Hindawi
Mathematical Problems in Engineering
Volume 2021, Article ID 9946573, 11 pages
https://doi.org/10.1155/2021/9946573

Research Article

Comparative Study on Feature-Based Scoring Using Vector Space Modelling System

Tarandeep Singh Walia,[1] Tarek Frikha ,[2] Omar Cheikhrouhou ,[3] and Habib Hamam[4]

[1]*School of Computer Applications, Lovely Professional University, Phagwara, Punjab, India*
[2]*Université de Sfax, CES Lab, Sfax 3038, Tunisia*
[3]*College of Computers and Information Technology, Taif University, Taif 21944, Saudi Arabia*
[4]*Faculty of Engineering, Uni de Moncton, NB, E1A3E9, Canada, IIT, Sfax, Tunisia*

Correspondence should be addressed to Tarek Frikha; tarek.frikha@enis.tn

Received 21 March 2021; Revised 25 April 2021; Accepted 8 May 2021; Published 28 May 2021

Academic Editor: Dr. Dilbag Singh

This paper shows the importance of automated scoring (AS) and that it is better than human graders in terms of degree of reproducibility. Considering the potential of the automated scoring system, there is further a need to refine and develop the existing system. The paper goes through the state of the art. It presents the results concerning the problems of existing systems. The paper also presents the semantic features that are indispensable in the scoring system as they have complete content. Moreover, in the present research, a huge deviation has been exhibited by the system which has been shown later in performance analysis of the study, and this clearly indicates the novelty and improved results of the system. It explains the algorithms included in the methodology of this proposed system. The novelty of our work consists in the use of its own similarity function and its notation mechanism. It does not use the cosine similarity function between two vectors. This paper describes and develops a more accurate system which employs a statistical method for scoring. This system adopts and integrates rule-based semantic feature analysis.

1. Introduction

In educational sectors, nearly every institute conducts various examination processes to evaluate the abilities of students. In this examination, student responses are evaluated for given questions. These questions can be both subjective answers or objective answers. In this research, evaluation of objective answer is not integrated and is a much trivial task.

Unlike multiple-choice questions (MCQ) in constructed-response (CR) questions, students write their own answers. They can express their own ideas and suitably support them to give their response to the text. Subjective question tests the adoptive ability of a student. But its assessment encounters several issues like synonymy, polysemy, and trickiness. The further categories of subjective answer-type evaluations include long answer essay and short answer essay. The long answer essays are also known as free-text answers in which contents and writing style are evaluated.

The scoring is generally done by extraction of grammatical and semantic relations from the student response and reference response [1]. The vector space model can be incorporated to correlate words as well as textual contexts from the student response with reference responses [2].

The remainder of this paper is organized as follows. In Section 2, we describe the state of the art of different scoring methods, both manual and automatic. Section 3 describes the existing system. It contains two important parts: the first one is predefined features (features construction, ranking, and selection). The second one is of related system. It contains the vector space model approach and other related concepts. Section 4 gives the proposed methodology. In section 5, a description of achieved objective is presented. We begin with the development of resources, the identification of predefined features, then the development of

statistical model, and finally the scoring mechanism. Section 6 is reserved for the results and the discussion. It contains the performance analysis and the cosine similarity comparison. Finally, in Section 7, we conclude and give recommendations and future work.

2. State of the Art

In this section, we will present the state of the art of different scoring methods, both manual and automatic.

2.1. Manual Scoring. Traditional mechanism carried out in the examination system was that the students were supposed to submit their answer sheets which were evaluated by the human rater. Since it is in use for long, its limitations cannot be overlooked. The answer sheets are provided to an examiner for scoring [3]. This process is both time consuming and greatly depends upon the examiner's availability [4]. Errors are likely to occur because different evaluators are employed for checking the answer sheets. Every human rater possesses their own perception for deeply looking into the answer as there are no standardized criteria for marking the answer. The results are then compiled [5].

With advancement in technology, advanced concepts of scoring an answer sheet were introduced in the examination system. The use of computerized tools overcomes the limitations of the manual process. In this system, the students are supposed to submit the answers written on the answer book [6]. The automated examination terminals are meant for transferring student's response to centralized database by electronic means, thereby restricting the physical movement of answer booklets. Intelligent software tools are advantageous in manifold as these are not only speedy but also overcome the human errors of omission and totalling mistakes [7]. The same inference mechanism for checking all the answers ensures the uniformity of marking scheme and speedy declaration of result [8].

Although the above process is partially automated, accuracy can be still enhanced more if the student answer is directly typed to the system and then automated scoring is done, giving score based on the content similarity. If the results produced by the automated scoring system correlate with the scores generated by human graders to a great extent, then this will make the system more consistent than the manual scoring [9]. One relevant solution to overcome the reliability and validity conflicts is to define external criteria against which human- and machine-generated scores can be validated. Another alternative is to define a true score against which these scores can be validated. The superiority of the AES system lies in the fact that it generates the same score for the same essay every time [10]. This employs reliability in test-retest and fairness in evaluation.

2.2. Automated Essay Scoring (AES) for Assisting Expert Human Raters. Expert human raters have to deal with issues like complexity in answers and subjectivity in evaluation. These limitations can be overcome using an automated scoring system. At the same time, effectiveness and convenience are obtained in scoring. Specifically, if automated tools exist, it would be advantageous to assist expert human raters to achieve the following objectives:

(1) To evaluate their own scoring criteria.

(2) To assess deviation from consistencies and undesirable tendencies while scoring.

(3) To study the steps of drawing summary conclusions from response features.

