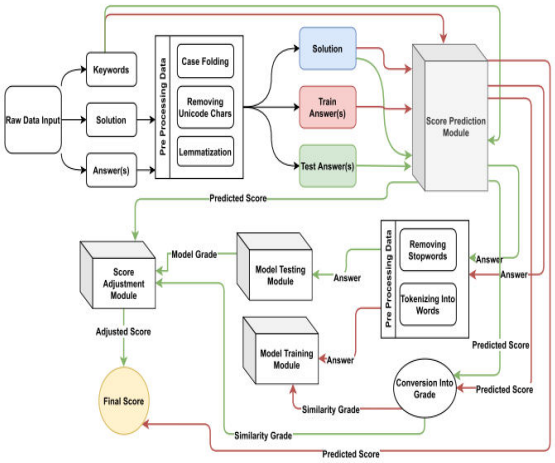


Figure .1 System Architecture

VI.IMPLEMENTATION MODULES

A. Keywords

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

B. Solution

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

C. Answer

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

D. Data Collection

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

E. Data Annotation

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

F. Preprocessing Module

After taking inputs from the user, both the solution and the answer go through some preprocessing steps, which involve tokenization, stemming, lemmatization, stop words removal, case folding, finding, and attaching synonyms to the text. Note that stop words are not removed when passing the data to word2vec because word2vec contains a vast vocabulary and can utilize those stop words to make better semantic sense of the text. However, stop words are removed before passing to a machine learning model such as Multinomial Naive Bayes because they hinder the machine's ability to learn the patterns.

G. Similarity Measurement Module

This module consists of WDM and Cosine Similarity functions which take two sentences or word vectors and return their Similarity. WDM tells us the dissimilarity while Cosine Similarity measures Similarity. Our approach uses both of these similarity measures one at a time and compares the results at the end. Various similarity (or dissimilarity) thresholds . 1) THRESHOLDS ANALYSIS Various thresholds

used in this paper have been experimentally deduced to produce the optimal result. WDM thresholds of WDM_LOWER and WDM_UPPER represent the dissimilarity between two sentences, where more dissimilarity represents high similarity. 0.7 threshold for WDM_LOWER was experimentally observed to represent semantically very similar sentences, and 1.6 thresholds for WDM_UPPER were observed to represent semantically less similar sentences. Anything beyond 1.6 is assumed to be too dissimilar to consider viable for comparison. Similarly, Cosine similarity thresholds COS_LOWER and COS_UPPER represent the similarity between two sentences. It should be noted that cosine similarity does not take the context of two sentences into account when measuring similarity as opposed to WDM, hence the usage of both of these similarity (or dissimilarity) measuring approaches.

H. 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 our module using either WDM or Cosine Similarity while considering the maximum matched solution/answer sentence pairs. This result can be compared to an actual score or fed into a machine learning model to be trained.

I. Machine Learning Model Module

This model consists of Machine learning models trained on the data obtained from the result prediction module.

Its working is as follows:

• Input data from Result Prediction Module.

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

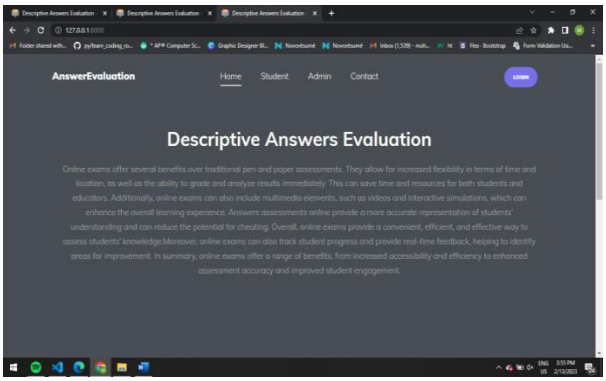
• Convert the overall score obtained from Result Prediction Module into some category. Four categories A, B, C, and D, are used in the paper, representing 1st, 2nd, 3rd, and 4th quarter of a 100. For example, A represents marks from 0 to 25, and B represents 26 to 50.

