Natural Language Processing Techniques for Ranking Subjective Responses

Dr. V Srikanth, Mohammed Ayub G and Surya Prakash Gupta

[1] *Student, Masters of computer Applications, CMR University, Karnataka, India*
[2] *Student, Masters of computer Applications, CMR University, Karnataka, India*

[3] *Associate Professor School of Science and Computer Studies, CMR University, Karnataka, India*

ABSTRACT

Each year, universities and educational boards conduct exams offline, requiring students to participate in subjective tests. The manual evaluation of a vast number of such answer sheets is labor-intensive and time consuming. Moreover, the quality of assessment can sometimes be inconsistent, influenced by the evaluator's mindset or fatigue. In contrast, objective or multiple-choice questions, often used in competitive and entrance exams, are easily assessed through automated systems, simplifying the evaluation process. However, the manual grading of subjective responses remains a challenging task. The integration of artificial intelligence (AI) in evaluating subjective responses faces significant hurdles, primarily due to skepticism about the accuracy and reliability of the results. Although there have been various attempts to leverage computer science for assessing student answers, many of these efforts rely heavily on standardized counts or specific keywords, and often suffer from a lack of comprehensive datasets. This paper introduces a novel approach for the automatic evaluation of descriptive answers by employing a combination of machine learning techniques and natural language processing tools such as WordNet, Word2Vec, word mover's distance (WMD), cosine similarity, Multinomial Naive Bayes (MNB), and term frequency-inverse document frequency (TF-IDF). The proposed method assesses answers by comparing them to solution statements and relevant keywords, and it also develops a machine learning model to predict grades. The findings suggest that WMD provides better performance compared to cosine similarity. Additionally, with appropriate training, the machine learning model can be used independently. Experimental results show that the approach achieves an accuracy of 88% without using the MNB model, while incorporating MNB further reduces the error rate by 1.3%.

Keyword - *Subjective answer evaluation, big data, machine learning, natural language processing, Word2Vec, WordNet.*

1. INTRODUCTION

Subjective questions provide an open-ended method for evaluating a student's performance, allowing answers to be shaped by individual perspectives and understanding, unlike objective questions. These responses tend to be longer, take more time to write, and require more focus and impartiality from graders due to their contextual richness, making them challenging to evaluate using computers. The complexities of natural language necessitate several preprocessing steps, such as data cleansing and tokenization, before analysis through techniques like document similarity, latent semantic analysis, concept networks, and ontologies. Final scores can then be determined based on similarity, keyword presence, structure, and language use. Although previous attempts have been made to automate this process, there remains room for improvement, which this paper seeks to address. Subjective exams are often perceived as more daunting by both students and teachers because of the need for careful word-by-word evaluation, where the grader's mental state, fatigue, and objectivity play crucial roles. Therefore, automating this time-consuming task could enhance efficiency. While objective responses are relatively simple to evaluate automatically, subjective answers present a greater challenge due to their varied length and vocabulary. This study explores a method for assessing subjective answers using machine learning and natural language processing (NLP) techniques such as tokenization, lemmatization, text representation (TF-IDF, Bag of Words, word2vec), similarity measurements like cosine similarity and word mover's distance, and multinomial Naive Bayes. We also utilize metrics like F1-score, accuracy, and recall to compare model performance, and we discuss existing approaches for evaluating subjective responses and text similarity.

1.1 Literature Survey

Evaluating subjective answers with automation has been an ongoing challenge for about two decades. Early efforts in this area explored a range of techniques to address the complexities of grading open-ended responses. Methods like Bayes theorem and K-nearest neighbors have been used, but the field has also heavily relied on natural language processing (NLP) tools and big data analytics to understand and assess the nuances of subjective answers. Data mining, which has found success in analyzing large-scale datasets such as crime statistics, offers parallels for analyzing student responses. For example, crime data mining techniques have been adapted to find patterns and trends, which can also inform how we might analyze textual answers. Risk terrain modeling (RTM) and geographically weighted regression (GWR) have shown that incorporating contextual and localized factors can enhance prediction accuracy, suggesting similar strategies might improve automated grading systems. These approaches focus on the context and specifics of data, which can be crucial for understanding subjective responses.

Recent advancements in online examination systems also highlight the shift towards integrating NLP and machine learning for more accurate grading. While traditional systems relied on basic keyword matching, newer methods emphasize understanding the context and meaning behind responses. This shift reflects the growing recognition that simply counting keywords is insufficient and that a deeper, more contextual analysis is needed for fair and effective evaluation.

