Journal of Scientific & Industrial Research
Vol. 81, January 2022, pp. 76-83

Evaluation of Descriptive Exam Answer Scripts using Word Mover's Distance

M Ramakrisna Murty[1]*, B Tarakeswara Rao[2], Y Anuradha[3] and Hyma J[4]

[1]ANITS, Visakhapatnam, AP, India; [2]KHIT, Guntur, A P; [3]GVP C E (A), Visakhapatnam; [4]GITAM, Visakhapatnam

Received 11 February 2021; revised 22 December 2021; accepted 29 December 2021

The knowledge and competency assessment have paramount significance in the education system. Recent scenario of COVID-19 witnessed the need of migrating from traditional education system to a modern online learning environment. Currently in the online assessment process, descriptive exam answer scripts evaluation is one of the tedious tasks to the teachers. The knowledge assessment may sometimes lead to biasing based on the mood of the evaluator and other circumstancing parameters. In general, though the evaluation process is well defined, still when two evaluators evaluate the same scripts, there are very less chances to award the same marks. The proposed model aims to address such real time issues and outer performs of the evaluation of descriptive answer scripts by using text semantic similarity measure. The proposed model works based on the word mover's distance, whose purpose is to measure the semantic similarity among the actual answer and the answer given by the students. In this work, the data set is generated from the descriptive on-line examination platform. The data set contains student's answers, which can pre-process initially and measure the semantic similarity among key answer and student's answers. The given automatic evaluation procedure, could guarantee the impartiality and concealment of the evaluation.

Keywords: Machine learning, Semantic similarity, Skip gram model, Text mining

Introduction

Student assessment on descriptive type of answers is the most complicated and time-consuming task in the education system. Descriptive answer assessment methodologies have been under research since long back and as a result, various algorithms have subsequently been proposed and implemented. Most of the exiting methods are focused on syntax, vocabulary, size of the content but not on the semantic meaning of the text given by the student. This work aimed to propose a machine-learning model to analyse text semantic based evaluation procedure for the descriptive answers in the examinations. The existing few semantic models namely Latent Dirichlet Allocation (LDA), Latent Semantic Analysis(LSA), Content Vector Analysis partially addresses the semantic issues in the text analysis. Latent Semantic Analysis[1], Code Explanation (CE) is few of the techniques, which are widely experimented before the concept of word embedding.[2]

Bag of Words [3] and Term Frequency Inverse Document Frequency (TF-IDF) are the basic underlying principles for the techniques like LSA and CE using which, frequency count of the words is computed and thus, their representation is obtained.[4] The shortcoming of these methods is that the order of words and context of words is ignored. For example, "A Good nutrition is important for leading healthy life style" and "Healthy weight reduces the risk of chronic diseases" will be treated orthogonal and makes them independent though they intend same meaning and same context.[5]

Learning of the new things is facilitated by a proper data interpretation by the human beings. The huge data availability demands the same sort of data interpretation by the machine for better data utilization, understanding and learning of many things. Machine Learning is one of the challenging areas that have the ability to automatically learn and focus various experiences by extracting the data. Most of the available real-world data will be in textual form. However, the algorithms available for working over this data cannot be applied directly over the text instead; they need a proper representation of the words in the data.

Word representation in the data is the foremost and essential task for language processing by the machines. Word embedding is the term coined for the method that is used to represent the word or text to discrete numbers or symbols. The numerical representation of

*Author for Correspondence
E-mail: ramakrishna.malla@gmail.com

the data now can be easily understandable by the machine learning algorithms. Document clustering, Feature generation, Text classification and Natural language processing are various applications of this word embedding. Word2Vec is one of the noticeable better and efficient products developed by Google.[6] In this semantic analysis model, the words in a text are represented in vector space and are placed in such a way that similar words appear closer and dissimilar words are located far way. This semantic relationship among words facilitates the learning algorithms to work in a more efficient way.[7]

The main objective of this work is to come forward with a model to automatically evaluate descriptive answer scripts of the students by using text mining and machine learning methods. In this proposed model, word mover's distance (WMD) is used for word embedding and vector representation of words. This semantic similarity measure finds the distance between individual words from the schema provided by the examiner and students answer. Word Embedding plays a vital role to reach the aimed task with its syntactic and semantic relationship among words in the answer scripts.

