Intelligent Automation And Soft Computing, 2020

Vol. 26, no. 1, 169–180
https://doi.org/10.31209/2019.100000138

An E-Assessment Methodology Based on Artificial Intelligence Techniques to Determine Students' Language Quality and Programming Assignments' Plagiarism

Farhan Ullah[1,4], Abdullah Bajahzar[2], Hamza Aldabbas[3], Muhammad Farhan[4], Hamad Naeem[1], S. Sabahat H. Bukhari[4,5], Kaleem Razzaq Malik[6]

[1]College of Computer Science, Sichuan University, Chengdu 610065, China
[2]Department of Computer Science and Information, College of Science at Zulfi, Majmaah University, Zulfi 11932, Saudi Arabia
[3]Prince Abdullah bin Ghazi Faculty of Information and Technology, Al-Balqa Applied University, Al-Salt- Jordan
[4]Department of Computer Science, COMSATS University Islamabad, Sahiwal Campus, Sahiwal 57000, Pakistan
[5]College of Computer Science, Chongqing University, Chongqing 400044, China
[6]Department of Computer Science & Engineering, Air University, Multan Campus, Multan 60000, Pakistan

ABSTRACT

This research aims to an electronic assessment (e-assessment) of students' replies in response to the standard answer of teacher's question to automate the assessment by WordNet semantic similarity. For this purpose, a new methodology for Semantic Similarity through WordNet Semantic Similarity Techniques (SS-WSST) has been proposed to calculate semantic similarity among teacher' query and student's reply. In the pilot study-1 42 words' pairs extracted from 8 students' replies, which marked by semantic similarity measures and compared with manually assigned teacher's marks. The teacher is provided with 4 bins of the mark while our designed methodology provided an exact measure of marks. Secondly, the source codes plagiarism in students' assignments provide smart e-assessment. The WordNet semantic similarity techniques are used to investigate source code plagiarism in binary search and stack data structures programmed in C++, Java, C# respectively.

KEYWORDS: Electronic-Assessment, Machine Learning, Artificial Intelligence, WordNet, Technology Enhanced Assessment, Semantic Similarity.

1 INTRODUCTION

CURRENTLY, data science is an emerging field getting more consideration by researchers. It is an interdisciplinary science, including statistics, mathematics, programming, problem-solving, data gathering and reasoning, etc. It has the solutions for multidisciplinary issues like information retrieval, data analysis, structured and unstructured data. It facilitates deep learning, artificial intelligence, data mining and research domains in an innovative way. Artificial intelligence generally prominences on automating the procedure of predicting problems and providing solutions. Consequently, a large volume of data is needed to accomplish machine learning and deep learning techniques. This process also supports technologies associated with big data and data mining. E-assessment and electronic learning (e-learning) got more attention in the last few years, but effectiveness and usefulness regarding students' e-assessment are not well understood. The e-assessment is used to assess students automatically based on the syntactic or semantic matching of the teacher's question and student answers. Secondly, the plagiarism detections in students programming tasks greatly improve the e-assessment process. Currently, Students directly convert the same logic of one programming language to another by using online tools. Because of this, students often do not try to get the logic of the assigned projects (Bakker, 2014). Further, the text mining and information retrieval techniques are used to provide text similarity. These techniques may be implemented on different domains i.e. source code plagiarism and text similarity, textual description (Sidorov, Gelbukh, Gómez-Adorno, &

CONTACT Farhan Ullah farhankhan.cs@yahoo.com, https://orcid.org/0000-0003-2422-575X)

Pinto, 2014; Sidorov, Gómez-Adorno, Markov, Pinto, & Loya, 2015) and so on. Various types of text represent objects, and these objects' features are transformed to information using Vector Space Model (VSM) with Term Frequency-Inverse Document Frequency (TF-IDF) value in Natural Processing Language (NLP), (Al Otaibi, Safi, Hassaïne, Islam, & Jaoua, 2017). However, these methods focus mainly on keyword-based similarity and ignore the semantic relations between texts. WordNet (Sidorov et al., 2014), or some other semantic database is used to capture the semantic connections between words. In these databases, all words with complete information like root word, synonyms details, tagging information are given. Annotated corpus is used to translate the keywords of semantics. Answer selection in question answering framework is an open research challenge to rank the relevant answer to the user. For example, yahoo answer in which a user input a query and then experts give answers after that the user search the appropriate response manually.

With the rapid development of the internet, there is a need to process a massive volume of data to retrieve meaningful text against the user's query. Teacher uploads a question online in e-learning education, and then students give replies in response to the teacher's question. Now there is a need to process all these replies and rank the most appropriate reply repeatedly in terms of semantically matching keywords. The conventional technique is, that teacher studies all replies and then gives grades to relevant reply. It is time consuming and erroneous job. There is a need for an e-assessment method to read all replies automatically using machine learning methods and assign marks to relevant replies through semantics. WordNet lexical database is used to match the keywords by synonyms or semantic similarity functions. In the proposed research, an e-assessment methodology has been proposed to assign marks automatically to students' replies against the teacher's question based on WordNet semantic similarity techniques. The WordNet similarity techniques are applied to notice source code similarity in students' projects given in pilot study 2. The teacher can check the plagiarism in source codes performed by students and then assign marks. It encourages the learning behaviour of students in programming. The automatic grading of students' answers saves time and requires less effort from the teacher. It does not remove the teacher, and it is an alternate automatic methodology to assess the students with less effort and time. The main goals of the proposed methodology are as:

- A methodology for e-assessment of students' replies to determine the language quality of answers.
- WordNet semantic similarity measures have been used to mark the students' replies based on semantics matching.
- A comparison is made between teacher's, and SS-WSST calculated marks, and it has been proved that our proposed SS-WSST methodology is much better to mark students' replies automatically.
- Plagiarism detection in students' programming assignments for smart e-assessment

The rest of this paper is planned as follows. The related work is explained in section II. Material and methods of our proposed methodology are discussed in section III. Results and discussion are given in part IV. Finally, part V includes conclusion and future work.