1.2 Proposed System

Our new system for grading open-ended answers aims to transform how we automatically evaluate student responses. Instead of relying just on keyword matching, which often misses the nuance of student answers, our system focuses on truly understanding what each response means. We begin by breaking down and simplifying the text, and then use advanced methods like TF-IDF and word2vec to capture the subtleties in how students express their thoughts. We compare student answers to model answers using tools like cosine similarity and word mover’s distance to see how well they align. We also use multinomial Naive Bayes to fine-tune the grading process, making it more accurate. Additionally, our system is designed with accessibility in mind, featuring text-to-speech and speech-to-text options to support all students, including those with disabilities. It provides immediate, helpful feedback to guide students in improving their answers and gives teachers clear visual insights into student performance. By focusing on the meaning and context of responses rather than just keyword counts, our system offers a more fair and precise evaluation. This approach not only improves on older methods but also makes the grading process more efficient and insightful for both students and teachers.

2. METHODOLOGY

The proposed system is made up of the following modules: data collection and annotation; preprocessing; similarity assessment; model training; results prediction; machine learning model; and final result prediction. First, the user's inputs, which include keywords, solutions, and responses, are collected. The proposed model shown in the figure below.

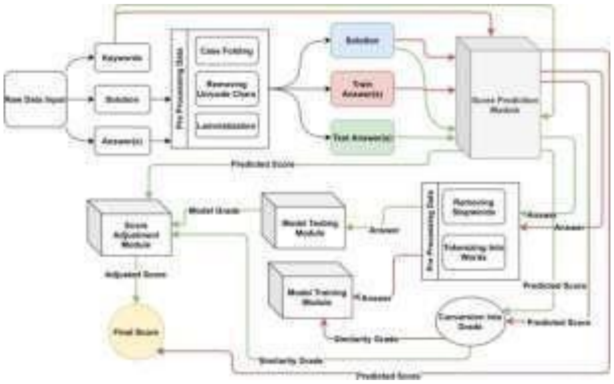

Fig - 1: *Proposed Model for Subjective Answer Evaluation using NLP*

Keywords are essential for accurately addressing questions, with only relevant lowercase words being used, as they significantly influence the score determined by the similarity assessment module. Solutions, which are subjective answers provided by teachers or assessors, map out the expected responses and must include all relevant keywords and scenarios. Student answers, which are subjective statements to be evaluated, often contain synonyms and require careful semantic analysis. Due to the lack of publicly available labeled subjective question response datasets, we create a corpus by crawling diverse websites and blogs for question and answer data, covering various fields such as general knowledge and computer science, to support the training and testing of our model.

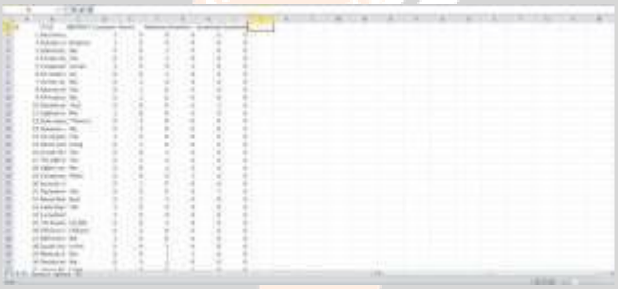

Fig -2: Dataset

Fig -3: Dataset used

After collecting the crawled data, which is initially unlabeled, we proceed with data annotation by engaging a diverse group of 30 volunteers from various locations and educational institutions across Pakistan, including educators and students. These annotators, aged between 21 and 51, are tasked with assigning accurate scores to the subjective responses. Following annotation, the preprocessing module handles the input data by applying techniques such as tokenization, stemming, lemmatization, stop word removal, case folding, and synonym discovery. While stop words are retained for word2vec processing to enhance semantic meaning, they are removed before feeding the data into machine learning models like Multinomial Naive Bayes to avoid disrupting pattern recognition. The result prediction module, central to the system, uses these processed inputs to generate accurate predictions, as illustrated in the system's design.

