M. Mohanedova *et al*, International Journal of Advances in Agricultural Science & Technology,
Vol.11 Issue.4, April 2024, pg. 51-57

ISSN: 2348-1358

Impact Factor: 6.901

NAAS Rating: 3.77

Enhancing Educational Assessment Efficiency: An AI-based Approach to Automated Examination Evaluation

Dr.M.Radhika Mani [1] , Nirogi Venkata Naga Sai Vardhan [2], Lavanya Kota [3], Lanka Vatapatra Sai Pathrudu [4], Batchu Yuvaraju [5], Panchandhi Kasi Visweswararao [6] , [1]Professor , [2, 3, 4, 5, 6]B.tech Students Department of Computer Science Engineering, Pragati Engineering College, Surampalem , Andhra Pradesh, India
Email: drradhikamani@gmail.com

ss

Abstract:

Manually grading subjective papers is a difficult and time-consuming task. Lack of comprehension and acceptance of the results is one of the biggest obstacles to employing artificial intelligence (AI) to analyse subjective articles. There have been several attempts to grade responses from pupils using computer science. However, most of the work uses specific words or traditional counts to achieve this. Furthermore, verified data sets are not enough. Using a range of natural language processing, machine learning, and toolkits, including Wordnet, Word2vec, word mover's distance (WMD), cosine comparison, multinomial naive bayes (MNB), and others, this study presents a novel approach for automatically analysing descriptive responses and TF-IDF, or term frequency-inverse document frequency. Answers are evaluated using keywords and solution statements, and grades are predicted using a machine learning algorithm. Overall, the findings show that WMD and cosine are more comparable. The machine learning system may be used independently once it has undergone the required training. Trial and error results in an 88% accuracy when there is no MNB model. Using MNB reduces the inaccurate prediction rate by 1.3%.

Keywords: Large data sets, machine learning, word2vec, natural language processing, and subjective response evaluation.

I. Introduction

It may be possible to evaluate an individual's results and abilities by using open-ended, personal inquiries and replies. There are no restrictions on responses, so learners can feel free to write whatever that best expresses their understanding of the subject matter and point of view. Subjective and objective responses differ in a few more important ways, though. First of all, they are far lengthier than the objective questions. Secondly, writing them requires more time. They also have a lot of past information, so the teacher evaluating them must pay great attention to them and evaluate them impartially.

Because spoken language can be ambiguous, evaluating requests like this with computers can be challenging. A number of processing actions, including cleanup and tokenization, which is must be completed before working with the data. Subsequently, the textual data can be compared via concepts graphs, ontologies, latent structural structures, and document similarity, among other methods. Similarity, the existence of keywords, organization, and language all affect the final score. Even though this subject has been tackled in the past in a number of ways, there is still opportunity for enhancement, some of these are covered in this book. Because subjective assessments rely so heavily on context, both teachers and students view them as more challenging and intimidating. When a response is subjective, the checker must actively score each word in the response; the total score is significantly impacted by the checker's objectivity, weariness, and mental state.

Thus, letting a system conduct this time-consuming and sometimes crucial work of assessing subjective responses is far more economical in terms of both time and resources. Machine evaluation of objective responses is a relatively simple and practical process. One-word answers to questions can be entered into a program to rapidly map students' responses. Subjective responses, however, are far more difficult to address. They range widely in

ISSN: 2348-1358

Impact Factor: 6.901

NAAS Rating: 3.77

length and have a large vocabulary. Additionally, people frequently utilize handy abbreviations and synonyms, which further complicates the procedure.

II. LITERATURE SURVEY

A new latent semantic indexing method was introduced by Hu and Xia [6] for assessing arbitrary online inquiries. To create a k-dimensional LSI spatial matrix, they employed subjective ontologies and Chinese automatic segmentation algorithms. A vector-based semantic space was produced by the single value decomposition (SVD) of the term-document matrix following the encoding of the answers in TF-IDF embedding matrices. When it came to solving problems with synonyms and polysemy, LSI worked well. Lastly, the resemblance of the responses was calculated using cosine similarity. 850 instances from 35 classrooms which were graded by academics have been included in the dataset. The differences between the teacher-marked instances and the recommended technique were found to be 5%. Word Mover's Distance (WMD) is a novel technique put forth by Kusner and Associates [7] for determining the degree of divergence between two texts. The system did not employ any hyper-parameters and instead loosened the vector space bounds using a relaxed WMD technique. Eight real-world collections were included in the dataset, including emotion data from Twitter and BBC sports articles. Two more custom models were trained to add to the News from Google Word2vec model. The KNN classification technique was used to classify the testing data. Therefore, reducing WMD resulted in fewer mistake rates and between 2- and 5-times faster categorization. Kim et al. Due to its excellent performance with the Korean language's complex morphology, [32] suggested a way to use the lexico-semantic pattern (LSP) to score succinct descriptive responses. To better comprehend the user's intentions, LSP might organize the answer's semantics. Additionally, to help the keywords fit differently, a list of alternatives was used.