(4) To identify the immediate and evolutionary changes in automated scoring.

(5) To determine the causes of scoring differences between humans and automated grades.

(6) To locate and correct the automated grades for answers that require manual intervention.

Although there exist numerous AES systems, the focus of most studies is on the agreement between automated scores and human-assigned scores on a single essay. Furthermore, the agreement does not tell much about what is measured by automated scores. There is no sufficient evidence for validating AES. Hence, it does not contribute in AES validation construction. Table 1 shows the strength of AES over manual scoring.

2.3. Automated Answer Scoring Methods. There are rules- and statistics-based automated short answer scoring methods which are graphically shown in Figure 1 and explained in the subsequent sections.

2.3.1. Rule-Based Approach. Every student answer has some inherent lexical rules or concepts in their answers. Such rules can be lexically matched, and certain features can be extracted by a few rules-based methods although they cannot be proved statistically. So, surface form of text is used in which the student answer is matched lexically with reference answers. This approach helps to get more accurate score.

2.3.2. Statistical Approach. This approach identifies the probabilities of assigning score values for the given reference answers. The probabilities are calculated to extract features to score the answer. Compared with the fully rule-based mode, the probabilistic and mathematical model produces more accurate score.

The existing automatic essay grading system relies on two aspects, namely, machine learning techniques and grammatical measures of quality techniques. However, none of them identifies meanings (propositions) in the text. Therefore, it proves to be inappropriate for scoring the contents of an answer.

2.4. Automatic Scoring Challenges. The automated scoring has been developed and adopted for English language. There might be few instances where it is used for foreign language but not for the Indian language. It integrates development and demonstration of one of the important systems for

Table 1: Comparison between manual and automated scoring.

Sr. No.	Manual Scoring	Automated Scoring
1	**Measurement weaknesses** Manual scoring has measurement weaknesses: (i) Subjectivity (ii) Lack of reproducibility (iii) Inconsistency errors	**Measurement strengths** Automated scoring is able to achieve: (i) Consistency (ii) Reproducibility (iii) Traceability
2	**Logistical weaknesses** (i) No quick rescoring (ii) Takes more time to score (iii) Not cost effective	**Logistical strengths** (i) Quick rescoring (ii) Time saving and possibility of immediate feedback (iii) Reduced cost
3	**Other weaknesses** It requires: (i) Attention to basic human needs (ii) Recruiting, training, calibration, and monitoring (iii) Intensive direct labour and time	**Other strengths** It requires: (i) No basic human needs once the system is set (ii) No more recruitment, training, calibration, and monitoring (iii) Only one trained operator is sufficient. Negligible labour and time.

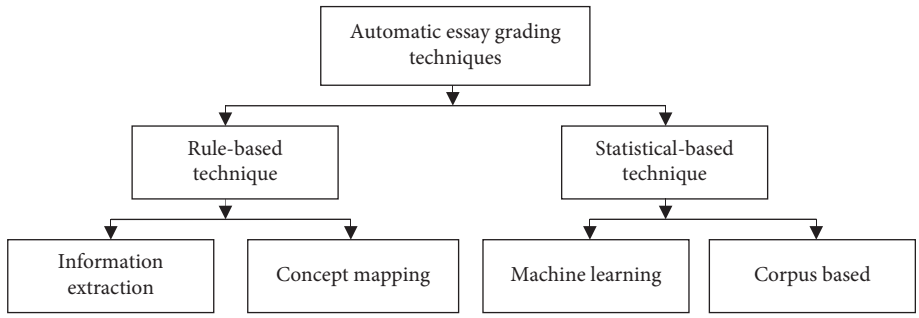

Figure 1: Methods of automatic short answer scoring.

Hindi language in Devanagari script. Few of the challenges related to Hindi language are use of compound or complex sentences and frequent use of polysemous words that are available [11]. Therefore, this system is more suitable with other languages which will broaden the scope of this proposed system.

The main advantages of automated scoring over manual scoring include efficiency and the application of the same evaluation criteria with greater consistency.

3. Existing System

In order to evaluate a number of varying features, there are various AES systems covering various aspects. Currently, there are four major developers of automated essay scoring which are widely used by universities, schools, and testing companies: Project Essay Grader (PEG), Intelligent Essay Assessor (IEA), E-rater, and IntelliMetric [12].

There are many advantages of automated assessment over manual one. These advantages include efficiency, application of the same evaluation criteria with more consistency, etc. Moreover, its ability to provide spontaneous feedback is its primary strength. Automated scoring achieves greater objectivity than manual scoring [13] as computers are not affected by external and emotional factors.

Majority of automated scoring systems generate nearly real-time performance feedback on various aspects of writing. For example: ETS e-rater model provides feedback on grammar, use of words, word mechanics, state, and organization of a written typed text. Similarly, Pearson's IEA covers the different aspects of writing for feedback. The aspects include ideas, organizations, conventions, fluency, and choice of words. This advantage of AES is a limitation of human rating which is not able to provide such analytical feedback for huge quantities of essays. Also, human raters usually need to train several grade ranges linked with a specific rubric and certain tasks. It requires adequate training for shifting to a new grade. Such training is not at all required for AES which is able to evaluate the essays at different grading levels (for example, the e-rater, IEA, and IntelliMetric). Comparison of the AES system is shown in Table 2.

3.1. Predefined Features. In this research, scores are based on extraction of syntactic and semantic features. This research incorporates feature-based grading. Grading is also focused on the similarity among the given answers by the extraction of various features like semantic, syntactic, and lexical.