Literature Study

Similarity measurement at various levels like word level, sentence level, paragraphs and documents level is a crucial task for various major applications like information retrieval, medical data analysis, news analysis etc. The descriptive work by Yuejin & Reynolds[8] has given a detailed survey report over various text similarity approaches namely String-based approach, Corpus-based approach and Knowledge-based similarities. Their detailed report elevated various algorithms in each of the category.

The recent Earth Mover Distance (EMD) is an efficient metric for comparing discrete probability distributions. However, it has the limitation of high computational cost. At the other end, linear complexity approximation algorithms with improved scalability are limited to both low dimensional vector space and become inefficient when there is high overlap between the probability distributions. The work proposed by Wan & Angryk[9] comes up with a new novel approximation algorithm that overcomes above disadvantages. Their practical results have shown the improved efficiency and accuracy of EMD in both high and low dimensions.

Identifying the parallel sentences required for statistical machine translation and their process of application to bilingual word embeddings addressed in[10] their work succeeded in attaining minimum distance while travelling through various translation paths in one-to-one variation.

Another work proposed in attempted to use a Relation Network (RN) module along with a Long Short Term Memory (LSTM) for performing string comparison needed for checking redundant data in sentence paragraphs.[11,12] Their work mainly concentrated on detection of duplication and paragraph detection. The approximation of Earth Mover Distance is performed by calculating word importance and their flow optimization is done using LSTM and RN modules.

Another work proposed in developed a system that could intelligently identify revision relationships with in a collection of text documents.[13,14] This revision detection mainly relies on comparing the two documents and assessing their similarity score. This work explored two new document distance measures and shown how they are able to capture the semantics of the words. Their work mainly attempted to retrieve an optimal revision sub network with the findings of minimum branching. Supervised word movers distance is proposed by Gao *et al*. where author improve the performance of the word movers distance measure.[15] Topic modeling on user stories using word mover's distance by Gülle *et al*.[16] and "Word Mover's distance for agglomerative short text clustering" by Franciscus *et al*.[17] the author used word mover's distance to short text clustering efficiently. The document classification carried out by Wu and Li, with modified word mover's distance works better than other existing methods.[18]

Proposed Model

In the proposed model both examiner answer/key and student answer will go through stop word elimination and stemming process. The examiner answer or key can be stored as repository file for the purpose of automatic evaluation. The proposed model is shown in the Fig. 1. In the proposed model, word mover's distance is chosen as semantic similarity measure and its in-depth details are presented in following sub sections.

Data Preprocessing

Data preprocessing is the vital task in this work to handle raw data. In this model generated answers scripts, which contains different formats of texts. Here definitely need preprocessing of the data set.

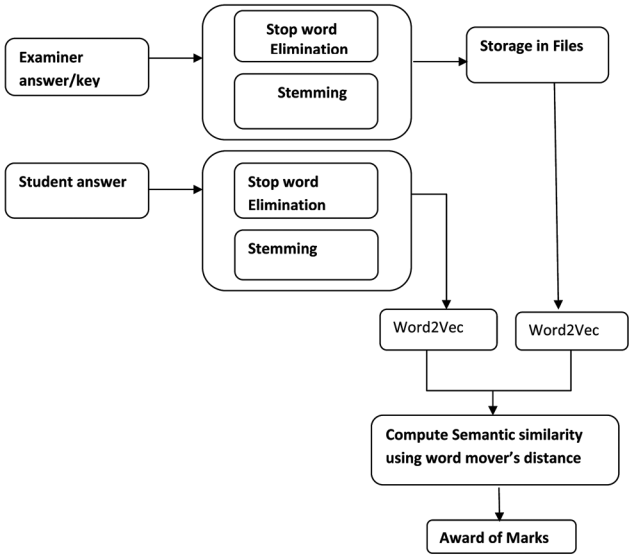

Fig. 1 — Proposed model for evaluation of student descriptive using word mover's distance