After gathering and converting the datasets of 89 pupils into LSP, the correct response was determined by comparing the LSP replies with the dataset. The system so performed 0.137 better than the current system. A pair-wise resemblance metric was presented by Oghbaie and Zanjireh [33] to assess the degree of similarity between two papers based on keywords that appear in a minimum of one of the documents. The in pairs document similarities measure, or PDSM, is a modified form of the desired properties technique that was introduced in this article. The suggested similarity metric was used in text mining applications like word recognition, clustering with K-means, and k Nearest Neighbour (kNN) for single-label classification. The PDSM method outperformed other metrics, including the Jaccard coefficients, by 0.08 recall when an accuracy measure was used. Words were represented on a fixed-sized vector-based model by Orkphol and Yang [34] using the word2vec technique, and the similarity of sentences was evaluated using a cosine similarities measure. Using the Google tool Word2vec, the phrase vector was generated by combining the phrases inside the sentence. It was deemed acceptable if the similarity results had a score between 0 and 1. With and without a probability of sense distribution, the system scored 50.9% and 48.7%, respectively, when recall and accuracy were employed as assessment metrics. To find commonalities among different legal papers, Xia et al. [8] employed the word2vec method when paired with the corpus of legal documents. Numerous text vectors were compared for similarity using the cosine similarity measure. Consequently, word2vec achieved a 0.2 accuracy advantage over the Bag of Words method. The word2vec model can be trained on legal texts to augment this improvement by a factor of 0.06-0.11. A multi-criteria decision-making perspective was put forth by Wagh and Anand [35] to assess the degree of resemblance across legal documents. An artificial intelligence-based similarity score between many publications was to be determined, along with aggregation techniques such ordered weighted mean (OWA). The data was compiled from Indian Supreme Court decisions rendered among 1951 and 1994. Evaluation criteria included recall and F1 score. Consequently, the based on concepts similarity approach suggested in the study achieved a F1 score of up to 0.8, surpassing previous techniques like TF-IDF.

ISSN: 2348-1358

Impact Factor: 6.901

NAAS Rating: 3.77

III. SYSTEM ANALYSIS

A. EXISTING SYSTEM

In this work, we investigate a machine learning and natural language processing method for analysing subjective responses. The foundation of our work includes tokenization, lemmatization of text representation methods like TF-IDF, Bag of Words, and word2vec; similarity measurement methods like cosine comparison and word mover's distance; and classification strategies like multinomial Naive Bayes. We evaluate several models according to several metrics, such as F1-score, Accuracy, and Recall, and compare their performances. Additionally, we go over several historical methods for evaluating subjective answers and overall text similarity.

Some of the major disadvantages of adopting subjective replies are as follows:
• Synonyms have been extensively used in previous study.
• The current lengths of research sometimes vary greatly.
• Existing studies often feature sentences placed in arbitrary sequence.

DISADVANTAGES OF THE EXISTING SYSTEM

1. **Model performance:** is greatly influenced by the quality and representativeness of training data. If the training data is biased, incomplete, or does not cover the entire range of subjective responses, the models may produce inaccurate or biased results.

2. **Difficulty with Complicated Sentences:** Subjective comments usually contain complex phrase structures, colloquial phrases, and sophisticated terminology. Machine learning models may be unable to accurately identify and evaluate such complexity, resulting in errors or misinterpretation.

3. **Limited Generalization:** While models perform well with training data, their capacity to generalize to new data or circumstances may be limited. Language, subject, and domain differences can all affect model performance and reliability.

4. **Lack of Interpretability:** It can be difficult to understand how complex machine learning models, such as neural networks, generate decisions. This lack of openness can cause problems, especially in applications that require explanations or justifications.