One of the key techniques for handling and organizing text data is text categorization. It is important because more and more documents are now available in digital form, and at the same time, online information is growing rapidly. It should be noted that the statistical classification methods and the machine learning techniques are also used in text

Table 2: Comparison of existing AES systems.

Existing systems	Approach	Focus
PEG	Statistical	Style
IEA	LSA	Content
E-rater	NLP	Style and content
IntelliMetric	NLP	Style and content

categorization. Since in the proposed system, the domain and the reference answer are fixed, text categorization can be implemented on a continuous basis and efficiently.

Text categorization involves feature extraction which is the most important part of any machine learning task. In this research, to build effective essay scoring algorithm, the aim is to develop model attributes like language fluency, grammatical and syntactic correctness, vocabulary and types of words used, essay length, domain information, and so on. The existing systems follow the following for feature extraction.

3.1.1. Feature Construction. Features are measurable attributes in a text, and they are used as input to the machine learning (automated) software. Feature construction is a process in which possible features are defined.

3.1.2. Feature Ranking. This procedure determines how important each feature is for categorization. The ERT algorithm provided by Scikit-learn [14] is used to generate the feature ranking which means placing feature in order of their importance. These features are considered to determine their ranks in this algorithm.

3.1.3. Feature Selection. In this process, ranked features are used as the input for feature selection algorithm. It is a grid-search process in which respective classifiers are also taken into consideration. The feature with the lowest ranking is eliminated, and the cross-validation error is computed after performing a classification. Ultimately, minimum number of features is reached in this process.

Chen and He[15] defined four different types of predefined features that indicate the essay quality including lexical features, syntactical features, grammar and fluency features, and content and prompt-specific features. These features have been appropriately refined and modified to achieve the objective of the study. The four classes of features used in this system are described below:

(i) Syntactical features

(ii) Lexical features

(iii) Content and prompt-specific features

(iv) Grammar and fluency features

3.2. Related System. The overall objective is to assess the shortcomings of earlier techniques. First, traditional automated systems have been discussed. Thereafter, other approaches have been discussed which are specifically related to the proposed research. The mechanism has been applied to short question answering.

Leacock and Chodorow [16] defined an automated scoring engine called C-rater which was developed to grade answers to content-based short answer questions, and C-rater utilizes morphological analysis, synonyms, and predicate argument structure for assigning full or partial credit to a short answer questions; it cannot be referred merely as a string machine program. C-rater agrees with human raters to a larger extent of 84% of the time.

Song et al. [17] explained the user interactive question answering by applying short-text similarity assessment. The various applications of interactive question answering are IR and text mining like text summarization, text categorization, content-based image retrieval, and machine translation. It should be noted that the short-text question-answers are used.

Kaur and Jyoti [18] explained short one-line free-text answers through automated assessment in the field of computer science. In their research, they have defined a segment of criteria for evaluation, covering all the relevant areas of a short-text evaluation system.

Gomaa and Fahmy [19] compared a different number of corpus-based and string-based similarities in order to explore text similarity approaches for automated short answer scoring in the Arabic language. The comparison between similarity measures reveals immediate feedback to the student. On analysis, resulted correlation and error rate findings proved that this system is useful for its application in a real scoring environment.

Rababah and Al-Taani [20] forwarded a proposal of automated scoring technique for Arabic essay questions in short answers. For this purpose of applying scoring process, we used cosine similarity measure. It was based on the similarity between the student's answer and standard one. The experimental results showed that the competitive scores were achieved when compared to other such approaches.

3.2.1. Vector Space Model (VSM) Approach

Tsatsaronic and Panagiotopoulous [21] discussed a generalized vector space model for text retrieval based on semantic relatedness. The most difficult task is the modification of the standard interpretation of the VSM and others which deals with incorporating the semantic information in a theoretically sound and rigorous manner.

Ekba et al. [22] elaborated plagiarism detection in the text using vector space model. In order to detect external plagiarism, they proposed a technique based on textual similarity. Further it identifies the set of source documents from where the copying of suspicious

document is carried out. This approach was based on the traditional VSM for selection.

Singh and Dwivedi [23] studied vector space model information retrieval for analysis. It is one of the best traditional applied retrieval models for evaluating web page for its relevance. Various approaches of vector space model to compute similarity score of the search engine hits were important.

Jahan and Ragel [24] discussed plagiarism detection on electronic text-based assignments using the vector space model. On analysis, even though trigram utilizes enough time, it is more suitable for detecting plagiarism using cosine similarity measure in all text documents. The vector space model was used in retrieving information using query processing. Cosine similarity measure showing higher results was preferred over Jaccard similarity measure. The future work is to concentrate lesser time for dealing with a large amount of assignments with long length document and detect plagiarism optimally.

Alzahrani et al. [25] developed and compared number of NLP techniques that accomplish the task of automating scoring. They presented the multivector model which is closer to human judgement and gives more accurate and reliable results. They also plan to apply their methodology in different languages.

Lilleberg et al. [26] performed demonstration for classification of text with semantic features on the support vector machines and word2vec. They assumed that word2vec brings extra semantic features helping further in text classification. Based on this, effectiveness of word2vec was demonstrated by showing that TF-IDF and word2vec combination can outperform TF-IDF. Their approach was incomplete as it only scratches the surface; ideal results can still be expected. Recommendations for a future work depend on the ways to bring much improvement in consistency which can be achieved in many ways such as modification of stopword list or changing the weights.

