Industrial Engineering Journal
ISSN: 0970-2555
Volume : 53, Issue 8, No.3, August : 2024

SUBJECTIVE ANSWERS EVALUATION USING MACHINE LEARNING

Dr. Nilima Warade, Dept. Of Electronics & Telecommunication Engineering, AISSMS Institute of Information Technology, Pune.
Nikhil Yadav, Electronics & Telecommunication Engineering, AISSMS Institute of Information Technology, Pune.
Anurag Kotasthane, Electronics & Telecommunication Engineering, AISSMS Institute of Information Technology, Pune.
Gaurav Tayade, Electronics & Telecommunication Engineering, AISSMS Institute of Information Technology, Pune.

ABSTRACT
Subjective answer evaluation is an interesting and tedious errand to do by physical work. Lacking comprehension and acknowledgment of information are vital difficulties while examining abstract papers utilizing machine learning. A few endeavours have been made to score student's responses utilizing software engineering, such as NLP, which can be utilized in abstract answer assessment in different ways. One of the most well-known approaches is to utilize NLP procedures to naturally score the nature of a composed reaction in light of its language highlights, like sentence structure, grammar, jargon, and intelligence. Another methodology is to utilize NLP methods to dissect the substance and design of the reaction, to recognize key ideas and contentions and survey their importance and cognizance with the inquiry brief. Be that as it may, the greater part of the function utilizes customary counts or explicit words to accomplish this assignment. Moreover, there is an absence of arranged information collections also. This paper proposes an interesting methodology that uses different AI, procedures and instruments. For example, WorldNet, Cosine Closeness, Word Mover's Distance (WDM), Term Frequency-Inverse Document Frequency (TF-IDF) and Multinomial Naive Bayes (MNB) to assess clear responses consequently. Arrangement proclamations and watchwords are utilized to assess replies, and an AI prototype is prepared to anticipate the marks of replies. Results prove that WDM works better than cosine closeness by and large. With enough preparation, the AI model could be utilized as an independent too. Trial and error creates a precision of 89% except the MNB technique. The blunder rate is additionally decreased by 1.4% utilizing Multinomial Naive Bayes.

Keywords:
NLP, WDM, PDSM, LSA, TF-IDF, KNN, MNB, LSP, RNN, CNN.

I. Introduction
Theoretical questions and solutions can be used to evaluate the aptitude and performance of a student in an unassuming way. The responses, normally, are not attached to any imperative, and students are allowed to think of them as indicated by their attitude and comprehension of the idea. All things considered, a few other indispensable contrasts separate emotional responses from their objective partner. For one's purposes, they are significantly detailed than the objective type of questions. Besides, they find an opportunity to compose another that conveys substantially more setting and takes a ton of focus. Furthermore, objectivity from the instructor assessing them. Assessment of such inquiries utilizing PCs is an interesting errand, for the most part, since regular language is equivocal. A few pre-processing steps should be performed, like cleansing the information and tokenization prior to performing operations on it. Then the text-based information can measure up utilizing different procedures like report similarity, inactive semantic designs, ontologies. The last score can be assessed by view of similarity, catchphrases, presence, structure, language. A few endeavours have been made in the past to tackle this issue, yet there is still an opportunity to get better. Subjective exams are viewed as additionally complicated and terrifying by the students and teachers both due to one crucial element,

context. Such questions and answers demand the evaluator to check each expression of the solution for scoring effectively, and the evaluator's emotional well-being, weariness, and objectivity assume a gigantic part in the general outcome. Hence, it is substantially more time and asset proficiency to allow a framework to deal with this drawn-out and fairly basic assignment of assessing abstract responses. Assessing objective responses with the help computer is extremely simple and doable. A code/program can be taken care of with query and single word addresses that can rapidly plan student's responses. Nonetheless, theoretical answers are considerably more demanding to handle. They are different, long and consist of huge measure of vocab. Besides, individuals will generally utilize equivalent words and helpful abbreviations, which make the interaction very uncertain. Much of the work has been finished on the theoretical answers assessment in some structure, like estimating comparability between various texts, words, and even reports, tracking down the setting after the text also, planning it with the arrangement's unique circumstances, including the thing expression in the reports, matching catchphrases in the answers, etc. But still, issues, for example, TF-IDF losing context for semantic, absence of hyperboundary tuning, expensive training, and the requirement for better datasets for advance performance still exist. We are investigating an AI, NLP Natural Language processing focused approach to theoretical answer assessment. This work depends on regular dialect handling methods such as tokenization, lemmatization, text addressing procedures like TF-IDF, Pack of Words, comparability estimating strategies like cosine similitude, and word mover's distance. We utilize different assessment measures like score, precision, review to consider the presentation of different models in contrast to one another. A better than ever approach to assessing unmistakable inquiry responds to consequently utilizing Artificial Intelligence and Natural Language Processing. It utilizes ways to deal with tackling this issue. The responses are assessed utilizing the arrangement and gave catchphrases utilizing different comparability-based strategies like ex. Word Mover's Distance (WMD). After that the outcome is utilized to prepare a prototype that will assess answers in the absence of solutions and keywords without any requirement.