2 RELATED WORK

IN (Luaces, Díez, Alonso-Betanzos, Troncoso, & Bahamonde, 2016), the author attempted to evaluate the knowledge of students on the Massive Open Online Courses (MOOCs). Automatic e-assessment is done with many assignments through multiple choice questions. VSM technique is used to provide similar text and further, to find from answers (Mackness & Pauschenwein, 2016; Watson et al., 2016). The monolingual and cross-lingual semantic textual similarity in e-assessment is a big issue in question answering framework. In (Agirre et al., 2016), the author introduced a question answering forum, in which text similarity is done in English lingual data and cross-lingual Spanish data. Author calculated semantic similarity in snippet text pairs. In (Kastner, Antony, Soobiah, Straus, & Tricco, 2016), the author applied an e-assessment methodology to select the most relevant answer in research question answering framework. The synthesis method is used to rank the similar text in response to a specific query (Shurygin & Krasnova, 2016). A conceptual algorithm is designed to analyze the information of a research query (Kang, Moon, Jang, Lim, & Kim, 2016). The algorithm is investigated using e-assessment of 6-12 years Korean children to evaluate the life quality in terms of allergic rhinitis. The 277 number of nominated from middle schools. Further, the students are allocated into three groups' i.e. allergic-rhinitis (AR), non-allergic rhinitis (non-AR), and controls. Moreover, it is defined that the e-assessment by questionnaires is beneficial for judging the worth of life in Korean kids. The e-assessment man includes multiple choice options or short queries that are easy to comprehend. In (Burrows, Gurevych, & Stein, 2015), the automatic short answer grading (ASAG) methodology is designed for e-assessment . It is described that short answer questions methodology requires relevant reply, text length and text content.

In (Kim, Chern, Feng, Shaw, & Hovy, 2006), the author proposed e-assessment in speech act analysis in web forums. Online discussions were conducted through a set of speech of activity patterns. It

described how different speech patterns identify the discussion threads and how they facilitate the automatic question answering framework. Knowledge-based information retrieval plays a significant role in e-assessment of question answering framework. In (Otegi, Arregi, Ansa, & Agirre, 2015), the WordNet ontology is used to analyze the words relationship for e-assessment. The pseudo-relevance feedback (PRF) technique is designed for query expansion. The author showed better results by using Wikipedia as a data repository. In (Partalas et al., 2015), To assess the text classification in question answering, the author proposed Large Scale Hierarchal Text Classification (LSHTC) for a large number of classes in a dataset. The corpora of Wikipedia and web directory were used to assess the hierarchal text classification. The training dataset of LSHTC is available online and may be downloaded for further experiments. In (Cigdem & Oncu, 2015), The TAM2 model is used to conduct e-assessment by e-quizzes among military vocational college students. The equation modelling method is designed to convey the grade and age of scholars. Further, results show that behavior objective has dramatically improved by the perception of the question's content. Cosine text similarity with VSM broadly used to rank the related text in a document. In (Chen & Van Durme, 2016), the author proposed a model that enter text into a context-sensitive environment and then calculate similarity in text pairs. They used an unsupervised approach, which ranked a similar version by using cosine similarity measure. In the last few years, the tree Edit Distance (TED) got more attention to calculate resemblance in various documents. In (Sidorov et al., 2015), the author described that, the TED uses syntactic n-grams to compute the similar text. Further, it calculates soft similarity among text documents (Pawlik & Augsten, 2016). The syntactic n-grams are a non-linear tree shaped, and TED algorithm is used to regain similarity among tokens (Piernik & Morzy, 2017). Further, the authors broadly explained the applications of TED in different scenarios for extracting similar text (Bringmann, Gawrychowski, Mozes, & Weimann, 2017; Spaendonck, de Vries, & Gieseke, 2016). In NLP text similarity plays a vital role like question answering, entity disambiguation, author attribution, and so forth. In (Sidorov et al., 2014), the author presented an idea to use soft cosine similarity measure using VSM between syntactic n-grams. In previous years words and n-grams used for VSM to extract similar text. The author used machine learning algorithms for translation of VSM to calculate similarity (Lacey et al., 2017). In (Ullah et al., 2018; Xu et al., 2015), the author matched four similarity measures Latent Semantic Analysis (LSA) for tokens, LSA for words, VSM for words and VSM in terms of different conditions. These are used to calculate the similarity among academic papers and patents. The author also showed that term based VSM measure gave more accurate result than others. In (Jiang, Kim, Banchs, & Li, 2015), the authors proposed an idea of infrequently question answering in the Chinese language. Pairwise objects presented in VSM in different dimensions.

The proposed research fills the gap in e-assessment methodology to automate the grading of student's answers based on semantics and also to detect plagiarism in students' source codes.

3 SS-WSST METHODOLOGY

TEXT similarity is a basic and essential task in NLP because the user is interested to see the most relevant text. WordNet (Sidorov et al., 2014), or some other semantic database is used to capture the semantic relations between words. In (Delen, 2015; Shum et al., 2016), the author described the student's behavior in computer based training and testing in two different conditions and then compared the resultant scores. It was shown that we could enrich the results of the computer-based test if the student has an optimum response time in this experiment. An e-assessment is investigated among undergraduate students from Virtual University of Pakistan (VU) in Open Source Web Application Development course. The case study is divided into different stages for semantics and marks ranking analysis. The semantic resemblance is measured concerning the teacher's query and the student replies to classify the most relevant response. An algorithm has been proposed for semantic similarity between sentences using different linguistic information in question answering framework as shown in Algorithm 1. In SS-WSST methodology keywords are extracted from teacher's question and students' replies and converted to Question Keywords Vector (QKV), and Students Replies Vector (SRV). The preprocessing parameters root word, stemming, the frequency of each token is used to extract QKV and SRV. A machine is used to apply WordNet semantic similarity techniques on QKV and SRV to measure the similarity scores. The Path Length, Lin, Wu & Palmer gave a rating in the range of 0 to 1, and Hirst & Onge gave a score in the field of 0 to 16. Normalization method is used to scale the semantic relatedness score in each range using equation 1.

$$Nval = (Cval / SM_{\max}) + SM_{\min} \quad (1)$$

The Nval denotes normalized value, Cval is Computed Value by Similarity Measures, SMmax represents maximum range of Similarity Measure and SMmin minimum range of Similarity Measure. Semantic measure score table is calculated using equation 2.

$$ScoreTable = \sum_{i=1}^{n} \sum_{k=1}^{k=4} ScoreTable_{i,k} \quad (2)$$

A comparison has been made to validate the resultant scores in both tables. It has been observed that the grading score table presents improved scores as compared to manually allocated marks because the instructor has only four choices to grade the student. The proposed SS-WSST methodology gave a more accurate score by using semantic similarity techniques as shown in Figure 1. The undergraduate students' dataset is collected from LMS of VU, Pakistan. The dataset contains an online evaluation of students in the undergraduate course, i.e. Open Source Web Application Development, Spring semester, 2016. The examination is conducted from August 11, 2016, to August 12, 2016. The teacher uploads a question on LMS from the Open Source Web Application Development course, and students give replies. Then the teacher read all responses manually and gave marks in 1 to 5 range.