Word frequency and inverse-term frequency are the widely used ways to find the similarity among pieces of texts, but this method fails to address the semantic similarity among the text documents. The proposed model works with the help of the word mover's distance. In this model, there are two inputs; one input is examiner answer or key of the question paper. The second input is student answer, which need to be assessed to award marks. The examiner input or key should be stored in a file, after eliminating stop words and stemming process.[19] The second input, which is otherwise called student answer, also will go through the same process of stop words elimination and stemming.[20] The stop word elimination and stemming are the pre-process step, during which common words in a language like "a" "as" "to"," the" etc are eliminated. These words do not imply significant meaning in the sentences and are usually removed from the document. After stop word elimination, it will go for the stemming process. Stemming is the process to find the root word of common words. For example 'running', 'runs', 'ran' are the words have the root word as 'run'. Stemming procedure eliminates related words and find only root word of the common words.

The semantic similarity finding rather than just going with frequency of words is of immense need in subjective paper evaluation. In this context, word mover's distance (WMD) is aimed to compare the student answer with key or examiner answer semantically. In this work used Database Management Systems(DBMS), Operating Systems(OS), Software Engineering(SE), Fundamentals of Computer Science(FCS), Computer Networks(CN)

Word Mover's Distance

Words Mover's Distance (WMD) was introduced in the year 2015 by Kusner and his fellow researchers.[21] The advantage of this technique is, it leverages word embedding's by targeting at both syntactic and semantic similarity distances among text documents. Earthmover distance [21] is the motivation to this Word mover's distance measure and it works better by overcoming the limitations of other basic distance measure like Euclidean distance, Cosine etc., with bag word model. WMD uses word embedding to estimate the distance so that, it can find even when there were no common words in the documents. There are few more word embedding models in the literature namely GloVe, Word2Vec etc.[22] This work mainly uses Word2Vec model to calculate the semantic similarity between examiner answer and key and students answer.

In the word mover's distance, the comparison among two documents and the distance among them is evaluated by using the mathematical equations given in 1,2 and 3.This denotes how much of word i in the examiner answer/key document (denoted by d) travels to word j in the students' answer (denoted by d').

Then the problem becomes the minimization of the document distance, or the WMD, and is formulated as: $\min_{T\geq 0} \sum_{i,j=1}^{n} T_{ij} c(i,j)$ … (1)

Given the constraints: $\sum_{j=1}^{n} T_{ij} = d_i$ … (2)

and

$\sum_{i=1}^{n} T_{ij} = d'_j$ … (3)

This is really a simplified case of the Earth Mover's distance (EMD), or the Wasserstein distance.

Word2Vec Embedding

Word embedding is the process of generating word vectors of the documents. The reason to generate word vectors is that the computer can understand only numeric values and algorithmic approaches like machine learning can do linear algebra operations on numbers in place of words. This work mainly used Word2Vec model to calculate the semantic similarity among student examiner answer/key and students answer with Word2Vec word embedding procedure.[23]

Word2Vec is one of the biggest developments in the field of text mining and natural language processing research area. This idea is simple, but makes good