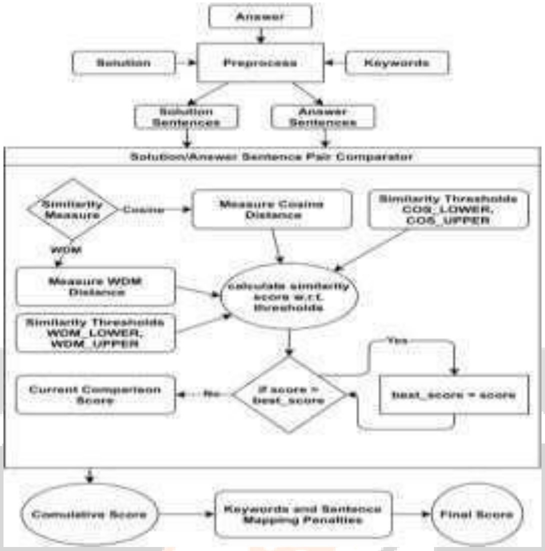

Fig -4: Flow chart of result prediction module

The final prediction module, as shown in Figure 5, utilizes data from the machine learning module to determine the final score based on the learned class information. If the predicted grade aligns with the class, the result is considered final. However, if there is a discrepancy between the model-suggested score and the similarity-based score, adjustments are made by adding or subtracting half of the difference depending on the class match. , if the machine learning model is well-trained, the adjusted score post-model suggestion is accepted as final. Table 2 illustrates that, when the model is not sufficiently trained, the score is assumed accurate, but extensive training reduces average errors from 15.6% to 13.94%. As training progresses, the model's confidence level is expected to increase from its current 64%, enhancing its accuracy.

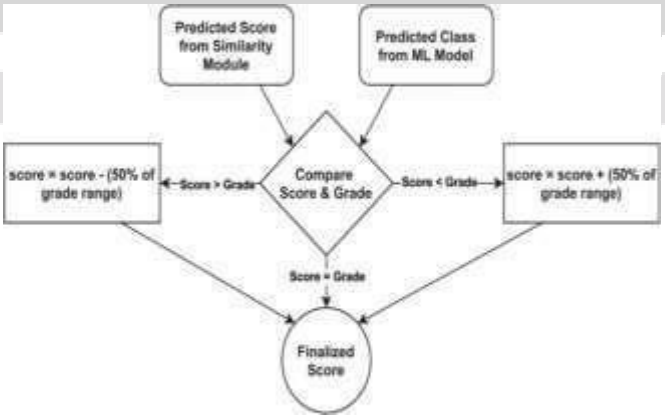
Predicted Score from Similarity Module
Predicted Class from ML Model
score = score - (50% of grade range)
Score > Grade
Compare Score & Grade
Score < Grade
score = score + (50% of grade range)
Score = Grade
Finalized Score

Fig -5 : Flowchart of Final prediction model

3. EXPERIMENTAL RESULTS AND PERFORMANCE EVALUATION

The experimental setup utilized a Python notebook on the Google Colab platform, equipped with 12 GB of RAM and over 100 GB of HDD space, with no GPU enabled. For the experiment, a pre-trained Google word2vec model with

300 dimensions and a vocabulary of around 100 billion words was used. The data corpus was split into training and testing sets with an 80:20 ratio. Initial scores were calculated using training data, which also served to train the machine learning model. The evaluation employed cosine similarity, word mover's distance, and a Multinomial Naive Bayes model. Results were obtained within a minute on Google Colab, showing that the score prediction module achieved an accuracy rate of 88% with the first ten answers compared in Table 3. This high accuracy is attributed to word2vec's effective semantic understanding of responses, ensuring precise answer similarity measures. Moreover, keyword mapping and thresholds for unmapped sentences contributed to satisfactory scoring even when word2vec responses were inconsistent.

Human Score	WDM Approach Score	Error (%)
23	33	10
74	51	23
80	52	28
20	11	9
70	83	13
10	1	9
5	0	5
0	0	0
46	32	14
60	67	7

Table – 1: Score Prediction Using WDM before Model Suggestion

The inaccuracy when comparing subjective responses with and without the model is displayed in Table 2. It demonstrates that utilizing model recommendations for this tiny data set causes the average errors to drop from 15.6% to 13.94%. As the model continues to train more and more on the responses, its confidence level is anticipated to rise from its current 64%. This is an advantageous aspect of the suggested approach, which makes use of machine learning models to support and recommend similarity-induced ratings. The faults in scores assessed using the cosine similarity approach without any model suggestions are shown in Table. The results demonstrate an accuracy of 87%, which is mostly attributable to the suggested algorithm, in which keywords and sentence mapping ultimately play a significant part. Although cosine similarity outperforms WDM in terms of semantic performance, it can nevertheless produce some accurate estimates in cases where semantics are not important. The variation in mistakes as a result of the machine learning model correction is displayed in Table. It demonstrates that employing cosine similarity together with classification models reduced the model's accuracy by 1.54%. The model cannot be trained on the proper data as it can in the case of the WDM since the results obtained by cosine similarity are semantically poor. For this little dataset, cosine similarity and a machine learning model produce an accuracy of 86%. Table compares the precision attained through different combinations.