3.2.2. Other Related Concepts

Keller [27] conducted a comparative study of the generalizability of scores produced by automated scoring systems and expert graders. In addition to the available information, their paper description is based on the performance of AES systems through various reports collected from expert raters and computer-produced scores. After analysis, performance was checked for physician's patient management skills through computer-delivered assessment. Final results exhibit a relatively positive outcome regarding performance of the regression-based scoring algorithm.

Hajeer[28] conducted a study on various statistical similarity measures for their effectiveness. The use of different statistical measures in information retrieval (IR) is very effective for document retrieval using a unified set of documents. Two issues were addressed: firstly, to study the different statistical measures for its effectiveness on a unified set of documents and secondly, to find the most appropriate one to classify documents through comparing them in an orderly manner. After analysis, it was concluded that the cosine similarity measure is the best for the document retrieval technique. In future work, he hopes to extend this project to test other measures.

Weigle [29] presented numerous considerations which are critical for English language learners and automated scoring of essays. His study projected various considerations to use automated scoring systems in evaluating second language writing. There were other aspects like challenges and opportunities which were listed in this presentation. His article analyses the extent to which system developers can assess the particular needs of learners in English language. It concludes that the greater the evaluators and authorities possess knowledge regarding automated scoring system, the more will be the chance of this technology to be used widely to meet the ever-growing demands of huge population.

Paskaleva et al. [30] developed a new set of similarity functions for information retrieval. Records were considered as multisets of tokens which map records into real vectors. In their research, for bridging the gap between set-based models and vector space model, consistent extensions of set-based similarity functions were developed.

McNamara et al. [31] explained in their study the significance of approach based on hierarchical classification approaches which are meant for computing essay scores involving a set of text variables. On analysis, 55% exact accuracy between predicted essay scores and the human scores is revealed along with 92% adjacent accuracy. Although features which inform the overall assessment will differentiate depending on the specific problem, this approach is able to get performance models with high accuracy and information in comparison to simple one-shot regression.

Sultan et al. [32] discussed student's short answer question which is given with the correct answer; the principle of grading student response is derived from its semantic similarity with the correct answer. Key measure employed in their supervised model utilizes the recent approach of identifying the short-text similarity features. In addition, the term weighting mechanisms are needed to identify important answer words in many cases. Accuracy for answer scoring can be achieved by evaluating a simple base model that can be easily extended with new features.

Wang et al. [33] conducted a study on identifying current issues in short answer grading (SAG). In order to observe the issues involved in SAG, they analyzed the results of a simple SAG approach. They used KNN to score query answers, where vector representations of answers are generated from weighted, pretrained word

embedding. By analyzing the errors in the given approach, it was shown how the diversity and short length of answers caused problems to SAG. Properties of short answer scoring such as diversity of answers were statistically analyzed.

Raczynski and Cohen [34] in their research article "Appraising the scoring performance of automated essay scoring systems—some additional considerations," they provided useful validation framework for assessment of the automated scoring system. They determined the type of essays which can be used to calibrate and test automated essay scoring (AES) systems. They also discussed what human scores should be used when there are scoring disagreements among multiple human raters.

Wang and Brown [35] discussed validation on manual and automated scoring of essays against "true" scores. Raters were divided into two groups (14 or 15 raters per group), and they rated 250 essays in two sets which were all written in response to the same prompt, thereby providing an approximate true score to the essay. Training on the datasets was provided to an automated essay scoring (AES) system in order to score the essays using a cross-validation scheme. We concluded that the correlation between automated and human scores is of the same order as the correlation between manual graders.

4. Proposed Methodology

Undoubtedly, this system is based on the vector space model, but it is incorporated by further changes for gaining better results:

(1) Our vector incorporates syntactical features and semantical features. It shows how the document is vectorized. Two arrays are there for each document as shown in Figure 2. First column referred to as predefined feature is inserted with term (T1,...,Tn). Second column is inserted with the weight (W1,. . .,Wn) with respect to the term feature.

Whenever any new document is added, the columns are incremented in the matrix and the number of rows is incremented when new term is to be added.

(2) Generally, cosine similarity is used to find similarity among vectors. Well-defined new similarity measures are proposed for the scoring of the answers which includes syntactic and semantic features. This technique definitely will produce better result than cosine similarity. Equation (1) represents similarity.

$$\textbf{Similarity} = \sum_{\mathbf{i}=1}^{\mathbf{n}} \mathbf{f_i}. \tag{1}$$

(3) Term weighting is the key in the vector space method. In addition, several researches on term weighting techniques have been conducted. There is still a conflict regarding which method is more appropriate.

D_1	
T_1	W_1
T_2	W_2
T_3	W_3
:	:
T_n	W_n

Figure 2: Vector representation.

4.1. Term Weighting. The advanced text retrieval systems view term weighing as an important component. The major content of the text or literature is well defined in terms of words, phrases, or other units of indexing. Each and every word of the text has its own importance and worth. This phenomenon is indicated as term weighing represented by the following equation:

$$\mathrm{Idf}_t = \log\left(\frac{N}{\mathrm{df}_t}\right), \tag{2}$$

where N is the total set of documents and df_t is the document frequency.

4.2. TF-IDF Weighting. It is now attained by combination of term frequency and inverse document frequency, and also it produces a combined weight of every term in each of the document. This scheme is represented mathematically as TF-IDF, and this assignment of weight to terms t in the document is the basis proposed system. Equation (3) illustrates TF-IDF formula.

$$\mathbf{Tf - idf_{t,d}} = \mathbf{tf_{t,d}} * \mathbf{idf_t}. \tag{3}$$

The Hindi literary document is such document with focus on information retrieval using proposed theory. This document proves to be much easier and interactive for Hindi literates and all students as they get the pictures along with the rhymes. It also learned and remembered content for beginners. Performance tuning is another important feature of this system which supports around 41 categories. Additionally, more documents can further be added which would be useful concept in future.

