SJIF Impact Factor 2021: 7.13| ISI I.F.Value:1.241| Journal DOI: 10.36713/epra2016 ISSN: 2455-7838(Online)
EPRA International Journal of Research and Development (IJRD)
Volume: 6 | Issue: 2 | February 2021 - Peer Reviewed Journal

A REVIEW OF MACHINE LEARNING TECHNIQUES FOR SUBJECTIVE ANSWERS EVALUATION WITH ONTOLOGY

Sourabh Singh, Ramdeo Yadav, Mayank Kumar Singh

ABSTRACT

The current system of evaluating subjective answers is adverse. As it takes a lot of time for the answers evaluation and it becomes boring also to check the answers of the same question multiple times. Currently, it can be seen that the use of Computer Aided Assessment(CAA) in checking objective answers is used for Objective Answers evaluation, but checking descriptive answers with CAA is a challenging task because the main aim of subjective evaluation is to get insight of student's learning and their knowledge enhancement. The challenges involve getting the semantic meaning of the language written by students, understanding the natural language of humans, assessing appropriately the knowledge obtained by students. Researchers have performed many ML and NLP techniques to apply CAA on descriptive answers. Among those techniques, hybrid techniques (combination of LSA and GLSA with fuzzy logic) have given better results than the rest of techniques. By using java programming language, open source libraries in java, MatLab and Wordnet, the hybrid technique is implemented. Along with this, there are few algorithms included in the preprocessing steps like word tokenization, stop words removal, synonym search and stemming. In 2016, Dr. Himani Mittal et. al has experimented these techniques with and without using Ontology. And they came to a conclusion that with the use of Ontology few algorithms performance improved extremely well but many have given consistent results.

KEYWORDS: ***Latent Semantic Analysis, Ontology, Subjective Evaluation, Bilingual Evaluation Understudy, Cosine Similarity, Maximum Entropy, WordNet***

1. INTRODUCTION

Subjective Evaluation of answers is an important method for understanding the learnings of students and their knowledge enhancement. The manual evaluation of subjective answers takes a lot of time, sometimes results are also delayed, etc. In this 21st century, technology has progressed very fast. At the time lockdown, many organizations have opted to work from home, even schools and colleges are taking their classes online and examinations too. But, the process of evaluating the answers is still manual. So there must be some intelligent software that can solve these problems efficiently and also give accurate results when compared to manual evaluation.

There are several Machine Learning (ML) and Natural Language Processing (NLP) techniques that are able to find the semantic similarity between two words, sentences, or paragraphs. Among those techniques, few of them are explored in this paper like Latent Semantic Analysis (LSA), Generalized Latent Semantic Analysis (GLSA), Bilingual Evaluation Understudy (BLEU), Maximum Entropy and Hybrid Technique. The outputs of all these techniques vary between range 0 to 1. If the output is 1, it means the student answer and the model answer are highly similar and if output is 0 then there is no similarity.

Computerised Evaluation of Answers is not the new concept. Researchers have been experimenting with various algorithms to overcome this problem for decades. Project Essay Grader [23] was developed by Ellis Page in 1994. But he focused on the surface structure more and ignored the semantic aspect of the essay because of which it was criticized at that time. Thomas K. Landauer et.al [1] developed Latent Semantic Analysis- the technique that is mostly used in automated assessment. There are also many Automated Essay Scoring(AES) widely used -- Intelligent Essay Assessor(IEA) [5], E-Rater and Criterion [15], C-Rater [14], IntelliMetric and MyAccess and Bayesian Essay Test Scoring System(BETSY) [12]. In 2016, M.S Devi and Himani Mittal,[3] in their

SJIF Impact Factor 2021: 7.13| ISI I.F.Value:1.241| Journal DOI: 10.36713/epra2016 ISSN: 2455-7838(Online)
EPRA International Journal of Research and Development (IJRD)
Volume: 6 | Issue: 2 | February 2021 - Peer Reviewed Journal

experiment they applied Latent Semantic Analysis(LSA), Generalized LSA, Bilingual Evaluation Understudy and Maximum Entropy on subjective answers with and without Ontology, they concluded that the results are more accurate with the use of Ontology.

The challenging problem that we are going to face in this project is to compare the answers of users with that of the stored one. Because, everytime we do not answer the question in the same manner. Suppose, if the question is "Where were you last night?". The answer stored is " Cinema Hall". Next time, when the application asked him the same question and user answered that "I was watching a movie in the theatre". We can see that the both answers are different but we understand it, both are semantically correct. Our main goal in this project is that even our application will also be able to mark these types of answers as the correct one.