II. Literature Survey

As referenced previously, the assessment of theoretical answers is certainly not a novel insight, what's more, it has been worked upon for practically twenty years. Different methods have been implemented to take care of this issue, like huge information Latent Semantic Analysis (LSA), Bayes theorem, Natural Language Processing, K-closest classifier, and surprisingly formal methods, for example, Formal Idea Examination. They are ordered into three primary classifications: Information Extrication, Statistical, and Full NLP.

[1] Kusner et al. introduced a unique thought of utilizing Word Mover's Distance (WMD) to see the distinction among two texts. This process used no hyperparameters and it utilized a freed WMD technique to soften up the vector space bounds. The datasets consisted eight real world groups, including X (formerly known as Twitter) sentiment data and sports articles of British Broadcasting Corporation. The Word2vec model of the Google News was used, what's more, two other specially made custom models were prepared. KNN grouping technique was utilized to distinguish the testing data. As seen, freed WMD minimized the errors. It lead to higher efficient program and better results which led to 3 to 5 times' faster classification.

[2] Hu and Xia suggested an Idle Semantic Ordering attitude for the evaluation of theoretical questions on the internet. They utilized Chinese programmed division strategies and abstract ontologies to make an LSI space lattice. The solutions were introduced in TF-IDF implanting matrices, and afterward Singular Value Decomposition (SVD) was utilized to the term-report framework, which shaped a semantic space of vectors. LSI assumed the part of diminishing issues with equivalent and equivocacy. Finally, the link between solutions was determined utilizing cosine similarity. The data set comprised of 35 classes and 850 occasions set apart by educators, and the outcomes showed about 5% distinction in evaluating carried out by the instructor and the proposed framework.

[3] Kim et al. suggested a method for grading short descriptive answers that is Lexico Semantic Pattern (LSP) because of its great compatibility with morphologically hard language. LSP can be used to define the semantics of the solution to assist and comprehend what are the true objectives of the user. A similarity list was also employed to help increase the keywords so that they can find various solution styles. Dataset was collected from different test subjects and was made suitable to be used with LSP, it was then compared to the LSP solution in order to grade the response. Subsequently, the framework did better job compared to the current framework.

[4] Oghbaie and Zanjireh presented a couple of resemblance measure to count the similarity between two docs depending on the keywords occurring in at least one of the docs. This work gave a new comparability measure called Pair-wise Document Similarity Measure (PDSM), an updated version of the best properties approach. This suggested similitude measure to be applied for text mining apps such as document detection, (KNN) k Nearest Neighbours for single-label classification, and K-implies clustering/grouping. An evaluation proportion of precision was utilized, and accordingly, the strategy of PDSM created improved results than different measures.

III. Implementation

This proposed system involves data grouping and explanation, pre-handling module, comparability assessment module, model readiness module, result predicting module, ML model module, and outcome expecting module. Most importantly, the data is abstracted from the user, which contains keywords, solutions and answers.

Assessment Models: First it begins by portraying the standards which are expected to assess theoretical responses. These principles could integrate components like significance, clarity, profundity of assessment, clearness, and inventiveness, dependent upon the possibility of the responses you're checking.

Explanation Cycle: Accumulate a dataset of responses close by human remarks that exhibit the quality or grade for each standard. Human annotators should be ready to ensure consistency and constancy in their choices. Ideally, various annotators should survey every answer for get substitute perspectives and lessening inclination.

Include Extraction: Focus on features from the text that are relevant to the evaluation principles. These features could consolidate word frequencies, syntactic plans, feeling exam scores, intelligibility. From that point, anything is possible. Consider procedures like pack of-words, TF-IDF, word inserting or significant learning-based approaches for featuring extraction.

Model Determination: Taking a AI models for this is necessary. Including relapse models, characterization models, or more refined models like (RNNs) and (CNNs) is good for overall system performance. Final decision will rely upon the idea of the assessment measures and the work's intricacy.

Preparation: Dataset is split into preparation, approval, and test sets. After that comes the training of AI model on the information preparation consisting of fitting capabilities and streamlining sums.

Checking: Checking the exhibition of framework on the test set by fitting assessment measurements. Exactness, accuracy, review, F1 score, relationship coefficients are measured depending upon the requirements.

Iterative Improvement: Repeat your approach by refining your explanation cycle, highlighting extraction methods, model engineering, and hyperparameters based on experience acquired from assessing model execution. This might include gathering more information, further developing component portrayal, or exploring different avenues regarding different model designs.

