**RESEARCH PROPOSAL FOR MS (DS)**

**COMPUTER SCIENCE DEPARTMENT, FAST KARACHI**

Hybrid Approach for Autonomous Assessment of Domain Specific Comparative Questions

**STUDENT NAME - MUHAMMAD QASIM**

**SUPERVISOR - DR NOUMAN DURRANI**

**ABSTRACT**

In developing countries, due to shortage of resources, teachers are colossally over burden. Social problems, limitation of existing literature on English Language and limited approaches addressed in exiting literature for autonomous assessment of domain specific comparative questions are the key factors which required a novel and refined technique to satiate. Our proposed work addresses the problem of autonomous assessment by using hybrid approach of using Peer Assessment and Latent Semantic Indexing (LSI) alongside domain specific questions of computer science, Chemistry and Biology, thus it would solve the teachers work load in an appropriate and effective way. Evaluation of the work will be done through subjective and objective evaluation. Our proposed work will share the workload of teachers especially in developing countries like Pakistan where teachers are immensely overburden/ overtaxed with number of classes, number of quizzes, plethora of student’s projects and assignments.

1. **Chapter 1**

1.1 **Introduction**.

In developing countries, due to shortage of resources, teachers are colossally over burden. Number of students, number of classes, number of courses and number of quizzes are such factors upon which teachers are being substantially overtaxed. Even, Pakistan’s National Accreditation Council (PNAC) proposed that in each course, at least three quizzes are required to be taken. Thus, marking of all such quizzes has become a big challenge. Secondly, teachers in general do not follow traditional definitional questions and prefer comparatives questions. In current era, teachers are facing plethora of problems across the world especially in developing countries like Pakistan which warrants swift remedial actions to enhance the effectiveness of teachers subsequently. At present, the social problem which unnecessarily challenging teachers is that they are overburden with intensive workload such as large number of classes, quizzes alongside numerous exams/ quizzes marking simultaneously. Nexus above, all such tasks make a tough call for a teacher yielding lack of effectiveness. In developing countries like Pakistan, teachers do not get assistants in general whereas universities of developed countries like Stanford University offer upto eight assistants to streamline student’s projects alongside course content.

Another problem entailing flash addressing is that the existing literature mostly addresses English language only owing to the reason of use of dictionary and grammar. All work done in the past is mainly pertinent to essay evaluation which is based on English language; hence English language is mapped through dictionary which is not a valid phenomenon while handling domain specific question assessment e.g. Computer Science, since they have their own terms. Consequent upon which, it can be inferred that all the work done on essay assessment in the past has been negated thereby. However, domain specific assessment is sparsely addressed.

There is another problem which requires concurrent solution is that all present techniques are solely depending upon one approach that is either Term Frequency-Inverse Document Frequency (TF-IDF), Latent Semantic Indexing (LSI), Text classifications or some Graphs sans domain specific questions. Even if, some previous research work addressed domain specific questions assessment but they had either less accuracy or they only depended upon single technique. Notwithstanding the above, the proposed work addresses the problem of autonomous assessment by using hybrid approach of using Peer Assessment and Latent Semantic Indexing (LSI) alongside domain specific questions of computer science, Chemistry and Biology, thus it would solve the teachers work load in an appropriate and effective way. Peer Marking technique is a technique in which all quizzes attempted by students are evaluated systematically for that quiz to quiz comparisons be done in order to see the class overall level of understanding of content. Thus, highest similarity in the quiz would be graded accordingly. It is for the perspective to make sure the point that the whole class has understood the lecture/ content and addressed the question appropriately.

1.2. **Background/ Knowledge Gap**

Since, previous research works on assessment were formulated in such a way that domain specific questions were not solved and secondly comparative questions were never even addressed in any research work. Moreover, previous research work addressed only English language which was mapped on dictionary. Existing research has sufficiently contributed in solving assessment of English language. However, domain specific questions mainly addressed definitional questions rather than comparative questions. Even though, techniques for solving definitional and English language questions entail hybrid with the knowledge and mixed up approaches but their accuracy is questionable.