In this paper, there are a total of 6 sections including this introduction. In the second section, we have explained about some important previous words that got some recognition. In the third section, the general methodology has been explained based on reading all the research papers on Subjective Evaluation. In the fourth section, we have gone through the working plan briefly. In the fifth section, after testing all the techniques with and without Ontology, the results were discussed and in the fifth section, finally given a conclusion about the best techniques and some discussion related to works that can be done in future.

2. REVIEW OF RELATED WORK

In 1994, a tool was developed called Project Essay Grade [23] that evaluates the English Essay. The accuracy of output of this tool was 83-87%. However, this tool was only to check the similarity of words. It doesn't consider the semantic meaning of the content. It was only checking on the basis of word length, word similarity, etc.

In 1999, Foltz et al. [9] developed another tool for assessing the english essay by applying a mathematical technique called latent Semantic Analysis. The name of the tool was Intelligent Essay Assessor. In Latent Semantic Analysis, a matrix is constructed that is called term-document frequency (tdf) matrix. Then Singular Value Decomposition is applied on the tdf matrix. Correlation was calculated between output of LSA and human-assigned grades and the value varied from 0.59 to 0.89 whereas correlation value between two human graders varies from 0.64 to 0.84. So the performance of LSA and human graders is comparable.

In 2005, there came an author Diana Perez [15] who developed a system using LSA and BLEU (Bilingual Evaluation Understudy) technique. The output of these two techniques was combined by a linear equation. The success rate was 50%. This work was further extended by Himani Mittal in 2016, where she combined outputs using fuzzy logic and success rate improved from 50% to nearly 85%.

In 2010, Islam and Hoque [14], they extended the work of Foltz. and introduced a technique called Generalised Latent Semantic Analysis for evaluation. In LSA, we are considering only single but in the GLSA group of words are considered for matrix construction. The accuracy of the results varied from 89% to 96%.

In 2016, Himani Mittal et al. [3] introduced Hybrid Technique. This technique was developed by combining LSA and BLEU with the help of fuzzy logic. The drawback of BLEU is that it neither performs semantic analysis nor measures the grammatical structure. The drawback of LSA is that it doesn't consider the syntactic structure of answers but measures the semantic aspect thoroughly. So the best features of LSA and BLEU are combined and syntactic structure and word similarity are taken care of by the use of WordNet tools.

In 2018, Vaibhav Miniyar et al. [5] developed an android application for Students Marks Evaluation. The Machine Learning and Natural Language Processing techniques they have used have a high agreement (up to 90%) with human performance. In their system, they have used Naive Bayes Classifier and Cosine Similarity.

3. GENERAL METHODOLOGY USED FOR EVALUATION

The general approaches that are applied for subjective evaluation are shown in fig 1.

Preprocessing steps are done first on the input answers. And in the pre-processing steps, tokenizations are performed to get individual words. Then stop words (a, an, the, as, etc.) are removed that are common in every sentence. After these steps, synonym search is performed on the rest of the words. Later one more step is performed called word stemming. Stemming algorithms like Porter's Algorithm reduce the words like argument, argues, arguing, etc. to argu i.e. stem words need not to be a word. Finally, Machine Learning and Natural Language Processing techniques like LSA, GLSA, MaxEnt, Cosine Similarity, Bilingual Evaluation Understudy(BLEU), etc. are applied along with the input of keywords provided as a model answer.

Latent Semantic Analysis (LSA) is a technique in natural language processing for the analysis of the relationship between a set of documents and their words. LSA is sometimes referred to as Latent Semantic Indexing(LSI) because of its application to information retrieval.

SJIF Impact Factor 2021: 7.13| ISI I.F.Value:1.241| Journal DOI: 10.36713/epra2016 ISSN: 2455-7838(Online)
EPRA International Journal of Research and Development (IJRD)
Volume: 6 | Issue: 2 | February 2021 - Peer Reviewed Journal