4 WORDNET SEMANTIC SIMILARITY TECHNIQUES

WORDNET is a dictionary of different words with synonyms details that can be used for semantic similarity. The WordNet semantic similarity techniques are used to match text semantically (Kutuzov et al., 2018). The Path Length, Lin, Wu & Palmer and Hirst & St-Onge techniques are applied to our dataset to mark replies in 1 to 5 range. Then, compare the teacher's marks with semantically calculated marks as shown in the results and discussion section.

The dataset contains the instructor's query and students' response. The text is preprocessed for the stemming, root word, the frequency of each token etc. The methodology is shown in algorithm one where QV represents Question Vector, and CV represents Comment Vector. The keywords are picked one by one from QV and CV and WordNet semantic similarity techniques are applied. The semantic calculated values are accumulated in a Score Table. After that, the Score Table is compared with instructor's marks. In summary, the semantic similarity calculation process can be described in Algorithm 1.

Algorithm 1 Semantic similarity calculation

Input: Question keywords Vector (QV), Comments keywords Vector (CV)
Output: Semantic Similarity Score table using WordNet

1. Start
2. For (i=0; i <= QV. length)
3. Pick a word from QV and a word from CV
   Check similarity of words using:
4. ScoreTable[i,1] = WordNet:Similarity::Path(QV[i], CV[i])
5. ScoreTable[i,2] = WordNet::Similarity::lin(QV[i], CV[i])
6. ScoreTable[i,3] = WordNet::Similarity::wup(QV[i], CV[i])
7. ScoreTable[i,4] = WordNet::Similarity::hso(QV[i], CV[i])
8. End of loop
9. Return Semantic Similarity ScoreTable

4.1 Path Length

It works on the counting of nodes along the shortest path between synset1 and synset2. The semantic relatedness is inversely proportional to nodes in the shortest path. It calculates the relatedness value in 0 to 1 range. If two synsets are same, then maximum relatedness value will be 1 (Pedersen, Patwardhan, & Michelizzi, 2004).

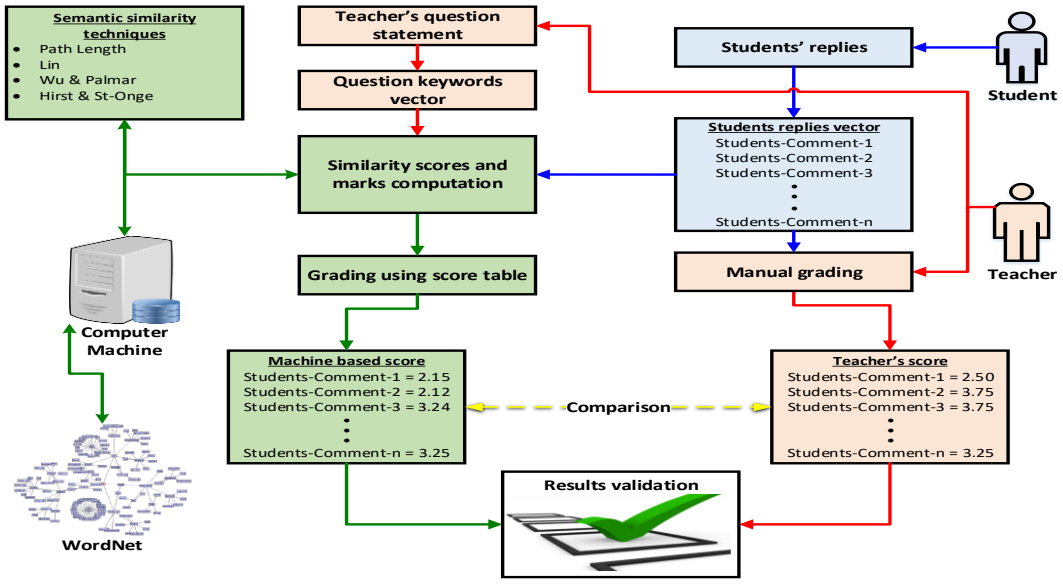

Figure 1. **SS-WSST methodology of Student Replies Assessment using WordNet-based similarity calculation and Teacher' marks**

4.2 Lin

The Lin measure works for information content. It gives the semantic relatedness value in 0 to 1 range as shown in Equation 3. If the information content of synset1 and synset2 is zero, then semantic relatedness value calculated from Lin measure is zero (Pedersen et al., 2004).

$$Lin = \frac{2 \times IC(LCS)}{IC(synset1) + IC(synset2)} \quad (3)$$

where IC is information content, and synset is synonyms' information given the word and LCS (Least Common Subsumer) of synset1 and synset2.

4.3 Wu & Palmer

It computes semantic similarity value by studying the depths of two synsets in WordNet nomenclatures to the depth of LCS as shown in Equation 4 (Pedersen et al., 2004).

$$Wu\ \&\ Palmer = \frac{2 \times depth(LCS)}{depth(s_1) + depth(s_2)} \quad (4)$$

The relatedness value is in the range of 0 to 1. In this technique, the score will not be 0 because the depth of LCS is never 0 between the two synsets, so it gave the better result in our experiment.

4.4 Hirst & St-Onge

It works by taking lexical information between the two-word meanings. The semantic relatedness score is in the range of 0 to 16. It has further three classes extra strong, medium strong and robust which is used for calculating semantic relatedness score (Pedersen et al., 2004).

5 RESULTS AND DISCUSSIONS

THE dataset consists of 210 pairs of words selected after preprocessing steps from different undergraduate students' replies. In our proposed SS-WSST methodology, four WordNet's semantic similarity techniques are applied to the dataset to assigned marks automatically in 1 to 5 range. In the lot study 42 pair words selected from 8 different students' replies. A comparison has been made with the manually allocated marks from the instructor. It is shown that our designed approach provided exact marks to students' replies compared with instructor's marks. To investigate the plagiarism in source codes, we have collected dataset contains C++, Java and C++, C# programmed in binary search and stack given in pilot study 2.