5. **Scalability concerns:** Training and deploying machine learning models for subjective responses can be resource-intensive, particularly for big datasets or real-time evaluations. Scaling the system to support more data or users may cause computational and infrastructure challenges.

B. PROPOSED SYSTEM

This work presents an enhanced method that leverages natural language processing and machine learning to automatically analyze descriptive question responses. It solves this problem using a two-step process. First, similarity-based techniques like word mover's proximity are used to compare the responses to the answer and specified keywords. After that, a model trained on the outcomes of this step is able to assess responses without the need for solutions or keywords. For instance, the following is the right response to the arbitrary question: "What is the capital town of Pakistan famous for?" Pakistan's capital city, Islamabad, is renowned for its mountainous landscape. The system gets the query and the response before assessing the student's response, along with a few essential keywords (Islamabad and mountain scenery, for example). The system assesses the student's response by contrasting its similarities (while keeping the context in mind) with the modal answer, as well as the inclusion or lack of keywords. Thus, a student's response, "Karachi is the capital city of Pakistan, it is infamous for mountain scenery," might receive 50% of the possible points; "Islamabad and mountain scenery," might receive 30% because the primary phrases are present despite the absence of context; and, in contrast to the correct response, "Islamabad is the city and it is renowned for mountain scenery," might receive 100% because it fulfils both contextual parallels and keyword presence.

ISSN: 2348-1358

Impact Factor: 6.901

NAAS Rating: 3.77

IV.SYSTEM DESIGN

SYSTEM ARCHITECTURE

Below diagram depicts the whole system architecture.

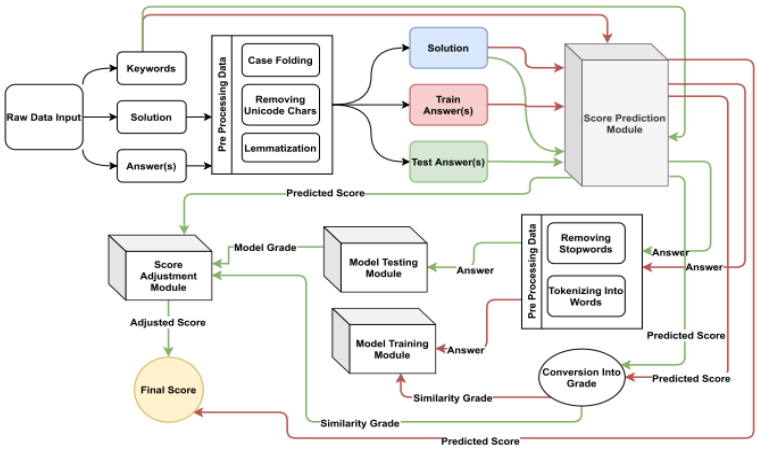

Fig 1. Methodology followed for proposed model

V. SYSTEM IMPLEMENTATION

MODULES

A. DATA COLLECTION

We are unaware of any publicly accessible labelled objective inquiry answer corpus, which would be required to create and test the proposed model on a large dataset of subjective question responses. In this study, we use corpus data to generate labelled subjective answers. To create a corpus, look for blogs and websites that contain random questions and answers. We monitor several websites and create a subjective question-and answers corpus from crawl data covering a wide range of topics, spanning general understanding and computer science.

B. DATA ANNOTATION

There needs to be more annotation because the data that was crawled is not labelled. We choose a sample of volunteers from our arbitrary question-and-answer corpus to annotate the data. We recruit thirty annotators from Pakistani universities and other organizations. They are mostly either teachers or pupils. Annotators can be any age from 27 to 51, although most are around the ages of twenty-one and twenty-five. Based on student responses, we assigned annotators the task of rating subjective question answers.

C. PREPROCESSING MODULE

Preprocessing operations like tokenization, stemming, lemmatization, stop word removal, case folding, and locating and relating synonyms to the text following user input are performed by both the answer and the solution. Because word2vec has a broad vocabulary and can use stop words to improve a document's semantic meaning, stop words are not removed when new data is introduced. Before being input into a machine learning model such as Multinomial Naive Bayes, stop words are removed since they impede the system's capacity to recognize patterns.