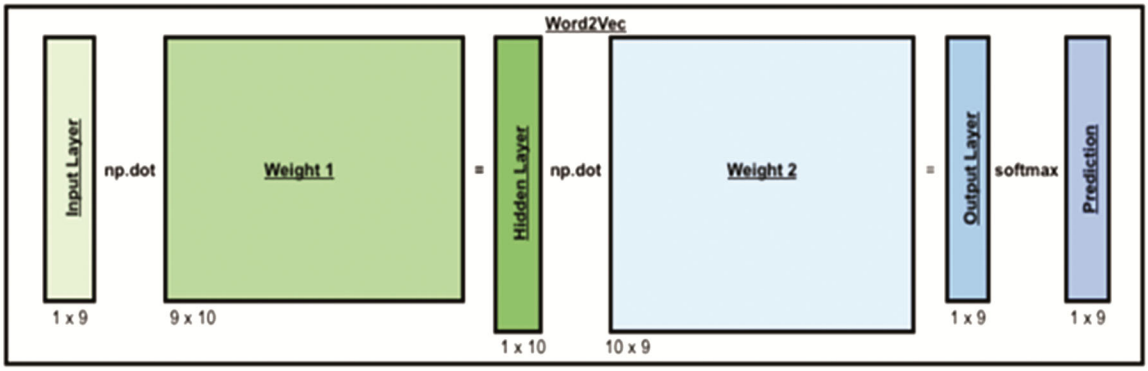

Fig. 2 — Word embedding training model using word2vec

impact on word processing and word vector generation.[24] Word2Vec purpose is to generate vector representation of words. Generally, every word vector has several dimensions and a vector is assigned to each unique word in the corpus. There are two kinds of Word2Vec implementation procedure namely Continuous Bag of Words - CBoW) and Skip-gram - SGM model. In the continuous bag-of-words it aims to reach target word from its neighbors, whereas skip-gram model looks for context words from target word. In this work, skip-gram model is implemented to guess the context word from the target word. Generally, Word2Vec is built on distributed hypothesis, where the circumstance for each word is in its neighboring words. Here, based on the neighboring words the target words are predicted. The skip-gram model works well, with small amount of training data also, and denotes well, even uncommon words.

The above word embedding training model consists of an input layer, projection process, weighted matrix connected to output layer.[25] The probability of neighboring words in the repository will get maximized with a proper training of each and every word vector. In the Fig. 2 showing word embedding model is taken from the towards data science website.

Skip-gram Model

Finding semantic similarity among the text is the vital component in text processing. It relies on the neural network model to train word vectors to predict neighboring words. In this model, the center word is going to be input and predictions are context words.[26] Word embedding concentrated on converting words into vectors while retaining the relationship among words. The skip-gram model takes word vector as input and proceeds to predict contextual words from the word repository as shown in the Fig 3.

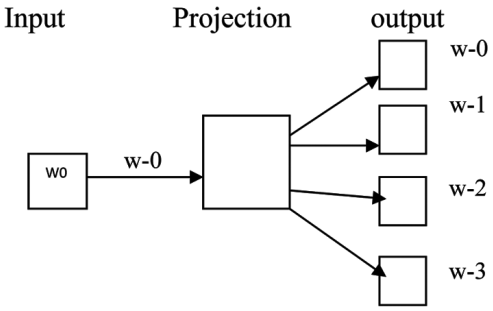

Fig. 3 — Skip-gram model to find context word

Here each word is represented as n-dimensional vector and each context is given as d dimensional vector.[27] Two matrices of random weights W, C are initialized to all vectors. The focus of this model is to extract the semantic similarity of the individual word pairs into the document distance matrix.

$$\frac{1}{N}\sum_{n=1}^{N}\sum_{j\in nb(n)} \log\ p(\frac{wj}{wn}) \quad \ldots (4)$$

Where *nb(n)* is the set of neighboring words of word wt and *p(wj/wi)*. The associated vector V_{wj} and V_{wn}.