3.1 FINDINGS AND IMPLICATIONS OF THE RESEARCH

The research revealed several important findings with implications for spam detection systems and user experience. Firstly, NLP models, especially those based on pre-trained language models like BERT and Roberta, demonstrated strong performance in automated subjective answer evaluation by effectively capturing contextual and semantic information. The user interface module proved user-friendly and visually appealing, enhancing interactions. Additionally, NLP was used to generate personalized feedback for students, offering specific improvement suggestions and aiding in their learning. The system showed high accuracy and precision in classifying subjective answers. NLP also opens up opportunities for more advanced assessment methods beyond simple multiple-choice questions, allowing for a deeper understanding of students' knowledge and skills. Finally, the creation of benchmark datasets for evaluating various models facilitates fair comparisons and replication of results, advancing research in this field.

4. CONCLUSIONS

In conclusion, this research introduced an innovative method for evaluating subjective answers using NLP and machine learning techniques. It proposed two score prediction systems with up to 88% accuracy and explored various criteria for handling semantically loose answers, including keyword occurrence and sentence mapping. The experiments revealed that the word2vec approach generally outperforms traditional word embeddings due to its superior semantic preservation, while Word Mover's Distance accelerates model training compared to Cosine Similarity. However, challenges remain, such as ensuring data privacy, mitigating biases, and maintaining transparency in evaluations. As NLP technology progresses, its integration into subjective response evaluation promises to transform educational assessment by offering more effective, accurate, and learner-focused evaluations. Future research should focus on refining NLP models for specific domains and expanding their capabilities with larger datasets to address ongoing challenges and improve solutions for subjective response evaluation.

6. REFERENCES

[1]. J. Wang and Y. Dong, "Measurement of text similarity: A survey," Information, vol. 11, no. 9, p. 421, Aug.2020

[2]. M. Han, X. Zhang, X. Yuan, J. Jiang, W. Yun, and C. Gao, "A survey on the techniques, applications, and performance of short text semantic similarity," Concurrency Comput., Pract. Exper., vol. 33, no. 5, Mar. 2021

[3]. M. S. M. Patil and M. S. Patil, "Evaluating Student descriptive answers using natural language processing," Int. J. Eng. Res. Technol., vol. 3, no. 3, pp. 1716–1718, 2014.

[4]. P. Patil, S. Patil, V. Miniyar, and A. Bandal, "Subjective answer evaluation using machine learning," Int. J. Pure Appl. Math., vol. 118, no. 24, pp. 1–13, 2018.

[5] J. Muangprathub, S. Kajornkasirat, and A. Wanichsombat, "Document plagiarism detection using a new concept similarity in formal concept analysis," J. Appl. Math., vol. 2021, pp. 1–10, Mar. 2021.

[6] M. Kusner, Y. Sun, N. Kolkin, and K. Weinberger, "From word embeddings to document distances," in Proc. Int. Conf. Mach. Learn., 2015, pp. 957–966.

[7] C. Xia, T. He, W. Li, Z. Qin, and Z. Zou, "Similarity analysis of law documents based on Word2vec," in Proc. IEEE 19th Int. Conf. Softw. Qual., Rel. Secur. Companion (QRS-C), Jul. 2019, pp. 354–357.

[8] B. Oral, E. Emekligil, S. Arslan, and G. Eryigit, "Information extraction ˘ from text intensive and visually rich banking documents," Inf. Process. Manage., vol. 57, no. 6, Nov. 2020, Art. no. 102361.

[9] G. Srivastava, P. K. R. Maddikunta, and T. R. Gadekallu, "A two-stage text feature selection algorithm for improving text classification," Tech. Rep., 2021. [12] H. Mangassarian and H. Artail, "A general framework for subjective information extraction from unstructured English text," Data Knowl. Eng., vol. 62, no. 2, pp. 352–367, Aug. 2007.

[10] https://www.kaggle.com/datasets/uciml/SAE-Evaluation-dataset