D. SIMILARITY MEASUREMENT MODULE

The WDM, also known and cosine similarities operations, which analyse two words or vectors of words and determine their similarity, are included in this module. While Cosine Similarity quantifies similarities, WDM educates us about dissimilarity. Our approach compares the outcomes after applying

ISSN: 2348-1358

Impact Factor: 6.901

NAAS Rating: 3.77

each of the two similarities metrics one at a time. distinct cutoff points for similarity and dissimilarity. 1) Analysis of thresholds. It has been empirically shown that a few of the study's parameters produce the best outcomes. The dissimilarity of two sentences is determined using the WDM_LOWER and WDM_UPPER criteria, where a greater degree of similarity is indicated by a higher dissimilarity. Sentences with a 0.7 threshold for WDM_LOWER and a 1.6 barrier for WDM_UPPER, respectively, indicated how comparable the sentences were in the trials. A difference of more than 1.6 is deemed too great to allow for meaningful comparisons. Similarly, the cosine similarity criteria COS_LOWER and COS_UPPER are used to determine how similar two sentences are to one another. It is noteworthy that cosine similarity, in contrast to WDM, does not include the meaning of the two words when calculating similarity; hence, both methods must be used to measure dissimilarity and similarity.

E. RESULT PREDICTING MODULE

The Result Predicting Module serves as the study's foundation. demonstrates the module's functionality.

VI . RESULTS AND DISCUSSION

A Python a notebook with an a web-based Google Collaborate site, a hard disk greater than 100 GB, and 12 GB of RAM are needed for the experiment. Throughout the test, the GPU is turned off. This study used a 300-dimensional, pre-trained Google word2vec model with about 100 billion words in its lexicon. To separate the training and testing data, the dataset was partitioned into 8:2 ratios. The train data was used to train the machine learning model, and the score prediction modules were assigned starting scores. Then, when more testing data was uploaded to the system, the machine learning model was gradually modified. Word mover closeness and cosine similarity are combined in a multinomial Naive Bayes model to produce the desired results. Both approaches with and without the model produced results at Google Collab in less than a minute. These are the results.

Human Score	Error Without Model	Error With Model
46	22	9.5
46	17	4.5
60	13	25.5
60	14	26.5
55	9	3.5
55	25	12.5
27	22	9.5
0	0	12.5
77	40	27.5
27	26	13.5