Algorithm

Input: 1. Examiner answer (Standard answer)-E_a
2. Student answer scripts – S_a
Output: Evaluate and award Marks - M_i
Step 1: Data Pre-processing of student answer scripts-S_a

1.1. Stop word elimination
1.2. Stemming

Step 2: Perform word2vec embedding using skip gram.

$$\frac{1}{N}\sum_{n=1}^{N}\sum_{j\in nb(n)} \log\ p(\frac{wj}{wn})$$

Step 3: Find semantic similarity among standard answer, student answer input file-using word mover's distance using the following

$$min_{T\geq 0}\sum_{i,j=1}^{n} T_{ij}c(Sa,Ea)$$

Given the constraints:

$$\sum_{j=1}^{n} T_{ij} = d_{sa}$$

, and

$$\sum_{i=1}^{n} T_{ij} = d'_{Ea}$$

Step 4: Measure and allocate marks accordingly

Step 5: Repeat step1 to Step 3 until no scripts in the storage device

Implementation Process

The experimentation process is carried using i3-5020 u with 2.20 GHz processor with 4 GB RAM on Windows 10 operating system 64 bit. Python 3.7 is used to implement the proposed model. The data set is generated from Moodle — online examination platform. The data set is completely pre-processed by dividing the continuous text into words, symbols, and various significant elements called tokens. The identified tokens considered as input for more processing such as parsing. The tokenization process is beneficial for data analysis. Data set has many different characters; a few characteristics are mentioned in the Table 1.

Experimental Results & Analysis

The experimentation is conducted on the synthetic data set and its important characteristics are mentioned in the Table 2. The evaluation process of the proposed model is done with various common evaluation metrics. The evaluation measures like f1-score, precession, and recall are used to estimate the proposed model efficiency. In Table 3 sample data set is presented and it presents the characteristic of the data used in experimentation.

Evaluation is measured by using F1-score.

$$\text{f1} - \text{score} = \frac{2*\text{precision}*\text{recall}}{\text{precesion}+\text{recall}} \quad \ldots (5)$$

where precession and recall are calculated as follows

$$\text{precession} = \frac{\text{tp}}{\text{tp}+\text{fp}} \quad \ldots (6)$$

Table 1 — Some of the sample characteristics in the data used in experimentation

Subject	Maximum Grade Points or marks award	Number of Answers in each subject	Average number of words in the answers (Range)	Size of training data set (No. of Scripts)	Size of test data set (No. of Scripts)
DBMS	40	5	500–600	120	80
O.S	40	5	530–650	100	90
S.E	40	5	560–700	120	90
FCS	40	5	480–660	800	600
CN	40	5	500–660	120	90

Table 2 — Some of the sample characteristics in the data used in experimentation

Data set subject wise	Manual Evaluator-1 score out of 40 marks					Manual Evaluator-2 score out of 40 Marks					Deviation percentage	Std Deviation
	S1	S2	S3	S4	S5	S1	S2	S3	S4	S5		
DBMS	34	32	29	28	26	30	27	21	20	19	25	16
OS	33	38	26	25	29	25	30	19	19	23	27	17.5
SE	30	36	28	26	26	21	30	22	20	22	24	15.5
FCS	38	37	34	33	30	29	30	29	27	29	18	14
CN	31	30	24	27	31	29	25	20	23	26	16	10
Average Deviation											22	

Table 3 — Some of the sample characteristics in the data used in experimentation

Data set subject wise	Proposed system iteration -1 for the 40 Marks					Proposed system iteration -2 for the 40 Marks					Deviation Percentage	Std Deviation
	S1	S2	S3	S4	S5	S1	S2	S3	S4	S5		
DBMS	30	29	23	30	25	33	30	23	30	25	3	2
O.S	33	32	26	26	28	30	35	28	27	28	3	1.5
S.E	32	35	31	29	29	32	35	30	28	28	2	1.5
FCS	39	37	36	35	34	36	36	35	34	33	4	3.5
CN	34	32	28	28	35	33	32	26	29	30	5	3.5
Average Deviation											3.4	

$$\text{recall} = \frac{\text{tp}}{\text{tp+fn}} \quad \ldots (7)$$

Here, tp = true positives, fp = false positives and fn = false negatives.

In the proposed model, the scores are supposed to be awarded to each question based on the semantic similarity among actual answer given by the examiner or key and student answer. The semantic similarity comparison among the key and student answer is performed on the word mover's distance score. Comparatively word movers distance metric provides better results than other semantic similarity measures. The experiment is conducted with two manual evaluators (examiners) and automation with the proposed model on the same data set. The following results showing that proposed model provides effective results than that of the manual evaluation process.