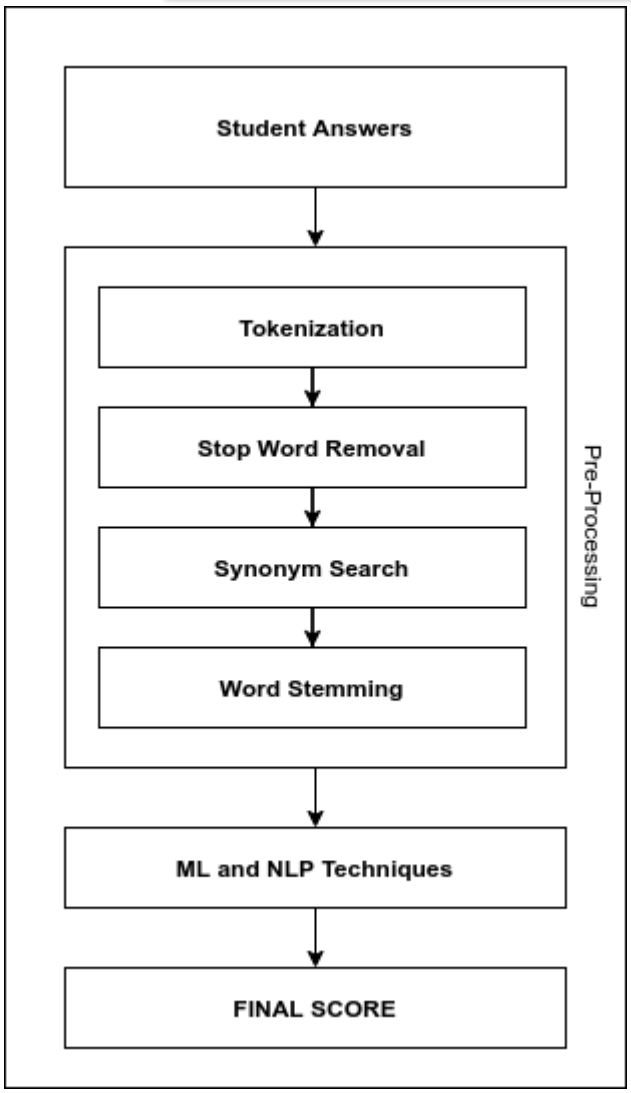

Fig. 1: General Approach of Subjective Evaluation Using ML and NLP Techniques

4. WORKING PLAN

This system can be widely used not only by students but anyone who wants to memorise something in question-answer form, they can use this as a reminder also. This can be implemented in different fields like for instant revision during examination, reminder for important works or meeting, etc. So, the techniques that we are going to use in this project will perform tasks like Tokenizing words and sentences, parts of speech tagging, chunking, chinking, lemmatizing words and wordnetting to evaluate the subjective answer.

Fig. 2-Workflow diagram

Users will give answers to the questions asked by application either in text form i.e. through keyboard or by speaking. After getting the answer an editing screen will be opened to correct if the speech-to-text API was able to detect their words correctly or not. After submitting, it will go for preprocessing of the text, i.e NLP techniques will be applied on them to compare the user's current answer with that of the model answer that was previously stored by the user in the database of the application. The model answers will then be trained by providing keywords and questions specific things(QST). In the user's answer, the algorithm will search for the presence of keywords . Grammar will be checked by an api as shown in table-1.

Table-1. Internalization of Grammar

Grammar Values	Numeric Values
Proper	1
Improper	0

Table-2. Comparative Study of the techniques for subjective evaluation of answers.

Author and Year	Tool	Techniques	Results	References
Landauer et al., 2003	Intelligent Essay Assessor	Latent Semantic Processing	59-88%	[1], [4]-[7]
Kakkonen et al., 2008	Automatic Essay Assessor	LSA, Probabilistic LSA, Latent Dirichlet Allocation	LSA better than the rest	[8], [9]
Islam et al, 2010		GLSA	80%	[11]
Rudner et al, 2002	Betsy	Bayes Theorem	80%	[12]

SJIF Impact Factor 2021: 7.13| ISI I.F.Value:1.241| Journal DOI: 10.36713/epra2016 ISSN: 2455-7838(Online)
EPRA International Journal of Research and Development (IJRD)
Volume: 6 | Issue: 2 | February 2021 - Peer Reviewed Journal

Libin et al., 2008		K-Nearest Neighbor	76%	[13]
Sukkarieh et al., 2012	C-Rater	Maximum Entropy	80%	[14]-[16]
Mittal et al., 2016		Hybrid Technique (combining LSA and BLEU)	72-99%	[3]
Miniyar et al., 2018	Student's Marks Evaluation Application	Naive Bayes, Cosine Similarity	up to 90%	[5]

The research is ongoing for many decades to find a solution for this problem i.e, subjective evaluation. Several machine learning as well as natural language processing techniques were applied to subjective answer evaluation. Table-2 contains comparative study between few techniques used in different tools by different authors.

LSA technique, proposed by Deerwester, is used to establish similarity between two contents. IEA means Intelligent Essay Evaluator was used in the TOEFL exam and accuracy of the results varied from 59 to 87 percent. Similarly, a tool was developed by Diana Perez called Atenea, using a hybrid of LSA and Bilingual Evaluation Understudy. Another was developed called Electronic Essay Rater (E-Rater) uses Natural Language Processing techniques to evaluate sentence structure. It was also used in an exam, GMAT and it has the accuracy of 84 to 93 percent.