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

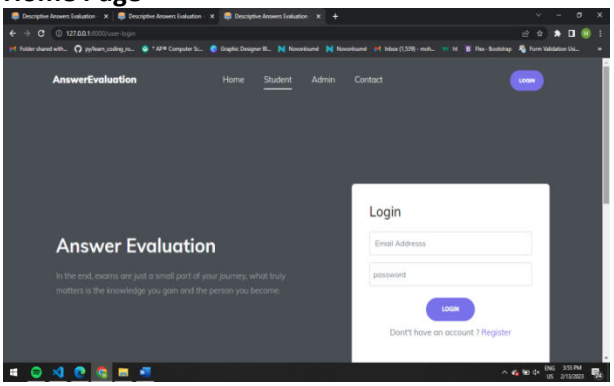
• A machine learning model such as Multinomial Naive Bayes, which performs well for multi-class classification, is chosen.

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

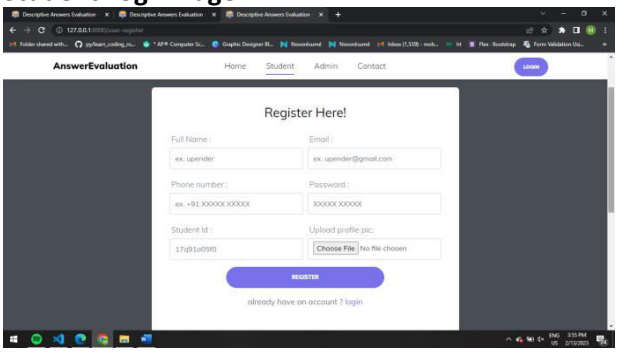
VII.RESULTS:



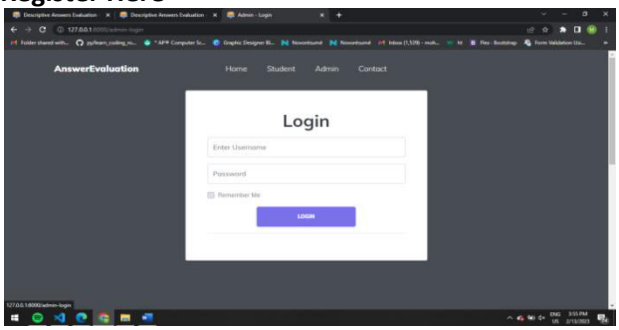
Home Page



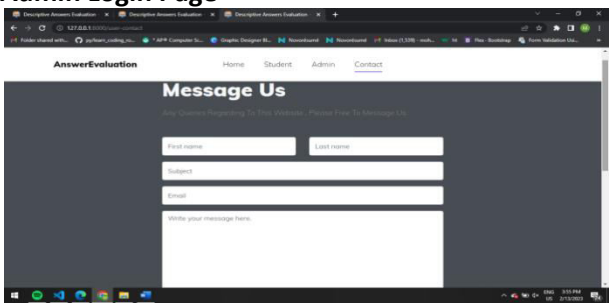
Student Login Page



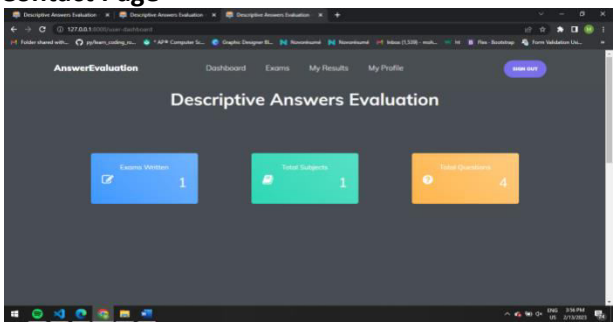
Register Here



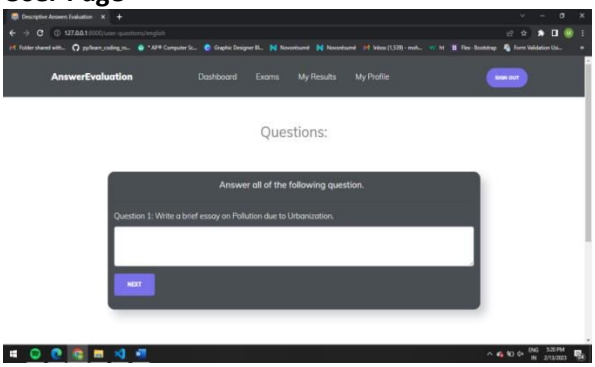
Admin Login Page



Contact Page

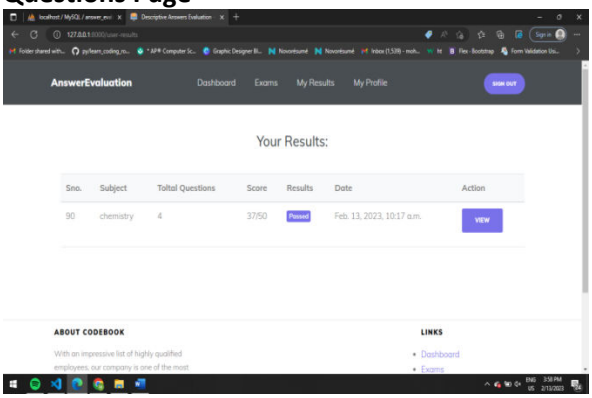


User Page

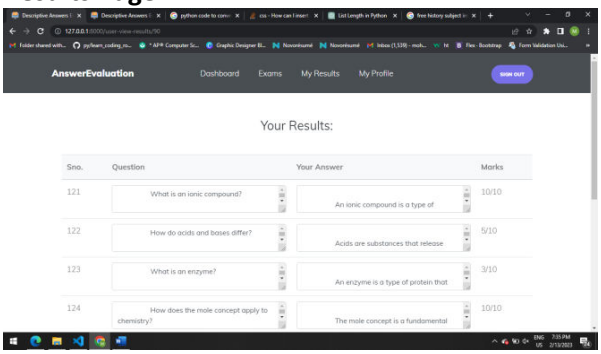


Questions Page

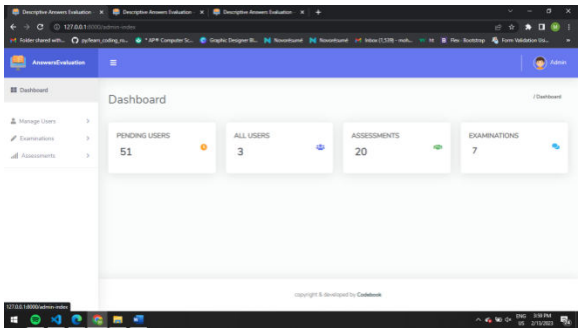
Questions Page



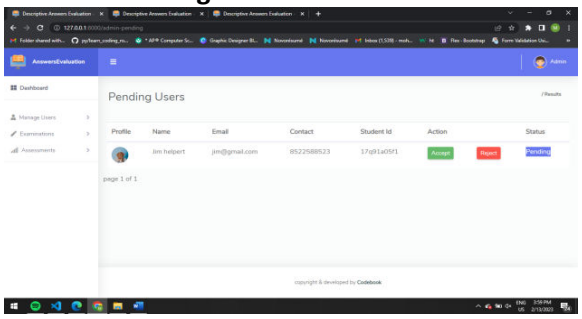
Results Page



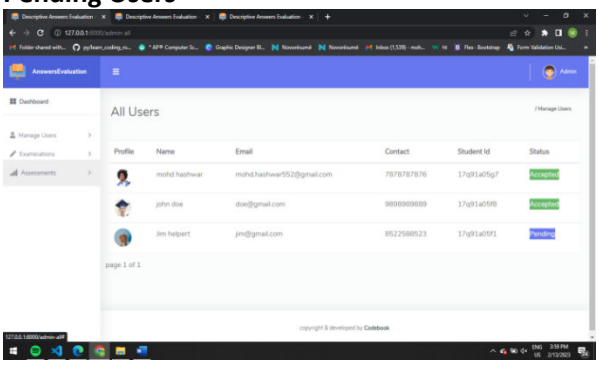
View Results Page



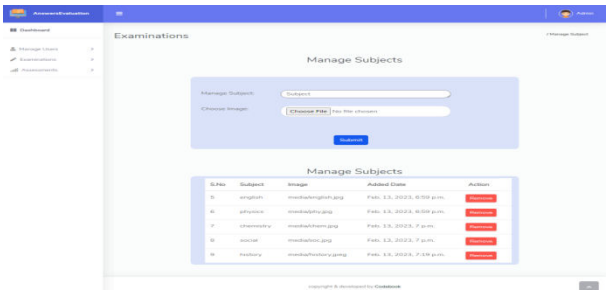
Admin Home Page



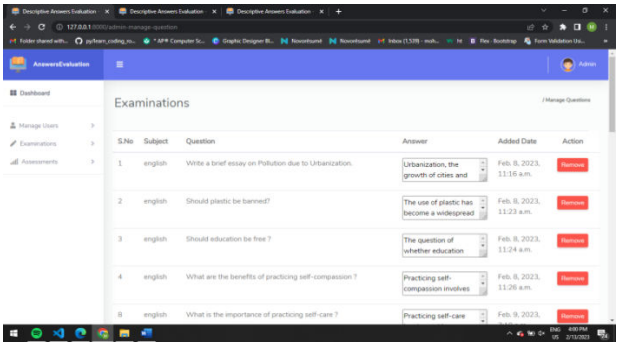
Pending Users



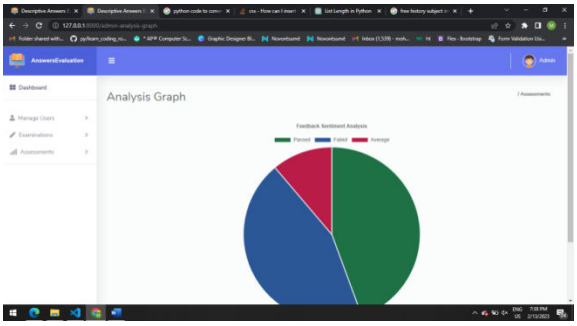
ALL USERS PAGE



Examinations Page



Add Questions Page



Analysis Page

VIII.CONCLUSION

Using natural language processing and artificial intelligence, we present a novel approach to assessing subjective replies in this research. Score predictions using the two proposed approaches are accurate up to 88% of the time. To address cases where answers may lack semantic coherence, we incorporate additional metrics such as keyword presence and sentence percentage mapping, alongside various similarity and dissimilarity criteria. Experimental results indicate that word2vec generally outperforms traditional word embedding methods by preserving semantic integrity.

IX.FUTURE ENHANCEMENT

In contrast to Cosine Similarity, Word Mover's Distance facilitates quicker training of machine learning models. Semantic verification becomes unnecessary as the model can autonomously predict scores with adequate training. With extensive datasets, the word2vec model can accommodate a broader range of classes or grades, making it suitable for domain-specific evaluations of subjective responses and a promising avenue for future enhancements.