The percentage deviation of manual evaluation of different subjects is depicted in Fig. 4 with percentage difference among two different manual evaluations and their standard deviation. From this experimental evaluation, it is observed that manual evaluation is giving more deviation from one evaluation to the other evaluation. Ignoring the discussion on parameters that will affect the manual evaluation, the results clearly depicts that an unacceptable deviation may be introduced with the manual evaluation.

The percentage deviation of proposed automated evaluation of different subjects is depicted in Fig. 5 with percentage difference among two different proposed automated evaluations and their standard deviation. The results have shown here elevating that, there is less deviation among number of iterations in automated evaluation. It is evident that the proposed method overcomes the impact of manual evaluation. The comparison of manual and proposed method is presented in Fig. 6. The analysis is performed with percentage deviation among both the methods. It is noticed that more deviation is obtained in manual evaluation than compared to proposed method, and it is evident that the proposed method outer performs with acceptable deviation.

The proposed model is also evaluated using quadratic weighted kappa, which can measure the likeness to the correlation coefficient. It is an evaluation measure used for text data analysis. Here in the given context, the quadratic weighted kappa measure is used to calculate the degree of agreement among two examiners (evaluators) one is the manual and, the other is proposed semantic evaluation method (automated). The kappa metric generally ranges from 0 to 1, here 1 indicates complete agreement between examiner and automated system and 0 is other case.

The proposed method is compared with the existing "Ramachandran Approach". The experimental results are showing that the proposed methods work in an optimized way comparatively with other existing methods in the literature, which is presented in Fig. 7. The proposed and Ramachandran