5.1 Pilot Study 1

In this research, keywords are extracted from 8 students' replies. In the next step, WordNet semantic similarity techniques are applied to calculate semantic relatedness score between question keywords and students' reply keywords. The information of reply-1s' keywords with questions' keywords with semantic relatedness scores as shown in Table 1.

Table 1. Similarity Scores of Reply-1

Standard answer's Keywords	Reply-1 Keywords	Path Length	Lin	Wu & Palmer	Hirst & St-Onge
Dominate	Demand	0.33	0.00	0.50	0.00
Server	Language	0.10	0.00	0.30	0.00
	Programmer	0.17	0.00	0.67	2.00
Script	Language	0.20	0.52	0.70	3.00
	Programmer	0.09	0.00	0.28	0.00
Language	Programmer	0.11	0.00	0.33	0.00
	Skill	0.25	0.52	0.80	4.00
Scaled values (1-5)		1.89	1.74	3.57	1.40

Is experiment, the Wu & Palmer technique gave better results than other methods as it considers depths of two synsets in WordNet taxonomies along the depth of LCS. It is analyzed in Table 2, that reply 6 is more relevant to the question and assigned highest value 4.47 by Wu & Palmer technique while response 2 is less like the question, which has 1.24 value by Lin technique, but teacher assigned 3.75 marks to reply two manually. This experiment has assigned highest value 4.47 to reply 6 and teacher also assigned high marks 3.75 to reply 6, which is proved that semantic relatedness score gave better relevancy results than the allotted marks from the teacher as shown in Table 2.

In Figure 2, reply details are given horizontally, and ranking of students' replies are given vertically. The ranking is provided in a range of 1 to 5. Semantic relatedness score for keywords is shown in different colours. Magenta colour shows teacher's marks, dark blue colour shows Hirst & St-Onge score, light blue shows Path Length score, Gray colour shows Wu & Palmer score and orange colour shows Lin-Score. It shows a comparison between teacher's marks and semantic relatedness scores from given techniques. The Wu & Palmer scores gave the better relevancy score compared with instructor's marks than other used techniques because it calculates semantic relatedness score between the depths of two synsets along the WordNet taxonomy of words. The dotted curve in Fig. 2 shows the moving average or running average in statistics, which analyzed data points by calculating a series of averages from different subsets of the dataset. It took the average of first two replies' scores, and the result is the moving average of first two points, then took the average of reply 3 and reply

Table 2. Comparison of WordNet semantic similarity measures with Teacher's marks

Students' Replies	Path Length	Lin	Wu & Palmer	Hirst & St-Onge	Teacher's Marks
Reply 1	1.89	1.75	3.57	1.40	2.50
Reply 2	2.08	1.24	3.66	1.50	3.75
Reply 3	3.06	2.67	4.35	2.88	3.75
Reply 4	1.83	1.38	3.42	1.42	2.50
Reply 5	2.41	2.47	4.36	1.99	3.75
Reply 6	2.57	2.45	4.47	2.56	3.75
Reply 7	2.77	2.71	4.29	2.19	3.75
Reply 8	2.88	3.51	3.57	2.72	5.00

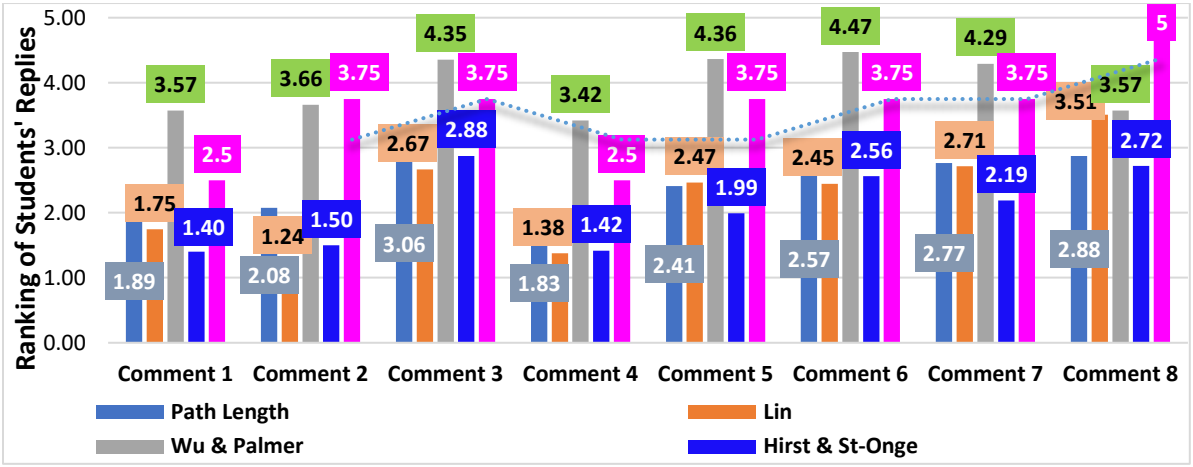

Figure 2. Comparison of Teacher and Similarity measures scoring

4 scores, and the result is the moving average of these two points and so on. Finally, average score's points in the dotted curve showing the comparison among replies' scores with teacher score. It generates different subsets from larger dataset to understand the overall behaviour of the dataset. Mean similarity scores are calculated by used techniques to compare with maximum score. In Table 3, it is shown that maximum semantic relatedness scores close to manually assigned marks to provide better results. The mean similarity measures with teacher marks to compare the overall performance of the proposed methodology as shown in Figure 3. The blue line represents the instructor's marks and the black line displays the mean similarity. Reply 4 gives the same value for mean value and teacher marks. Reply 5 and reply eight values are approximately the same values. The mean and sigma values' comparison is given in Figure 4. The sigma is also called the standard deviation. Wu & Palmar gave better results as compared to other similarity measures. The red line shows the teacher and Wu & Palmar values. The blue line shows the sigma values, and the black line shows the mean of similarity values. The red and blue line showing approximately the same behaviour but the mean line is far. Other related measures are also computed in the mean and because of this is behaving differently. Wu & Palmar is a better option to assess the students automatically. The percentage accumulative contribution of each reply for low similarity as shown in Figure 5. The right vertical line indicates the accumulative values, and the horizontal line shows the corresponding similarity contribution for each reply. Reply 3 contributes more in similarity, but the reply 4 contributes less as compared to other values.