(4) In the present study, the performance analysis is calculated using Pearson's correlation coefficient between human scores, which is computed using the following equation:

$$\mathbf{Corr}\ (\mathbf{X}, \mathbf{Y}) = \frac{\sum_{i=1}^{n}\left[\left(x_i - \overline{x}\right)\left(y_i - \overline{y}\right)\right]}{\sqrt{\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2}\sqrt{\sum_{i=1}^{n}\left(y_i - \overline{y}\right)^2}}. \tag{4}$$

It is more reliable and accurate as the performance analysis is correlated with human graders.

This study combines both approaches, that is, rule- and statistical-based methods, for obtaining the scores for the given answer. For making a vector, it uses NLP tools such as morphological analyzer and POS tagger. The superiority and better version of this system is visualized through synergistic effect and the accuracy of the output.

5. Objective Achieved

This research develops an automatic answer scoring system suitable for Indian languages specifically Hindi in Gurmukhi script. The subobjectives to achieve this aim are as follows.

5.1. Development of Resources. Resources developed and used to accomplish this automated scoring were

Predefined question: predefined questions developed by teachers are fed to the system. The system is flexible in the sense that questions can be deleted, modified, or added at any time.

Reference answers: standard reference answers defined by expert human raters are fed in the system. There can be more than one reference answers for a given question.

Corpus: a fixed corpus is developed and selected for a domain so that teachers can set or select fixed questions. This helps in easy extraction of information and in increasing the accuracy of the system.

Lexicons: words and their synonyms relevant to the domain are collected and their contextual meaning are defined. This dictionary is then added to the database. Unlike English, lexical material is not easily available in Hindi. So, a special dictionary with synonyms has been prepared.

Other standard NLP tools: NLP tools like morphological analyzer, stemmers, and part-of-speech tagger have been incorporated and successfully integrated in the system.

5.2. Identification of Predefined Features. To define and maintain quality of an essay, certain predefined features are first identified and then extracted for evaluating them with respect to the reference answers. So, high importance is given to predefined features of both reference answers and student answers.

The process of feature selection is used to determine and limit the dimensionality of features. Instances with higher probabilities are selected which have feature relevance. This helps in improving the performance of feature selection. A wide range of algorithms is used for text clustering in feature selection. A distance measure is selected in clustering which determines the similarities of two answers. Cosine similarity, Euclidean distance, Jaccard coefficient, and Pearson correlation coefficient concepts are some of the similarity or distance measures which have been used and widely applied in the study. The identification of predefined features involves the following three steps.

5.2.1. Preprocessing. Text preprocessing is used to transform the whole text into a viable form for learning algorithms. It involves tasks like treatment and refining of data. Preprocessing includes:

(1) Converting byte strings to tokens which can be called lexical analysis.

(2) Eliminating stopwords like the, and, of, and a.

(3) Changing different word forms of a word to a single "stem" form like ing, ed, pre, and sub.

(4) Selection of terms (feature) which can be individual words or noun phrases.

5.2.2. Extraction. In extraction, NLP (natural language processing) tools are used to extract feature terms. The process is also applied to feature reduction phase of the text classification process. Linguistic features are extracted from text and used as a part of their feature vectors. One of the methods for extracting features is the part-of-speech (POS) tagging. The document is tagged through the standard n-gram tagger. Besides the above NLP tools, there are other tools like morphological analyzer and spell checker, which can be used for feature extraction.

5.2.3. Feature Selection. Feature selection is performed after extracting features. Thereafter, standard predefined features are selected.

5.3. Developing Statistical Model. Statistical model has been developed and used, and it has the following advantages:

(1) It characterizes numerical data, describes measurements, and helps in the development of conceptual models of a system or process.

(2) It helps to estimate the uncertainties in observational data and its calculation.

(3) It characterizes numerical output from mathematical models. The information gathered from the model can be fed back to the system to enhance its performance.

(4) Input parameters can be estimated if more complex mathematical models are encountered.

VSM is essentially a statistical model. So, considering the above advantages and strengths, it has been adopted in present study. The required changes have been applied to remove the limitations in the proposed system.

5.4. Scoring Mechanism. It has been already discussed that the scoring mechanism is implemented by extracting grammatical and semantic relations between the student answer and the reference answers. The internal scoring

mechanism has been explained in the next chapter "Experimentation and Evaluation."

Semantic similarity is a metric which is used to determine the degree of distance between a set of documents or terms. It is based on similarity of their meaning or semantic content just like syntactical similarities. The comparative analysis of the different meanings that allows us to obtain the numerical description. Semantic similarity must be distinguished from semantic relatedness. Any relationships between two terms constitute semantic relatedness. On the other hand, semantic similarity is based on "is a" relation. For example, "car" is similar to "bus" but is also related to "road" and "driving."

So, the proposed statistical model greatly reduces the complexity of semantic relatedness as it is very hard to extract. However, these two terms are used interchangeably in much of the literature. It is true that basically these three terms, namely, semantic similarity, semantic relatedness, and semantic distance, mean "How much does term A have to do with term B?" The answer is usually a number between 0 and 1 or −1 and 1. If it is 1, it means very high similarity. Semantic similarity is a hot issue in NLP. Natural language processing **(NLP)** is a field of computer science and linguistics in which semantic similarity between concepts is a parameter to measure the semantic similarities or distances among the given answers. In other terms, semantic similarity is used to identify concepts that have common "features."