1.3. **Problem Identification**

In order to reduce the burden of teachers, there is a requirement of technique/ method that help out the teachers yielding autonomous marking of quizzes.

1.4. **Problem Statement**

There are three problems upon which our research will be working dynamically such as social problems, limitation of existing literature on English Language and limited approaches addressed in exiting literature.

1.5. **Objectives**.

There are mainly three major objectives of this proposal which are now being addressed. The first one is to propose a hybrid approach for autonomous assessment of domain specific comparatives questions. One of the objectives is to objectively evaluate the effectiveness of the proposed technique in the light of Accuracy, Precision, Recall and F-Measure respectively. Another objective of this proposal is to subjectively evaluate the effectiveness of the proposed technique in terms of correct marking. For that, we will hire a teacher who will conduct quizzes in his class. He will provide us his own solution alongwith notes/ helping material, consequent upon we will compute all quizzes systematically using LSI yielding a result sheet. Then we will again consult the teacher for marking the conducted quizzes for evaluation / comparison with the result sheet. That’s how we will be able to see how much accuracy we achieved consequently.

1.6. **Significance of Study**

Our proposed work will share the workload of teachers especially in developing countries like Pakistan who are immensely overburden/ overtaxed with number of classes, number of quizzes, plethora of student’s projects and assignments. Thus, our work will enable teachers to focus more on student’s grooming and use of latest tactics to enhance the performance and exposure of students.

1.7. **Limitations of Study**

Our work has a limitation that LSI does not handle negations such as “The person will” and “The person will not”. Semantically, these are two opposite action words; but conventional LSI, at present don’t differentiate it properly. Thus, this will be our future work, which could be sorted out by using string to string comparison and rearrangements of their weight-age as per context semantic meanings.

2. **Chapter 2**

2.1. **Literature Review**



3. **Chapter 3**

3.1 **Research Methodology and Model**

The methodology that we will adopt is to demonstrate a hybrid approach showing how to compare the quizzes with the solution and notes alongwith peer assessment simultaneously. For Quiz comparison, we may use, TF-IDF, Cosine Similarity, Latent Semantic Analysis (LSA) and Word2Vec technique depending upon best suitability and result accuracy. There are following modules which show our methodology thereby:-

3.1.1. **Data Set Formulation**

First part of any research is data collection. We will conduct real time quizzes from students and will also type their answers in e-format or we may use OCR for the purpose. At present, we are focusing on how assessment is being done efficiently but in future, we may be able to make electronic testing system for which students comes to prescribed labs and write down the quizzes in e-format. Detail of data sets is as under:-

* Quizzes will be conducted in real time from students.
* Solutions will be acquired from teacher.
* Notes will also be acquired from teacher which could be a book page or web page.

3.1.2. **Pre Processing**

At first, we will remove stop words because basic words like a, the, for, of etc do exist in the data which sans semantics or effects and also useless to keep. Secondly, we will do stemming because many words have the same meaning but exists in different forms e.g. keep or keeping, contribute or contribution; so stemming would consider one word for them. After that, we will do tokenization so that we be able to differentiate the text for vector space model and tf-idf etc.

3.1.3. **Assessment**

We may use tf-idf, Latent Semantic Indexing (LSI), Word2Vec, Cosine Similarity and Text Summarization techniques for assessment of quizzes. We may also drift towards machine learning as and when required.

3.1.3.1 **Frequency–Inverse Document Frequency (TF–IDF)**

TF–IDF is the most fundamental form of document representation and has the longest history among the three adopted representation methods (Robertson, 2004). It is based on the bag-of-word scheme in which a document can be represented by a collection of words used in the document. TF–IDF also assumes that if a word is important for a document, it should repeatedly appear in that document whereas it should rarely appear in other documents. The TF is associated with the former assumption whereas the IDF is associated with the latter assumption. The parameter tfij is defined as the number of times word i appears in document j; the larger the value, the more important the word is. The parameter dfi is the number of documents in which word ‘i’ appears at least once; the larger the value, the more common the word is. If word i can be considered important for document j, it should have a large TF (tfij) and a small DF (dfi). For example, articles such as ‘a’ and ‘the’ and pronouns such as ‘it,’ ‘this,’ ‘that,’ and ‘those’ would frequently appear in a document. However, they cannot be considered important words for that document because they are prevalent in all documents. Hence, TF–IDF is defined by Eq.(1):