Table 3. Mean and Max Similarity Measure vs Teacher's Marks

Students' Replies	Max out of all measures	Mean of Similarity Measures	Teacher's Marks
Reply 1	3.57	2.15	2.50
Reply 2	3.66	2.12	3.75
Reply 3	4.35	3.24	3.75
Reply 4	3.42	2.01	2.50
Reply 5	4.36	2.81	3.75
Reply 6	4.47	3.01	3.75
Reply 7	4.29	2.99	3.75
Reply 8	3.57	3.17	5.00

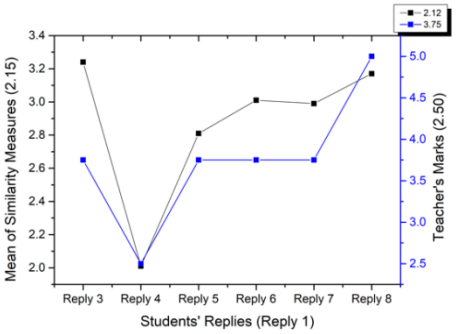

Figure 3. **Comparison of Teacher marks with mean similarity values**

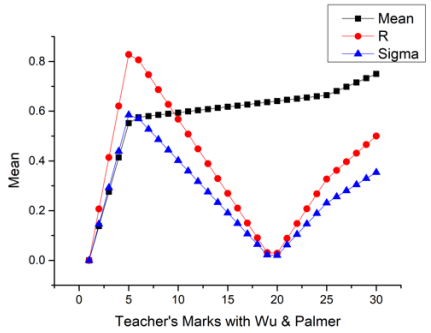

Figure 4. **Teacher marks comparison with Wu & Palmar measure with mean and Sigma values**

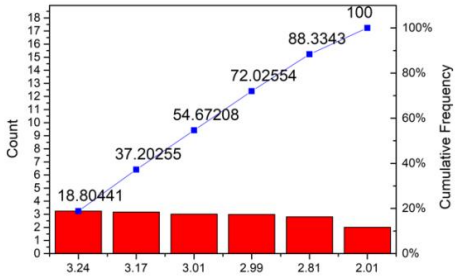

Figure 5. **The percentage contribution for each reply with a cumulative frequency**

5.2 Pilot Study 2

The plagiarism detection is students' programming assignments is a crucial task in e-assessment methodology. Most source codes are available on the internet due to the emerging development of software industry and open source software. Teacher gives programming assignments to students to exercise their programming skills, and they copy the source code of the given task from someone. Students do not use their logic to solve a given programming problem. It discourages the learning process in students (Joy & Luck, 1999). Teacher gives programming tasks to students in one programming language and asks to convert the same logic in another language. Some tools can transform codes from C++ to Java and vice versa (Laffra). So, students may use these tools to covert the source to the target code and submit to the teacher without understanding the inner logic of the code. The students' coding style can also be used to analyze similarity in programming assignments (Mirza, Joy, & Cosma, 2017). We have proposed an idea that is used to detect plagiarism in C++, Java and C++, C# respectively. Although, these are different computer languages and their syntactic and semantic structures are different but still, our proposed idea done well. To analyse our experiment, two case studies, i.e. stack and binary search programmed in C++, java and C#. The keywords are extracted from source codes using preprocessing techniques. The weighting technique is applied to analyze the value of tokens in terms of similarity (Cosma & Joy, 2012) (Evangelopoulos, 2013). The semantic similarity is calculated between each token of C++ and Java in a binary search case study using WordNet similarity techniques are shown in Table 4.

Table 4. **The similarity between C++ and Java (Binary Search) using WordNet Techniques**

C++	Java	Path Length	Wu & Palmer	Lin
Search	Find	0.334	0.800	0.516
Class	Class	1.000	1.000	1.000
Number	Value	0.250	0.727	0.437
Result	Search	0.250	0.625	0.514
Bottom	Last	0.250	0.5833	0.1745
Void	Void	1.000	1.000	1.000
Binary	System	0.142	0.667	0.400
Element	Number	0.142	0.571	0.218
Enter	Out	0.333	0.500	0.000
Value	Number	0.250	0.727	0.437
Find	Search	0.250	0.800	0.336
While	System	0.125	0.461	0.102
Top	Last	0.334	0.667	0.298
Center	Middle	1.000	1.000	1.000
Else	Else	1.000	1.000	1.000
Error	Wrong	0.167	0.706	0.757
Return	Result	0.200	0.706	0.444
Main	Main	1.000	1.000	1.000
Position	Location	1.000	1.000	1.000
Locate	Present	0.334	0.500	0.000
Object	System	0.250	0.714	0.403
Order	Out	0.250	0.667	0.267
Array	List	0.143	0.571	0.325
Order	Static	0.143	0.625	0.232
Percent Similarity		42.272%	73.410%	49.435%

The first and second columns tokens extracted from C++ and Java source codes respectively. The third, fourth and fifth column shows the similarity scores received from Path Length, Wu & Palmer and Lin techniques respectively. The last row shows the per cent similarity value for each technique for the same source code. The mentioned three techniques retrieve similarity values between each pair of tokens in a range of 0 to 1. The Wu & Palmer technique gives

73.41% overall similarity value which is better than others. The Lin gives 49.435% overall similarity value which is better than Path Length but lower than Wu & Palmer. The Path Length gives the 42.272% overall similarity value which is the lowest than others. So, Wu & Palmer is the better choice to investigate the similarity between different source codes. As we have two different source codes, but still the WordNet similarity techniques detected plagiarism based on semantics. The similarity between C++ and C# in stack source codes is given in Table 5. Wu & Palmer gave better similarity results among the three mentioned techniques.

Table 5. The similarity between C++ and C# (Stack) using WordNet Techniques

C++	C#	Path Length	Wu & Palmer	Lin
stack	Stack	1.000	1.000	1.000
class	Class	1.000	1.000	1.000
display	Out	0.250	0.667	0.243
return	Out	0.250	0.706	0.26
push	remove	0.250	0.572	0.391
pop	Empty	0.200	0.500	0.295
empty	current	0.077	0.334	0.000
private	console	0.067	0.417	0.000
public	Push	0.125	0.462	0.091
main	Main	1.000	1.000	1.000
void	Void	1.000	1.000	1.000
maximum	Peek	0.077	0.334	0.000
number	Value	0.250	0.728	0.438
top	Static	0.100	0.400	0.164
top	Peek	0.250	0.400	0.000
enter	Write	0.334	0.834	0.747
source	program	0.334	0.876	0.588
push	Out	0.250	0.706	0.255
push	Push	1.000	1.000	1.000
pop	Pop	1.000	1.000	1.000
Percent Similarity		44.07%	69.68%	47.36%

Further, the similarity between C++ and Java source codes is given in Figure 6. The terms are retrieved from both source codes are given horizontally while similarities values between each pair of tokens are given vertically. The blue, orange and indigo colours showing the similarity values for Path Length, Wu & Palmer and Lin respectively. As these source codes are different in terms of syntax and semantics but still the WordNet similarity techniques retrieving similarity on tokens by tokens comparison.