eISSN1303-5150

Moving forward, our aim is to discover more efficient solutions to the intriguing challenge of evaluating subjective replies.

X.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] X. Hu and H. Xia, ‘‘Automated assessment system for subjective questions based on LSI,’’ in Proc. 3rd Int. Symp. Intell. Inf. Technol. Secur. Informat., Apr. 2010, pp. 250–254.

[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 from 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] S. Afzal, M. Asim, A. R. Javed, M. O. Beg, and T. Baker, ‘‘URLdeepDetect: A deep learning approach for detecting malicious URLs using semantic vector models,’’ J. Netw. Syst. Manage., vol. 29, no. 3, pp. 1–27, Mar. 2021.

[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] Z. Lin, H. Wang, and S. I. McClean, ‘‘Measuring tree similarity for natural language processing based information retrieval,’’ in Proc. Int. Conf. Appl. Natural Lang. Inf. Syst. (NLDB) (Lecture Notes in Computer Science), vol. 6177, C. J. Hopfe, Y. Rezgui, E. Métais, A. D. Preece, and H. Li, Eds. Cardiff, U.K.: Springer, 2010, pp. 13–23.

[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] A. Schofield, M. Magnusson, and D. M. Mimno, ‘‘Pulling out the stops: Rethinking stopword removal for topic models,’’ in Proc. 15th Conf. Eur. Chapter Assoc. Comput. Linguistics (EACL) vol. 2, M. Lapata, P. Blunsom, and A. Koller, Eds. Valencia, Spain: Association for Computational Linguistics, 2017, pp. 432–436.