TF–IDFij = tfij × log (N / dfi+1)

where the IDF takes the logarithm of the ratio of the number of total documents in the corpus to the document frequency of word i with IDF smoothing to prevent it from being divided by zero, i.e.,   
idfi = log. We select those words with high TF–IDF scores on an average for the corpus.

TF–IDF has been the most commonly adopted document representation method for various document-processing tasks. It provides each word in a document a weight according to the following two criteria: (1) the frequency of its usage in the specified document (TF) and (2) the rarity of its appearance in the other documents in the corpus (IDF) (Baharudin, Lee, & Khan, 2010). Zhang et al. (W. Zhang, Yoshida, & Tang, 2011) performed a comparative study of text classification using TF–IDF, latent semantic indexing (LSI), and multi-words for text representation. Ranjan et al. [4] used TF–IDF to assign weights to the words and classified documents using long short-term memory (LSTM) neural networks. Moreover, a few studies have used variants or extensions of TF–IDF as the main document representation method to perform classification tasks (Sabbah et al., 2017)  
 (Y. Zhang, Gong, & Wang, 2005).

3.1.3.2. **Doc2Vec**

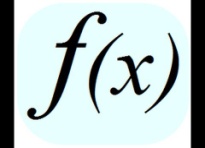
Recently, Doc2Vec has been successfully used for document classification tasks in various domains (Lau & Baldwin, 2016) (Pan, Xing, & Wang, 2016). Doc2Vec is a natural extension of Word2Vec, the main task of which is to determine an appropriate distributed representation for a single document by learning a neural network with the information of a target word and the words surrounding it in the document. Previous studies have demonstrated that Doc2Vec yields higher classification accuracy than other document representation methods in various domains, such as sentiment classification with the IMDB dataset (Lau & Baldwin, 2016), news categorization (Pan et al., 2016), and forum question duplication (Lau & Baldwin, 2016).

3.1.3.3. **Latent semantic analysis**

LSA is used due to its unique ability to uncover the conceptual content within unstructured data through a variety of mathematical dimension reduction techniques that estimate the linear combinations of the meaning of words and concepts (Kulkarni et al., 2014). It then uses that categorization method to allow subsequent in-depth analysis. LSA is a natural language processing method that extracts concepts from a sparse matrix of terms to produce an arrangement of terms that is bounded by the primary assumption that words similar in meaning will occur in analogous segments of text (Kulkarni et al., 2014; Deerwester et al., 1990). A collection of contexts, identified as either specific individual unique words or a collection of specific meaningful terms in the collected data, is extracted. A context can consist of unique terms and synonyms, as simple as “database” and “query” to describe “database applications” or a single multiple-word term such as “accounting information systems.” More complex contexts can be generated using a mixture of both single and multiple words such as “enterprise resource planning,” “ERP,” “implementation,” “plan,” and “system” to describe “ERP implementations.” Thus, in corollary with primary LSA assumptions, contexts with similar meaning will occur in similar meaning documents. Through a process analogous with traditional factor analysis, LSA uses singular value decomposition (SVD) to determine unique terms that represent the underlying concepts manifested within the data. A matrix containing counts of contexts per individual textual data segment (document, paragraph, or other designation of granularity) is constructed from the data. SVD then reduces this sparse matrix while preserving the similarity structure amongst columns representing each separate document. Like principal component analysis (PCA), SVD produces simultaneous principal components for two sets of variables, contexts (U) and documents (VT). Two separate sets of factor loadings, one for the U and VT matrices, are produced with each latent factor associated with both a set of corresponding high-loading terms and paired set of corresponding high-loading documents. These two sets of results can then be interpreted concurrently to develop the fundamental word usage and association patterns, which are termed Factors or Themes. Like traditional factor analysis, the researcher can indicate the number of factors within LSA to extract and therefore specify the level of granularity for theme extraction. The resulting LSA analysis approximates the relationship of a word or group of words (contexts from the U matrix), to the meaning of a specific passage of text, illustrated by the VT matrix, and vice versa. The relationship these words have on the passage is the interpretation of the derived associations between each individual word and word group utilized in the passage of text, and not the individual frequencies of the words in the given corpus. Due to similarities with traditional factor analysis, cross-loadings, and thus overlapping, of themes can occur.