The proposed idea for plagiarism detection will significantly help the teacher to perceive similarity in students' projects. The teacher will check the plagiarized source codes and then assign marks to students. It provides the smart e-assessment methodology for plagiarism detection in students' programming tasks. It will improve the learning process in students.

6 EXPERIMENTAL IMPLICATIONS

THE proposed experiment is applied to big dataset as well. A sample of the dataset is shown in Figure 7 in terms of WordNet semantic similarity values concerning teacher's marks. The horizontal line indicates the 25 students and the vertical line shows the similarity values. The black color line shows the Path Length, red for Lin, green for W& Palmer, dark yellow for Hirst & St-Onge, blue for teacher's marks and yellow for average values. The Wu & Palmer technique similarity values better similarity, but Hirst & St-Onge technique shows the worst similarity values as compared to teacher's marks. The average line indicates the mean of all similarity values. The replies 11, 23 and 35 shows closer values with teacher's marks by Wu & Palmer technique. These replies are mostly similar to the standard answer as well in terms of semantics. The Wu & Palmer technique is a better choice to use for automatic grading of students' replies because it shows good relevancy to manual values.

The circular graph as shown in Figure 8. represents each reply similarity assessment distribution to different WordNet similarity techniques. Each colour gave to every reply is further divided into a percentile from 0 to 100. The Wu & Palmer gave marks 32 as a whole and teacher assigned 29 which is quite close. Wu & Palmer has the highest similarity accuracy for given queries. Similarity, the Hirst & St-Onge technique gave the lowest score which is 17. It concludes that Wu & Palmer gave better results than other methods under discussion for the WordNet.

7 CONCLUSION

THE E-assessment of students' can be automated by WordNet semantic similarity techniques. The WordNet lexical database can be used to extract semantics from the text rather than just keywords matching. In this experiment WordNet, semantic similarity measures are used to mark the students' replies by matching text semantically. The dataset consists of 210 pairs of words from undergraduate students collected from LMS of VU. An algorithm has been proposed that show the overall working of the proposed methodology. The teacher has only four bins for the assignment of marks, but our proposed methodology gave more accurate scores through semantics extraction. This process does not remove the teacher but, it is an alternate method that can decrease the teacher's effort and time. Moreover, plagiarism detections in student's source codes play an important role in e-assessment. Two case studies are taken in pilot study 2 to detect plagiarism in students' programming tasks based on WordNet similarity techniques. The proposed research provides smart e-assessment in both subjective questions' as well as in programming assignments.

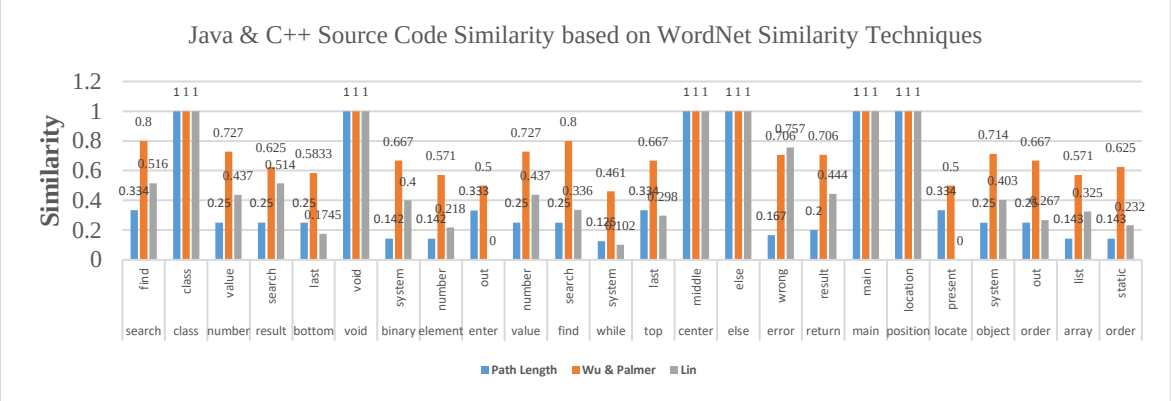

Figure 6. **Similarity between C++ and Java based on WordNet Techniques**

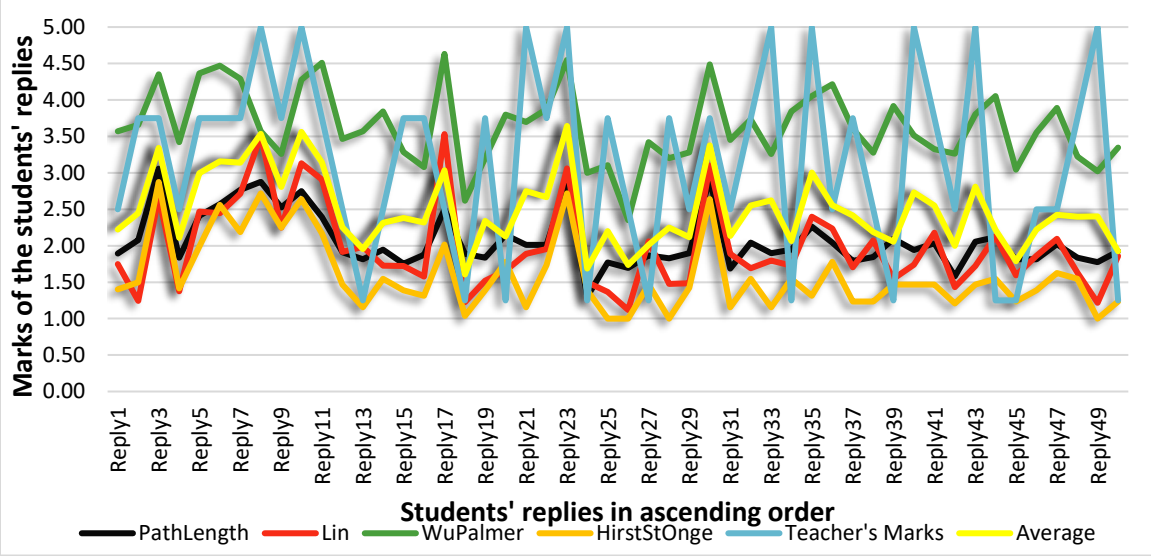

Figure 7. **Students' replies with their scores by the teacher and semantic similarity techniques**