6. Result and Discussion

The usage of cosine similarity function between two vectors is the novelty in scoring; It is both its own scoring function as well as its similarity formula. For better evaluation, the proposed model is classified into three weighting intervals for effective evaluation. The maximum value belongs to one of the following intervals: (i) [33%, 50%] (ii) [50%, 75%] (iii) [75%, 100%].

Word_Count_feature: to give complete demonstration, suppose the maximum grade of the word count is 0.5, which is obtained through the filling of questionnaire given by 50 expert human scorers. For better grading, it is further divided into three parts to get a score (point) of 0.5 of word count feature.

After employing the process of calculating the range of feature, next step is to give scoring by finding similarities among the total word count of the reference answer vector with that of total word count of the student answer vector. Target grade scores are awarded after matching and are calculated by

```
if ((Word_Count ≥ x) && (Word_Count ≤ y)) then
        word_count = 0.35;
elseif ((Word_Count ≥ y) && (Word_Count ≤ z)) then
        word_count = 0.45;
elseif (Word_Count ≥ z) then
        word_count = 0.5;
else
        word_count = 0;
End if
```

where x, y, and z represent total word count (TWC) percentage of reference answers (Ri). It will be computed as follows.

$$x = \text{TWC (Ri)} * 0.33$$
$$y = \text{TWC (Ri)} * 0.50$$
$$z = \text{TWC (Ri)} * 0.75$$

Feature-based scoring is incorporated in this research. The student answer and the reference answer similarities are calculated to evaluate the correct scores. It also extracts the different lexical, syntactical, and semantic features for giving the scores. Each feature has its own weightage towards the target grade. The implication is that if there is any absence of semantic feature, then it will affect the target grade more than the syntactic features. Other features have less weightage than semantic features because semantic features depend on the content similarities as shown in Figure 3.

6.1. Performance Analysis. The performance of the system is evaluated by comparing the output generated by the system with the result given by human raters. The correlation coefficient is calculated, and it proves that system score and human raters' score are highly correlated with each other, i.e., near to one (positive correlation). The value between human raters is quite close to the score agreement value achieved between a human rater and the system. Table 3 shows the final score of **100 students** graded by the system.

6.2. Cosine Similarity Comparison. In Table 4 and Figure 4, the observation shows that angle of cosine is not near to one. In this research, the reference answers and student answer similarities are not far from one which shows the efficiency of the proposed system. The cosine similarity function mentioned above is computed as shown in Table 3.

7. Conclusion, Recommendations, and Future Work

This paper contains a summary of the whole work, few recommendations, and future scope to extend the reported work. This work contributes to the active research area of automated answer scoring system. The focus of this work is to analyze the existing automated scoring systems along with its workflow system. The research objectives have been satisfied by developing reliable automated answer scoring system for Indian language.

7.1. Conclusion. Various approaches reported by researchers have been reviewed which show the feasibility of automated short answer system. Evaluating large number of students' answer in a given period of time with feedback is a trivial task. Manual grading proved to be a weakness with respect to resource requirement, fairness, cost, and timely feedback challenges. On the other hand, the automated system is

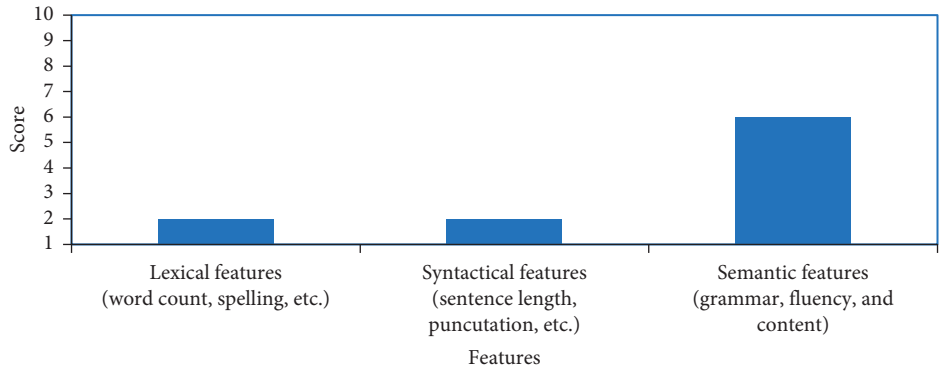

Figure 3: Feature-based analysis.

Table 3: Final score of 100 students.

No. of student	System grade	Human grader 1	Human grader 2	Human grader 3	Human grader 4	Human grader 5
1	8.8	9.1	7	8.6	7	8
2	7.3	6	7	6	7.6	7
3	9	10	9	8.5	8	7.5
4	5.8	8.5	6.1	6	6.2	7.5
5	1	1	2	0	1	2
6	7	7	6.5	8	6.5	7.5
7	5	4	4	3	2	5
8	9.5	9.5	10	9	8.5	8
9	9	7.5	8.5	8	8	9
10	2.5	1.5	0	1.5	1	2
11	8.5	8.5	7	8	8	9
⋮						
99	0	0.5	1	0	2.5	1.5
100	8	8.5	6.5	8.5	7	8.5
Correlation coefficient (r)		**0.91**	**0.90**	**0.88**	**0.85**	**0.95**

Table 4: Comparison of cosine similarity.