3.1.4. **Formation of Marking Equation**

Since, quizzes are being judged for their similarity assessment module, so their marking within a given range of teacher require a well defined assessment function or equation. Our possible equation require mean of quiz to solution similarity score and quiz to notes similarity score alongwith peer assessment evolution. Finally, this equation would result a numeric number within a given range of teacher.

Assessment Function =  = X (Scalar value)

3.1.5. **Evaluation**

Since, we have got marked quizzes from previous module and their marking number is within range of teacher. Now, we will go back to teacher and get his manual marking on all quizzes There will be two types of evaluations which we will be doing in our research such as objectively and subjectively:-

3.1.5.1. **Objective Evaluation**

There will be two types of evaluations such as subjective evaluation and objective evaluation. In objective evaluation, we will compute Accuracy, Precision, Recall and F-Measure which will be done through automated method. There are four parameters upon which we would be defining respectively:-

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Predicted Class** | | |
| **Actual Class** |  | **Class = Yes** | **Class = No** |
| **Class = Yes** | True Positive (TP) | False Negative (FN) |
| **Class = No** | False Positive (FP) | True Negative (TN) |

Accuracy is said to be the most intuitive performance measurement and is a ratio of correctly predicted observation to the total observations. High Accuracy means best model we have.

Accuracy = TP + TN

TP + FP + FN + TN

Accuracy means that how close the generated score by the system is with respect to teacher’s score. Accuracy can be measured by RMSE equation which is as under:-

RMSE = sqrt 1/n ∑e2

Greater the RMSE value, lower the accuracy, so accuracy is:-

Accuracy = 100-RMSE

Precision is said to be the ratio of true positive with respect to total predicted positives:-

Precision = True Positive

True Positive + False Positive

Precision = Number of Quizzes Correctly Graded by System

Number of Graded Quizzes

Precision = True Positive

Total Predicted Positive

Recall is to said to be the ratio of true positive with respect to total actual positive:-

Recall = True Positive

True Positive + False Negative

Recall = True Positive

Total Actual Positive

Recall = Number of Quizzes Correctly Graded by System

Number of Quizzes graded by Human

F-Measure is required when there is a requirement to seek a balance between Precision and Recall:-

F1 Measure = 2x Precision x Recall

Precision + Recall

3.1.5.2. **Subjective Evaluation**

In subjective evaluation, we will involve expert who will design the quiz and also mark the quiz. Thus, we will do blind evaluation from the expert. After that, we will take feedback of the teacher as well. We will also find Error Rate by comparing his assessment and our system’s assessment. Then we will report the output. That’s how our research will be accomplished.

3.2. **Research Model**

Our research model is quantitative in which quizzes are transformed into a scalar value.

3.3. **Hypothesis Statement/ Experimental Scenarios**

Within given range, we will be able to correctly mark quizzes in a more accuracy as well as handling of domain specific questions.

3.4. **Sample**

|  |  |  |
| --- | --- | --- |
| **Computer Science** | **Biology** | **Chemistry** |
| For 2 Questions:  25 Quizzes | For 2 Questions:  20 Quizzes | For 2 Questions:  15 Questions |

3.5. **Performance Measures**

There are following performance measures on which out system’s performance will be evaluated:-

* We will report accuracy
* We will report Error Report
* We will report Precision, Recall and F-Measure

3.6. **Procedures or Experimental Scenarios**



4. **Timeline and Budget.**

Time period to complete this research is 3-6 months. Cost on the project including hiring experts of 3 domains for conduction of Quizzes would be around Rs. 50,000/-

5. **References**.