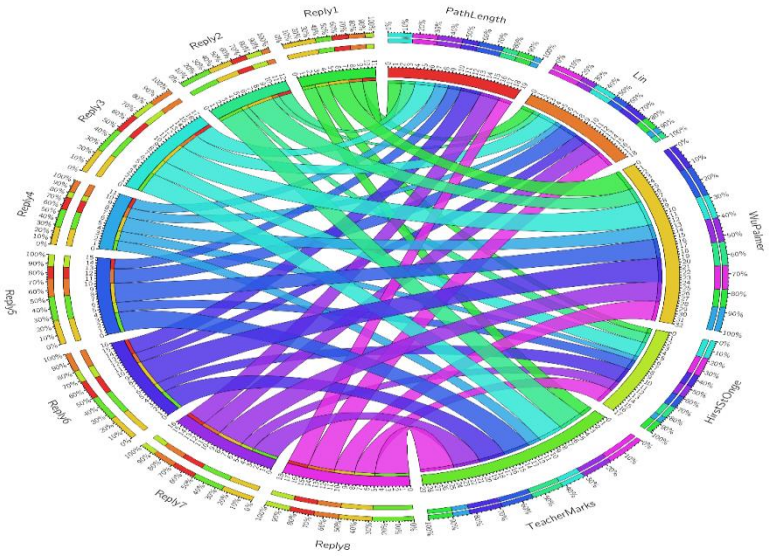

Figure 8. **Results validation for WordNet Semantic Similarity Techniques**

In future, the proposed methodology can be improved by automatic selection of keywords which can give the overall meaning and structure of the student's replies. An algorithm can be designed to select the best keywords automatically by using different semantic measure. The Soft cosine similarity can be used to calculate marks semantically.

8 ACKNOWLEDGEMENT

THIS work was supported by the National Key Research and Development Program (2016YFB0800605, 2016QY06X1205), and the Technology Research and Development Program of Sichuan, China (18DYF2039, 17ZDYF2583)

9 REFERENCES

Agirre, E., Banea, C., Cer, D., Diab, M., Gonzalez-Agirre, A., Mihalcea, R., & Wiebe, J. (2016). Semeval-2016 task 1: Semantic textual similarity, monolingual and cross-lingual evaluation. Paper presented at the Proceedings of the 10th International Workshop on Semantic Evaluation.

Al Otaibi, J., Safi, Z., Hassaïne, A., Islam, F., & Jaoua, A. (2017). Machine Learning and Conceptual Reasoning for Inconsistency Detection. IEEE Access, 5, 338-346.

Bakker, T. (2014). Plagiarism Detection in Source Code. Ph. D. dissertation, Universiteit Leiden.

Bringmann, K., Gawrychowski, P., Mozes, S., & Weimann, O. (2017). Tree Edit Distance Cannot be Computed in Strongly Subcubic Time (unless APSP can). arXiv preprint arXiv:1703.08940.

Burrows, S., Gurevych, I., & Stein, B. (2015). The eras and trends of automatic short answer grading. International Journal of Artificial Intelligence in Education, 25(1), 60-117.

Chen, T., & Van Durme, B. (2016). Discriminative Information Retrieval for Knowledge Discovery. arXiv preprint arXiv:1610.01901.

Cigdem, H., & Oncu, S. (2015). E-assessment adaptation at a military vocational college: student perceptions. Eurasia Journal of Mathematics, Science & Technology Education, 11(5), 971-988.

Cosma, G., & Joy, M. (2012). An approach to source-code plagiarism detection and investigation using latent semantic analysis. IEEE transactions on computers, 61(3), 379-394.

Delen, E. (2015). Enhancing a Computer-Based Testing Environment with Optimum Item Response Time. Eurasia Journal of Mathematics, Science & Technology Education, 11(6), 1457-1472.

Evangelopoulos, N. E. (2013). Latent semantic analysis. Wiley Interdisciplinary Reviews: Cognitive Science, 4(6), 683-692.

Jiang, R., Kim, S., Banchs, R. E., & Li, H. (2015). Towards improving the performance of Vector Space Model for Chinese Frequently Asked Question Answering. Paper presented at the 2015 International Conference on Asian Language Processing (IALP).

Joy, M., & Luck, M. (1999). Plagiarism in programming assignments. IEEE Transactions on Education, 42(2), 129-133.

Kang, H. Y., Moon, S. H., Jang, H. J., Lim, D. H., & Kim, J. H. (2016). Validation of" quality-of-life questionnaire in Korean children with allergic rhinitis" in middle school students. Allergy, Asthma & Respiratory Disease, 4(5), 369-373.

Kastner, M., Antony, J., Soobiah, C., Straus, S. E., & Tricco, A. C. (2016). Conceptual recommendations for selecting the most appropriate knowledge synthesis method to answer research questions related to complex evidence. Journal of clinical epidemiology, 73, 43-49.

Kim, J., Chern, G., Feng, D., Shaw, E., & Hovy, E. (2006). Mining and assessing discussions on the web through speech act analysis. Paper presented at the Proceedings of the Workshop on Web Content Mining with Human Language Technologies at the 5th International Semantic Web Conference.

Kutuzov, A., Panchenko, A., Kohail, S., Dorgham, M., Oliynyk, O., & Biemann, C. (2018). Learning Graph Embeddings from WordNet-based Similarity Measures. arXiv preprint arXiv:1808.05611.

Lacey, A., Lyons, J., Akbari, A., Turner, S. L., Walters, A. M., Fonferko-Shadrach, B., . . . Ford, D. V. (2017). Codifying unstructured data: A Natural Language Processing approach to extract rich data from clinical letters. International Journal for Population Data Science, 1(1).

Laffra, C. A C++ to Java Translator. Advanced Java: Idioms, Pitfalls, Styles and Programming Tips.