Student	Cosine similarity	Proposed system score
1	0.81	0.91
2	0.7	0.9
3	0.50	0.92
4	0.7	0.74
5	0.5	0.8
6	0.7	0.8
7	0.62	0.89
8	0.53	0.92
9	0.63	0.85
10	0.8	0.91

much helpful for providing grade as well as feedback of student's answer within the specified time frame. In order to ensure consistency and to overcome the problems of manual scoring, the automated system gives correct scores. These scores can be repeated several times with consistency at different times. Recognizing the importance and fairness of automated scoring system, there is need to refine and develop the system further for more accuracy. Other issues like computation and linguistic features are required to be handled for more effectiveness of the system. A number of AES are available for evaluation of different features. It still needs to be developed and refined so as to overcome the shortcomings.

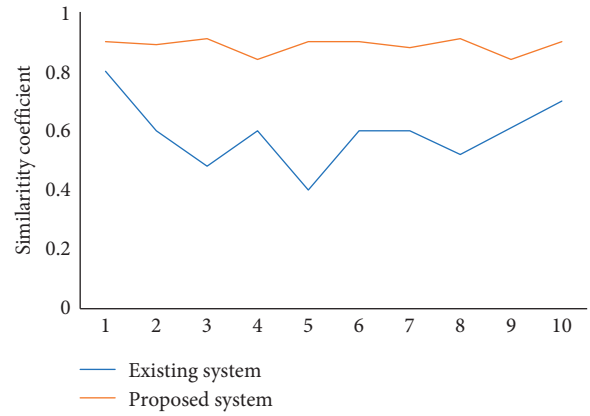

Figure 4: Comparison of the proposed system with the existing system.

7.2. Recommendations. Based on the computer-based automated scoring system, the following recommendations are suggested for reliable scoring. The validation strategy for AES in this system takes into consideration the following:

(1) The statistical relation between scores assigned by the AES system.

(1) Other aspects of AES like selection features and their weightage are also considered.

(2) It considered the agreement between independent measurements of students' writing skills and the scores assigned by an AES system.

7.3. Future Work. The work undertaken in this research can be expanded in diverse directions. The present system achieves its target of objectivity and fairness of evaluation in scoring. The achievements have already been mentioned. However, one may say that the system is rigid and lacks flexibility from the human perspective. But this charge, if at all it is to be levelled, can easily be addressed in the future development of the system. For example, the system can be modified so that the scoring can be graded or classified as strict, moderate, standard, etc., according to the requirements of the situation or teachers. The idea can surely be developed in the automated scoring system.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

Dr. Omar Cheikhrouhou thanks Taif University for its support under the project Taif University Researchers Supporting Project (no. TURSP-2020/55), Taif University, Taif, Saudi Arabia.

References

[1] M. Albared, N. Omar, and M. J. Ab Aziz, "Classifiers combination to Arabic morphosyntactic disambiguation," in *Proceedings of the 2009 International Conference on Electrical Engineering and Informatics*, pp. 163–171, Selangor, Malaysia, August 2009.

[2] L. Jovita, A. Linda, A. Hartawan, and D. Suhartono, "Using vector space model in question answering system," *Procedia Computer Science*, vol. 59, pp. 305–311, 2015.

[3] R. Wang, H. Yu, G. Wang, G. Zhang, and W. Wang, "Study on the dynamic and static characteristics of gas static thrust bearing with micro-hole restrictors," *International Journal of Hydromechatronics*, vol. 2, no. 3, pp. 189–202, 2019.

[4] M. Kaur, D. Singh, and V. Kumar, "Color image encryption using minimax differential evolution-based 7D hyper-chaotic map," *Applied Physics B*, vol. 126, no. 9, pp. 1–19, 2020.

[5] T. Wiens, "Engine speed reduction for hydraulic machinery using predictive algorithms," *International Journal of Hydromechatronics*, vol. 2, no. 1, pp. 16–31, 2019.

[6] S. Osterland and J. Weber, "Analytical analysis of single-stage pressure relief valves," *International Journal of Hydromechatronics*, vol. 2, no. 1, pp. 32–53, 2019.

[7] H. S. Basavegowda and G. Dagnew, "Deep learning approach for microarray cancer data classification," *CAAI Transactions on Intelligence Technology*, vol. 5, no. 1, pp. 22–33, 2020.

[8] S. Ghosh, P. Shivakumara, P. Roy, U. Pal, and T. Lu, "Graphology based handwritten character analysis for human behaviour identification," *CAAI Transactions on Intelligence Technology*, vol. 5, no. 1, pp. 55–65, 2020.

[9] B. Gupta, M. Tiwari, and S. Singh Lamba, "Visibility improvement and mass segmentation of mammogram images using quantile separated histogram equalisation with local contrast enhancement," *CAAI Transactions on Intelligence Technology*, vol. 4, no. 2, pp. 73–79, 2019.

[10] M. Kaur, D. Singh, and R. Singh Uppal, "Parallel strength pareto evolutionary algorithm-II based image encryption," *IET Image Processing*, vol. 14, no. 6, pp. 1015–1026, 2019.

[11] N. K. Matharu, V. Dasari, and R. K. Mishra, "Homeotic gene regulation: a paradigm for epigenetic mechanisms underlying organismal development," in *Epigenetics: Development and Disease. Subcellular Biochemistry*, T. Kundu, Ed., Springer, Dordrecht, Netherlands, 2013.

[12] S. Dikli, "An overview of automated scoring of essays," *The Journal of Technology Learning and Assessment*, vol. 5, no. 1, 2006, https://ejournals.bc.edu/index.php/jtla/article/view/1640.