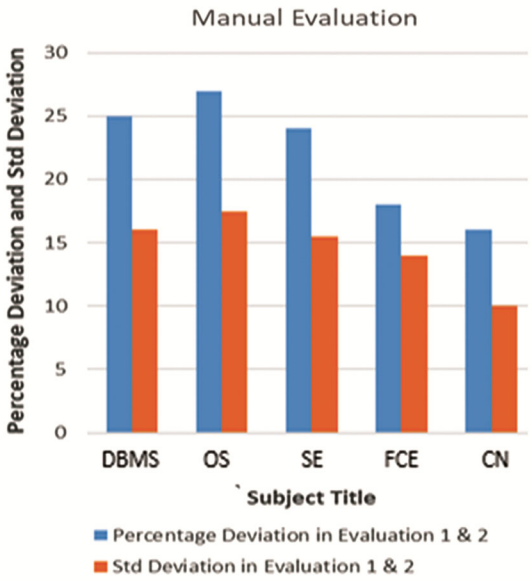

Fig. 4 — Deviation parameters on manual evaluation

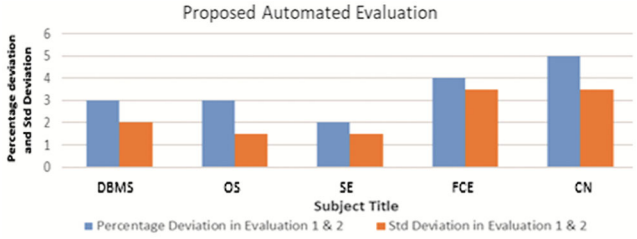

Fig. 5 — Deviation parameters on propose automated evaluation

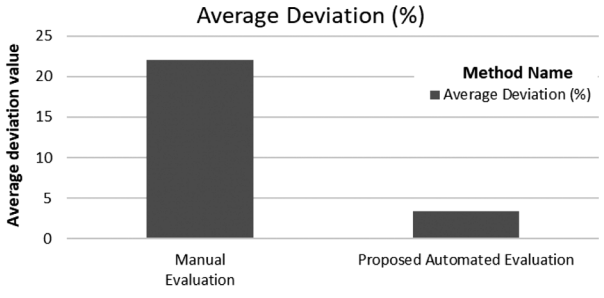

Fig. 6 — Average deviation among Manual vs. Proposed automated evaluation

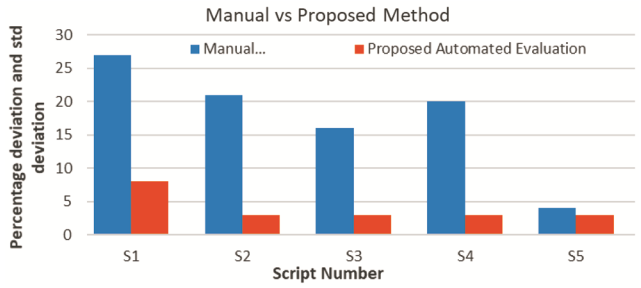

Fig. 7 — Individual script analysis on manual and proposed methods

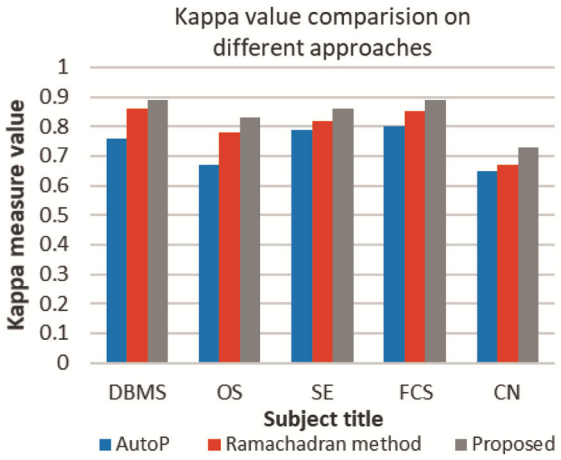

Fig. 8 — Comparative analysis using Kappa measure

methods works based on the semantic metrics and related words can show alternative answers. The comparative analysis results are presented in the Fig. 8, the results are showing that the proposed approach works better than existing methods.

Conclusions & Future Work

In the online examination pattern, it is very much essential to introduce the automated evaluation of the student answer scripts to avoid the biasing of the manual evaluation. The proposed model works on the automated evaluation of descriptive exam answer scripts using word mover's distance. It mainly aims at to measure the semantic similarity among the document vocabulary. After a proper pre-processing phase, the extracted input data from the online examination platform, semantic similarity among the student answer and answer key has been computed. With the obtained results, it is observed that the proposed model is succeeded in identifying the semantic similarity among the input documents. And it worked well, when compared to the manual evaluation and thus, it resolves the overhead involved in that process. To elevate these various statistical measures and likeliness measure of correlation coefficient are used and, with these, the proposed method witnessed as better than compared to the existing methods. The limitations here noticed are addressing the automated evaluation of mathematical equations and image similarity components. As a future scope, these limitations can be addressed and new innovative refined learning models can be proposed. More analysis implement machine-learning methodology in future.

References

1 Maguitman G, Menczer F, Roinestad H & Vespignani A, Algorithmic detection of semantic similarity, *Proc Int Conf World Wide Web*, (2005) 107–116.

2 Mohamed & Oussalah M, A hybrid approach for paraphrase identification based on knowledge-enriched semantic heuristics, *Lang Resour Eval*, **54** (2019) 457–485.

3 Li Y, Bandar Z & Mclean D, An approach for measuring semantic similarity using multiple information sources, *J Trans Knowl Data Eng*, **15** (2003) 871–882.

4 Sahami, Heilman T D, A web based kernel function for measuring the similarity of short text snippets, *Pro Int Conf on World Wide Web*, (2006) 377–386.

5 Chen H H, Lin M-S & Wei Y-C, Novel association measures using web search with double checking, *Int Conf on Comp Linguist Ann Meet ACI*, (2006) 1009–1016.