Luaces, O., Díez, J., Alonso-Betanzos, A., Troncoso, A., & Bahamonde, A. (2016). Content-based methods in peer assessment of open-response questions to grade students as authors and as graders. Knowledge-Based Systems.

Mackness, J., & Pauschenwein, J. (2016). Visualising structure and agency in a MOOC using the Footprints of Emergence framework. Paper presented at the Tenth International Conference on Networked Learning. Lancaster. http://www/.networkedlearningconference.org. uk/abstracts/mackness. htm.

Mirza, O. M., Joy, M., & Cosma, G. (2017). Style Analysis for Source Code Plagiarism Detection—An Analysis of a Dataset of Student Coursework. Paper presented at the Advanced Learning Technologies (ICALT), 2017 IEEE 17th International Conference on.

Otegi, A., Arregi, X., Ansa, O., & Agirre, E. (2015). Using knowledge-based relatedness for

information retrieval. Knowledge and Information Systems, 44(3), 689-718.
Partalas, I., Kosmopoulos, A., Baskiotis, N., Artieres, T., Paliouras, G., Gaussier, E., . . . Galinari, P. (2015). Lshtc: A benchmark for large-scale text classification. arXiv preprint arXiv:1503.08581.
Pawlik, M., & Augsten, N. (2016). Tree edit distance: Robust and memory-efficient. Information Systems, 56, 157-173.
Pedersen, T., Patwardhan, S., & Michelizzi, J. (2004). WordNet:: Similarity: measuring the relatedness of concepts. Paper presented at the Demonstration papers at HLT-NAACL 2004.
Piernik, M., & Morzy, T. (2017). Partial Tree-Edit Distance: A Solution to the Default Class Problem in Pattern-Based Tree Classification. Paper presented at the Pacific-Asia Conference on Knowledge Discovery and Data Mining.
Shum, S. B., Sándor, Á., Goldsmith, R., Wang, X., Bass, R., & McWilliams, M. (2016). Reflecting on reflective writing analytics: Assessment challenges and iterative evaluation of a prototype tool. Paper presented at the Proceedings of the sixth international conference on learning analytics & knowledge.
Shurygin, V. Y., & Krasnova, L. A. (2016). Electronic Learning Courses as a Means to Activate Students' Independent Work in Studying Physics. International Journal of Environmental and Science Education, 11(8), 1743-1751.
Sidorov, G., Gelbukh, A., Gómez-Adorno, H., & Pinto, D. (2014). Soft similarity and soft cosine measure: Similarity of features in vector space model. Computación y Sistemas, 18(3), 491-504.
Sidorov, G., Gómez-Adorno, H., Markov, I., Pinto, D., & Loya, N. (2015). Computing text similarity using tree edit distance. Paper presented at the Fuzzy Information Processing Society (NAFIPS) held jointly with 2015 5th World Conference on Soft Computing (WConSC), 2015 Annual Conference of the North American.
Spaendonck, P. v., de Vries, A. P., & Gieseke, F. (2016). Comparing Web Page Layouts using Tree Edit Distance.
Ullah, F., Wang, J., Farhan, M., Jabbar, S., Wu, Z., & Khalid, S. (2018). Plagiarism detection in students' programming assignments based on semantics: multimedia e-learning based smart assessment methodology. Multimedia Tools and Applications, 1-18.
Watson, S. L., Loizzo, J., Watson, W. R., Mueller, C., Lim, J., & Ertmer, P. A. (2016). Instructional design, facilitation, and perceived learning outcomes: an exploratory case study of a human trafficking MOOC for attitudinal change. Educational Technology Research and Development, 64(6), 1273-1300.
Xu, H., Zeng, W., Gui, J., Qu, P., Zhu, X., & Wang, L. (2015). Exploring similarity between academic paper and patent based on Latent Semantic Analysis and Vector Space Model. Paper presented at the Fuzzy Systems and Knowledge Discovery (FSKD), 2015 12th International Conference on.

10 NOTES ON CONTRIBUTORS

Farhan Ullah received his MS in Computer Science degree in 2012 from CECOS University Peshawar, Pakistan and BS in computer science degree in 2008 from University of Peshawar, Pakistan. He is currently pursuing Ph.D. in computer science degree from School of Computer Science, Sichuan University Chengdu, China. His research interests include information security and data science.
Email: farhankhan.cs@yahoo.com

Abdullah Bajahzar received PhD in Computer Science and Software Engineering degree from De Montfort University United Kingdom in 2014 and MSc Software Engineering degree from De Montfort University United Kingdom in 2008. He is currently working as an Assistant Professor at Almajmaah University Saudi Arabia. His research interest includes data mining, big data, IoT and data science.
Email: a.bajahzar@mu.edu.sa

Hamza Aldabbas received his PhD Degree in Computer Science and Software Engineering from De Montfort University United Kingdom in 2012 and M.Sc. Computer Science from Al-Balqa Applied University Jordan in 2009. He is currently working as an Assistant Professor at Al- Balqa Applied University Jordan. His research interests include Malware detection, IoT, machine learning and data science.
Email: Aldabbas@bau.edu.jo

Muhammad Farhan is working as Assistant Professor in the department of Computer Science, COMSATS University Islamabad, Sahiwal Campus, Pakistan. He has completed his Ph.D. in 2017 in the field of Computer Science from Department of Computer Sciences and Engineering in University of Engineering and Technology (UET), Pakistan. His interests include, Data Science, machine and deep learning and Internet of Things.
Email: farhansajid@gmail.com

Hamad Naeem has completed Ph.D. software engineering degree from college of computer science, Sichuan University, China, in 2019. He has published various articles in reputed SCIE and EI journals/conferences. His research interest includes malware detection, image processing, internet security, and machine learning.
Email: hamadnaeemh@yahoo.com

S. Sabahat H. Bukhari received MS in Computer Science from COMSATS Institute of Information Technology, Islamabad Pakistan in 2007. He is currently pursuing his Ph.D. degree in Software Engineering at Chongqing University, P.R. China. His research interests include cloud computing, edge computing and real-time systems.
Email: sabahatbukhari@cqu.edu.cn

Kaleem Razzaq Malik is working as Associate Professor in Department of Computer Science & Engineering, Air University, Multan Campus. He has completed Ph.D. Computer Science from University of Engineering and Technology, Lahore, Pakistan in 2011. His areas of research include but are not limited to Semantic Web, Data Modelling, Internet of Things, Big data, and Database
Email: krmalik@gmail.com