5. FINDINGS

To test all the subjective evaluation techniques, there is no standard database available. Therefore, by conducting a class test among 50 students is performed and all the answer sheets were first evaluated by human beings. Later, it was applied to all the techniques with and without Ontology. After finding correlations of performance of different techniques with humans, they have concluded that Maximum Entropy performance was improved extremely well and rest of techniques were giving consistent results.

6. CONCLUSIONS

What is an Onscreen Marking System? To evaluate the physical copies of the answer sheets in digital form, the Onscreen Marking System is useful. It helps remove location and physical answer sheet handling constraint for the examiners, moderators and result processing authority. Digital evaluation tools like digital annotations, assigning marks, total calculation, moderation is simplified and can be completed in quick time. The figure is well understood.Conclusion The Technique LSA and GLSA discussed and implemented in this project is having good accuracy more than 90% . This project works the same as human beings work considering for evaluation of answer sheets. The accuracy can be increased as per the data set collected. This system can be improved by taking continuous feedback from teachers and students. The Bilingual Evaluation Understudy technique worlds like if the number of keywords is less in the student's answers then it will give lower marks to it. So after applying the BLEU technique, the score that it generates can be considered as the minimum marks to be awarded to students. The technique BLEU i.e Bilingual Evaluation Understudy doesn't measure the semantic analysis of the sentences and also grammatical structures and expressions of the sentences. It is only checking exact word matches and dividing it with the total number of keywords. We get the best features of LSA and BLEU by combining them, also we say it is a hybrid technique. Using wordnet tools, we are finding the similarity between words and sentences and finding the syntactic structure of the sentences.. We can conclude that in the lower bound we can put BLEU technique and in Upper bound we can put the LSA technique. Using fuzzy logic, this technique actually helps to combine two scores.

REFERENCES

1. *Devi, M. S., & Mittal, H. (2016). Machine learning techniques with ontology for subjective answer evaluation. arXiv preprint arXiv:1605.02442.*
2. *Patil, P., Patil, S., Miniyar, V., & Bandal, A. (2018). Subjective Answer Evaluation Using Machine Learning. International Journal of Pure and Applied Mathematics, 118(24).*
3. *Guruji, P. A., Pagnis, M. M., Pawar, S. M., & Kulkarni, P. J. (2015). Evaluation of Subjective answers using GLSA enhanced with contextual synonymy. International Journal on Natural Language Computing (IJNLC), 4(1).*
4. *Mittal, H., & Devi, M. S. (2016). Computerized evaluation of subjective answers using hybrid technique. In Innovations in Computer Science and Engineering (pp. 295-303). Springer, Singapore.*
5. *Mihalcea, R., Corley, C., & Strapparava, C. (2006, July). Corpus-based and knowledge-based measures of text semantic similarity. In Aaai (Vol. 6, No. 2006, pp. 775-780).*
6. *Landauer, T. K. (2003). Automatic essay assessment. Assessment in education: Principles, policy & practice, 10(3), 295-308.*
7. *Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. Discourse processes, 25(2-3), 259-284.*
8. *P. W. Foltz, W. Kintsch, and T. K. Landauer, "The measurement of textual coherence with latent semantic*

SJIF Impact Factor 2021: 7.13| ISI I.F.Value:1.241| Journal DOI: 10.36713/epra2016 ISSN: 2455-7838(Online)
EPRA International Journal of Research and Development (IJRD)
Volume: 6 | Issue: 2 | February 2021 - Peer Reviewed Journal

analysis," Discourse Process., vol. 25, no. 2–3, pp. 285–307, 1998.

9. *T. Kakkonen, N. Myller, E. Sutinen, and J. Timonen, "Comparison of dimension reduction methods for automated essay grading," Educ. Technol. Soc., vol. 11, no. 3, pp. 275–288, 2008.*
10. *Hofmann, T. (1999, August). Probabilistic latent semantic indexing. In Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval (pp. 50-57).*
11. *Islam, M. M., & Hoque, A. L. (2010, December). Automated essay scoring using generalized latent semantic analysis. In 2010 13th International Conference on Computer and Information Technology (ICCIT) (pp. 358-363). IEEE.*
12. *Rudner, L. M., & Liang, T. (2002). Automated essay scoring using Bayes' theorem. The Journal of Technology, Learning and Assessment, 1(2).*
13. *Bin, L., Jun, L., Jian-Min, Y., & Qiao-Ming, Z. (2008, December). Automated essay scoring using the KNN algorithm. In 2008 International Conference on Computer Science and Software Engineering (Vol. 1, pp. 735-738). IEEE.*
14. *Miller, G. A. (1995). WordNet: a lexical database for English. Communications of the ACM, 38(11), 39-41.*