[13] D. M. Williamson, R. J. Mislevy, and I. I. Bejar, *Automated Scoring of Complex Tasks in Computer-Based Testing*, Lawrence Erlbau, Hillsdale, NJ, USA, 2006.

[14] F. Pedregosa, G. Varoquaux, A. Gramfort et al., "Scikit-learn: machine learning in python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[15] H. Chen and B. He, "Automated essay scoring by maximizing human-machine agreement," in *Proceedings of the 2013 Conference on Empirical Methods in Natural Language*, Seattle, NJ, USA, October 2013.

[16] C. Leacock and M. Chodorow, "C-rater: automated scoring of short-answer questions," *Computers and the Humanities*, vol. 37, no. 4, pp. 389–405, 2003.

[17] W. Song, D. Ren, W. Li et al., "SH2B regulation of growth, metabolism, and longevity in both insects and mammals," *Cell Metabolism*, vol. 11, no. 5, pp. 427–437, 2010.

[18] N. Kaur and K. Jyoti, "Automated assessment of short one-line free-text responses in computer science," *International Journal of Computer Science and Informatics (IJSCI)*, vol. 2, no. 1, p. 2, 2012.

[19] W. H. Gomaa and A. A. Fahmy, "Short answer grading using string similarity and corpus-based similarity," *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 3, no. 11, pp. 1–7, 2012.

[20] H. Rababah and A. T. Al-Taani, "An automated scoring approach for Arabic short answers essay questions," in *Proceedings of the IEEE 8th International Conference on Information Technology (ICIT)*, pp. 697–702, Amman, Jordan, May 2017.

[21] G. Tsatsaronic and V. Panagiotopoulous, "A generalized vector space model for text retrieval based on semantic relatedness,," in *Proceedings of the EACI 2009 Student Research Workshop*, pp. 70–78, Athens, Greece, April 2009.

[22] A. Ekba, S. Saha, and G. Choudhary, "Plagiarism detection in text using vector space model," in *Proceedings of the 2012 12th International Conference on Hybrid Intelligent Systems (HIS)*, pp. 366–371, Pune, India, December 2012.

[23] J. N. Singh and S. K. Dwivedi, "Analysis of vector space model information retrieval," in *Proceedings of the National Conference on Communication Technologies & its impact on Next Generation Computing 2012 CTNGC*, pp. 14–18, Ghaziabad, India, November 2012.

[24] M. J. M. A. J. Jahan and R. G. Ragel, "Plagiarism detection on electronic text based assignments using vector space model," in *Proceedings of the 7th International Conference on Information and Automation for Sustainability*, Colombo, Sri Lanka, July 2014.

[25] A. Alzahrani, A. Alzahrani, F. K. AlArfaj, K. Almohammadi, and M. Alrashid, "AutoScor: an automated system for essay questions scoring," *International Journal of Humanities Social Sciences and Education (IJHSSE)*, vol. 2, no. 5, pp. 182–187, 2015.

[26] J. Lilleberg, Y. Zhu, and Y. Zhang, "Support vector machines and Word2vec for text classification with semantic features," in *Proceedings of the 2015 IEEE 14th International Conference on Cognitive Informatics & Cognitive Computing*, Beijing, China, July 2015.

[27] L. A. Keller, B. E. Clauser, and D. B. Swanson, "Using multivariate generalizability theory to assess the effect of content stratification on the reliability of a performance assessment," *Advances in Health Sciences Education*, vol. 15, no. 5, pp. 717–733, 2010.

[28] I. Hajeer, "Comparison on the effectiveness of different statistical similarity measures," *International Journal of Computer Applications*, vol. 53, no. 8, 2011.

[29] S. C. Weigle, "English as a second language writing and automated essay evaluation," *Handbook of Automated Essay Evaluation: Current Applications and New Directions*, Routledge, New York, NY, USA, 2013.

[30] B. S. Paskaleva, S. E. Godoy, W.-Y. Woo-Yong Jang, S. Bender, M. M. Krishna, and M. M. Hayat, "Model-based edge detector for spectral imagery using sparse spatiospectral masks," *IEEE Transactions on Image Processing*, vol. 23, no. 5, pp. 2315–2327, 2014.

[31] D. S. McNamara, S. A. Crossley, R. D. Roscoe, L. K. Allen, and J. Dai, "A hierarchical classification approach to automated essay scoring," *Assessing Writing*, vol. 23, pp. 35–59, 2015.

[32] Md A. Sultan, C. Salazar, and T. Sumner, "Fast and easy short answer grading with high accuracy," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, San Diego, California, June 2016.

[33] T. Wang, Z. Jiang, T. An, G. Li, H. Zhao, and P. K. Wong, "Enhanced visible-light-driven photocatalytic bacterial inactivation by ultrathin carbon-coated magnetic cobalt ferrite nanoparticles," *Environmental Science & Technology*, vol. 52, no. 8, pp. 4774–4784, 2018.

[34] K. Raczynski and A. Cohen, "Appraising the scoring performance of automated essay scoring systems-Some additional considerations: which essays? Which human raters? Which scores?" *Applied Measurement in Education*, vol. 31, no. 3, pp. 233–240, 2018.

[35] J. Wang and M. S. Brown, "Automated essay scoring versus human scoring: a correlational study," *Contemporary Issues in Technology and Teacher Education*, vol. 8, no. 4, 2008, https://citejournal.org/volume-8/issue-4-08/english-language-arts/automated-essay-scoring-versus-human-scoring-a-correlational-study.