Baharudin, B., Lee, L. H., & Khan, K. (2010). A Review of Machine Learning Algorithms for Text-Documents Classification. *Journal of Advances in Information Technology*, *1*(1), 4–20. https://doi.org/10.4304/jait.1.1.4-20

Febrita, R. E., & Mahmudy, W. F. (2017). Pre-Processed Latent Semantic Analysis for Automatic Essay Grading. *Kursor*, (January 2016), 175. https://doi.org/10.28961/kursor.v8i4.110

Islam, M. (2010). Automated Essay Scoring Using Generalized. *Proceesings of 13th International Conference on Computer and Information Technology (ICCIT 2010)*. https://doi.org/978-1-4244-8494-2/10

Kim, Y. B., & Kim, Y. S. (2010). An autonomous assessment system based on combined latent semantic kernels. *Expert Systems with Applications*, *37*(4), 3219–3228. https://doi.org/10.1016/j.eswa.2009.09.058

Lau, J. H., & Baldwin, T. (2016). An Empirical Evaluation of doc2vec with Practical Insights into Document Embedding Generation, (2014), 78–86. https://doi.org/10.18653/v1/W16-1609

Liu, O. L., Rios, J. A., Heilman, M., Gerard, L., & Linn, M. C. (2016). Validation of automated scoring of science assessments. *Journal of Research in Science Teaching*, *53*(2), 215–233. https://doi.org/10.1002/tea.21299

Moharreri, K., Ha, M., & Nehm, R. H. (2014). EvoGrader: an online formative assessment tool for automatically evaluating written evolutionary explanations. *Evolution: Education and Outreach*, *7*(1), 15. https://doi.org/10.1186/s12052-014-0015-2

Mohler, M., & Mihalcea, R. (2009). Text-to-text semantic similarity for automatic short answer grading. *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics on - EACL ’09*, (April), 567–575. https://doi.org/10.3115/1609067.1609130

Pan, Y., Xing, C., & Wang, D. (2016). Document classification with spherical word vectors. *2015 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, APSIPA ASC 2015*, (December), 270–273. https://doi.org/10.1109/APSIPA.2015.7415518

Robertson, S. (2004). Understanding inverse document frequency: On theoretical arguments for IDF. *Journal of Documentation*, *60*(5), 503–520. https://doi.org/10.1108/00220410410560582

Sabbah, T., Selamat, A., Selamat, M. H., Al-Anzi, F. S., Viedma, E. H., Krejcar, O., & Fujita, H. (2017). Modified frequency-based term weighting schemes for text classification. *Applied Soft Computing Journal*, *58*(May), 193–206. https://doi.org/10.1016/j.asoc.2017.04.069

Zhang, W., Yoshida, T., & Tang, X. (2011). A comparative study of TF\*IDF, LSI and multi-words for text classification. *Expert Systems with Applications*, *38*(3), 2758–2765. https://doi.org/10.1016/j.eswa.2010.08.066

Zhang, Y., Gong, L., & Wang, Y. (2005). An improved TF-IDF approach for text classification. *Journal of Zhejiang University SCIENCE*, *6*(1), 49–55. https://doi.org/10.1631/jzus.2005.A0049

Muhammad Azeem Abbas, Shiza Hammad, Gwo-Jen Hwang, Sharifullah Khan & Syed Mushhad Mustuzhar Gilani (2020) An assistive environment for EAL academic writing using formulaic sequences classification, Interactive Learning Environments, DOI: [10.1080/10494820.2020.1789670](https://doi.org/10.1080/10494820.2020.1789670)

MuhammadAzeem Abbas, Gwo-Jen Hwang, Saheed Ajayi, Ghulam Mustafa, Muhammad Bilal,

Modelling and exploiting taxonomic knowledge for developing mobile learning systems to enhance children’s structural and functional categorization,Computers and Education: Artificial Intelligence,Volume 2,2021,100007,ISSN 2666-920X, https://doi.org/10.1016/j.caeai.2021.100007.(https://www.sciencedirect.com/science/article/pii/S2666920X21000011)