6 Chen F, Topic-based document segmentation with probabilistic latent semantic analysis, Conference: Proceedings of the 2002 ACM CIKM, *Int Conf Proc Info Knowledge Magt*, (2002) 4–9.

7 Anna Huang, Similarity measures for text document clustering, *Pro Int Conf proc NZCSRSC*, (2008) 53–65.

8 Yuejin X & Reynolds N, Using text mining techniques to analyze students written response to a teacher leadership dilemma, *Int J Comput Theory Eng*, **4** (2012) 575–578.

9 Wan S & Angryk R A, Measuring semantic similarity using WordNet-based context vectors, *Pro Int Conf Sys Man and Cyber*, **23** (2007) 908–913.

10 Williamson D, A framework for implementing automated scoring, *Anl Meet American Edu Res Assoc (AERA) & National Coun Measur Edu (NCME)*, **31** (2009) 2–13.

11 Attali Y, A differential word use measure for content analysis in automated essay scoring, *ETS Res Report Ser*, **36** (2011) 1–19.

12 Schultz M T, The Intelli Metric, Auto Essay Scoring Engine-A Review and an App to Chinese Essay Scoring, in *Handbook of Automated Essay Evaluation*, (2013) 89–98.

13 F B Hasim Sak, Senior A, Long short-term memory recurrent neural network architectures for large scale acoustic modeling, *Con Proc Ann Speech Comm Assoc*, (2014) 1–5.

14 Pawar A, Mago V, Calculating the similarity between words and sentences using a lexical database and corpus statistics, *J Tran Know Data Eng*, **23** (2018) 234–243.

15 Huang G, Guo C, Kusner M J, Sun Y, Weinberger K Q & Sha F, Supervised Word Movers Distance, *Proc Int Conf Neural Inf Syst*, (2016) 342–352

16 Gülle K J, Ford N, Ebel P, Brokhausen F & Vogelsang A, Topic Modeling on user stories using word mover's distance, *J Comp Lan*, (2020) 175–186.

17 Franciscus N & Ren X, Wang J & Stantic B, Word mover's distance for agglomerative short text clustering, *J Intel Informat Database Syst*, **35** (2019) 165–176.

18 Wu X & Li H, Topic mover's distance based document classification, *Proc Int Conf on Comm Tech (ICCT)*, (2017) 1998–2002.

19 Kaya T, Performance prediction reliability of computer-aided work simulations and employment tests: a case of selecting blue-collar employees for repetitive tasks, *J Sci Ind Res*, **80** (2021) 1096–1106

20 Sabarivani A, Ramadevi R, Pandian R & Krishnamoorthy N R, Effect of data preprocessing in the detection of epilepsy using machine learning techniques, *J Sci Ind Res* **80** (2021) 1066–1077.

21 Kusner Matt J, Sun Y, Kolkin N I & Weinberger K Q, From Word Embedding to Document Distances, *Proc Int Cofn Mach Learn* JMLR, **37** (2015) 234–245.

22 Ligthart A & Catal C, Systematic reviews in sentiment analysis: a tertiary study, *J Artif Intel Review*, **54** (2021) 997–5053.

23 Riyanarto N & Kelly A, Semantic Recommender System based on Semantic Similarity using Fast Text and Word Mover's Distance *J Intel Eng & Syst*, **14** (2021) 377–385

24 Murty M R, Murthy J V R, P V G D P Reddy & Satapathy S, A survey of cross-domain text categorization techniques, *Proc Int Conf Recent Adv Info Tech*, (2012) 499–504.

25 Shermis & Hamner M D, Contrasting state-of-the-art automated scoring of essays: Analysis, *Proc in Ann Nat Council Mgt Edu Meet*, (2012) 14–16.

26 Zupanc K & Bosnic Z, Advances in the field of automated essay evaluation, *J Informatica*, **39** (2015) 383–395.

27 Mehamood A & Won B, Prognosis essay scoring and article relevance using multi-text features and machine learning, *J Sym*, **9** (2017) 1–16.