Table 1. Score prediction using WDM with model suggestion

ISSN: 2348-1358

Impact Factor: 6.901

NAAS Rating: 3.77

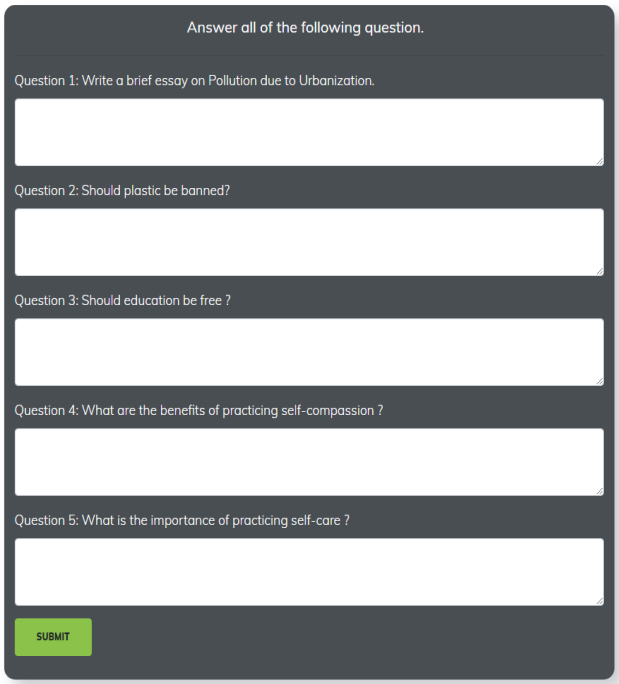

Fig 2. Input Subjective Answers For Evaluation

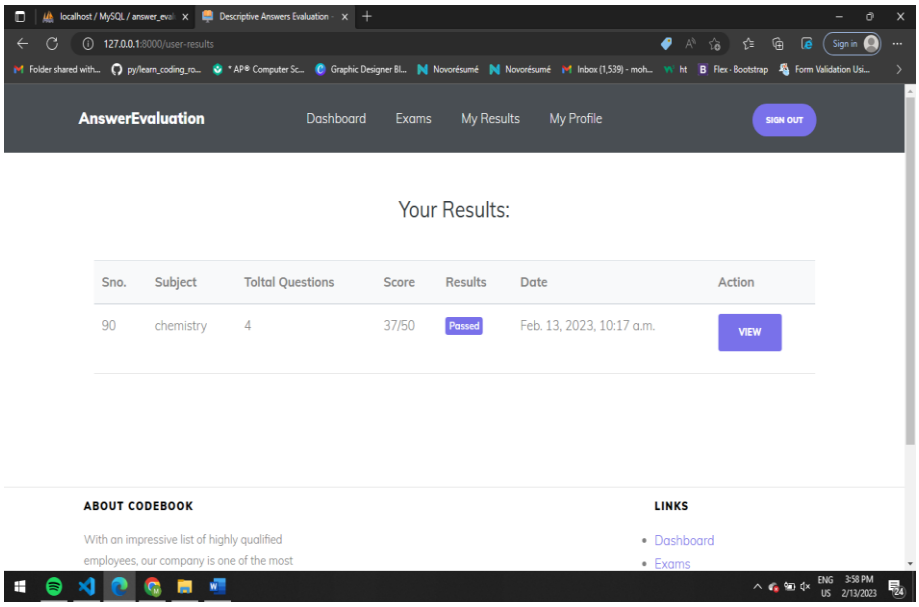

Fig 3. Results Analysis

VI I.CONCLUSION AD FUTURE WORK

This study introduces a novel way for evaluating subjective responses that combines machine learning and natural language processing techniques. Two score prediction methods with a maximum accuracy of 88% are proposed. To account for the irregular frequency of responses with loose semantics, several thresholds for similarities and differences are analysed. Additionally, other techniques such percentage sentence translation and keyword

ISSN: 2348-1358

Impact Factor: 6.901

NAAS Rating: 3.77

presence are used. The experimental results show that the word2vec method outperforms conventional word embedding techniques while maintaining semantics. In addition, Word Mover's Distance typically exceeds Cosine Similarity, this enables the machine learning model to be trained more quickly. If the system is trained appropriately, it can forecast scores without the need for semantic testing. At some point, the word2vec model will receive specialized training to evaluate subjective reactions to a given topic. It will also greatly expand the number of classes or grades in the model by adding enormous data sets. Subjective response evaluation is still a fascinating field of study, and we hope to discover more effective methods in the future.

REFERENCES :

A survey on the measurement of text similarity was conducted by J. Wang and Y. Dong. Information, August 2020, vol. 11, no. 9, p. 421.

[2] A survey on the methods, uses, and efficacy of short text semantic similarity was conducted by M. Han, X. Zhang, X. Yuan, J. Jiang, W. Yun, and C. Gao in Concurrency Comput., Pract. Exper., vol. 33, no. 5, Mar. 2021.

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

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

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

[6] "Automated assessment system for subjective questions based on LSI," by X. Hu and H. Xia, in Proceedings of the 3rd International Symposium on Intelligence, Information, and Technology, April 2010, pp. 250–254.

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

[8] "Similarity analysis of law documents based on Word2vec," by C. Xia, T. He, W. Li, Z. Qin, and Z. Zou, was published in the IEEE 19th International Conference on Softw. Qual., Rel. Secur. Companion (QRS-C), July 2019, pp. 354–395.

[9] In Appl. Artif. Intell., vol. 32, no. 1, pp. 85–95, Jan. 2018, H. Mittal and M. S. Devi, "Subjective evaluation: A comparison of several statistical techniques."

[10] Cutrone, L. A. and Chang, M., "Automarking: Automatic assessment of open questions," in Proceedings of the 10th IEEE International Conference on Advanced Learning Technologies, Sousse, Tunisia, July 2010, pp. 143–147.

[11] "A two-stage text feature selection algorithm for improving text classification," by G. Srivastava, P. K. R. Maddikunta, and T. R. Gadekallu Rep. Tech., 2021.

[12] A comprehensive framework for subjective information extraction from unstructured English text by H. Mangassarian and H. Artail, Data Knowl. Eng., vol. 62, no. 2, pp. 352–367, Aug. 2007.

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

[14] H. Khan, "Fake review classification using supervised machine learning," in Proc. Pattern Recognit. Int. Workshops Challenges (ICPR). New York, NY, USA: Springer, 2021, pp. 269–288. With M. U. Asghar, M. Z. Asghar, G. Srivastava, P. K. R. Maddikunta, and T. R. Gadekallu.