IV. Block Diagram

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 53, Issue 8, No.3, August : 2024

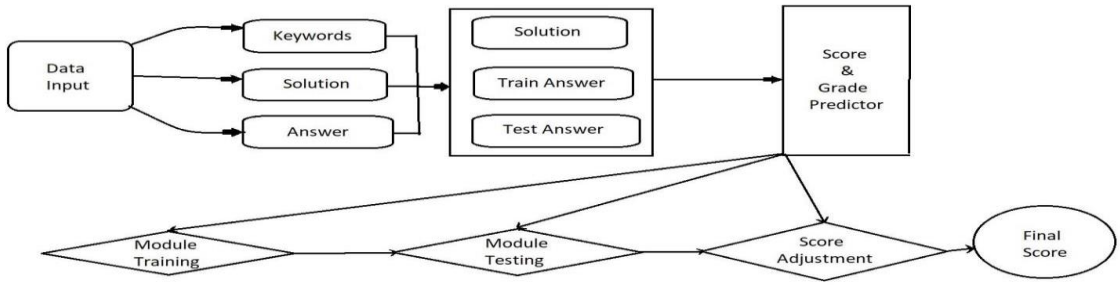

Fig. 1 Block Diagram

V. Methodology

The proposed framework comprises information assortment and annotation, pre-processing module, similitude estimation module, results foreseeing module, model preparation module, ML model module, and final-result anticipating module. To start with, the sources of info are gathered from the user, which comprises is solutions, keywords, and solutions. Then that data is processed and sent further for the end result.

Characterize Assessment Models: Begin by obviously characterizing the standards you need to use to assess emotional responses. These standards could incorporate elements like importance, lucidness, profundity of examination, clearness, and creativity, contingent upon the idea of the answers you're assessing.

Explanation Cycle: Gather a dataset of subjective answers alongside human comments that demonstrate the quality or score for every standard. Human annotators ought to be prepared to guarantee consistency and dependability in their decisions. In a perfect world, numerous annotators ought to assess each solution to catch alternate points of view and decrease bias.

Include Extraction: Concentrate on highlights from the text that are pertinent to the assessment standards. These highlights could incorporate word frequencies, syntactic designs, feeling investigation scores, comprehensibility measures. From there, the sky is the limit. Consider utilizing strategies like pack of-words, TF-IDF, word embedding or profound learning-based approaches for highlighting extraction.

Model Deciding: Pick fitting AI models for the undertaking. This could include relapse models, characterization models, or more refined models like RNNs and CNNs. The decision of model will rely upon the idea of the assessment measures and the intricacy of the task.

Preparing: Split your clarified dataset into preparation, approval, and test sets. Train your AI model on the preparation information utilizing fitting misfortune capabilities and streamlining calculations. Utilize the approval set to tune hyperparameters and screen execution.

Assessment: Assess the exhibition of your model on the test set utilizing fitting assessment measurements. These measurements could incorporate exactness, accuracy, review, F1 score, or relationship coefficients, depending upon the particular task and assessment standards.

Continuous Improvement: Rehashing methodology by refining clarification cycle, featuring extraction strategies, model designing, and hyperparameters for involvement obtained from surveying model execution. This could incorporate assembling more data, further creating part depiction, or investigating various roads in regards to various model plans.

VI. Results

Generate Synthetic Dataset: Make a synthetic dataset of subjective answers alongside related scores or names, demonstrating the nature of each answer in light of predefined assessment rules. For

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 53, Issue 8, No.3, August : 2024

straightforwardness, we should accept two assessment models: relevance and clarity. Create irregular subjective answers and allot scores from 1 to 5.

Fig. 2 Data Synthesis

Include Extraction: Concentrate highlights from the synthetic dataset. For this simulation, we should consider essential highlights, for example, word count, normal word length, and opinion score of the responses.

Model Determination: Pick a basic relapse model for this simulation. Straight relapse is a decent beginning stage.

Training: Separate the synthetic dataset into test sets and training. Train the straight relapse model on the preparation set involving the removed highlights as info and the allotted scores as target values

Fig. 3 Training

Evaluation: Assess the prepared model on the test set to recreate its presentation. Compute assessment metrics like mean squared mistake (MSE) or Pearson relationship coefficient to survey how well the model predicts the allotted scores in light of the info highlights.

Representation: Imagine the re-enacted results by plotting the anticipated scores against the genuine scores for the test set. This assists with understanding how intently the model's expectations line up with the genuine scores.

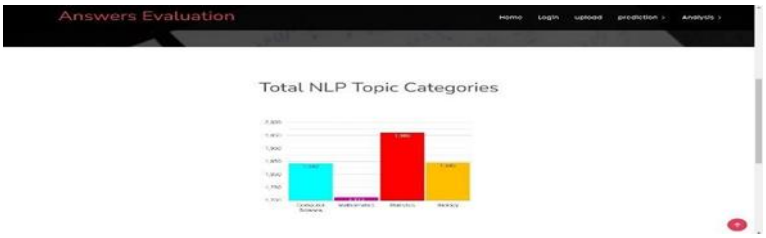

Fig. 4 Analysis

Examination and Understanding: Analyze the simulated results to acquire bits of knowledge for the model's exhibition. Recognize any examples or regions where the model might be struggling to make precise predictions.

Fig. 5 Accuracy Graph

Final Result: Sum up the results and get conclusions about the effectiveness of the picked approach for subjective answers evaluation utilizing AI and ML in this simulated situation.

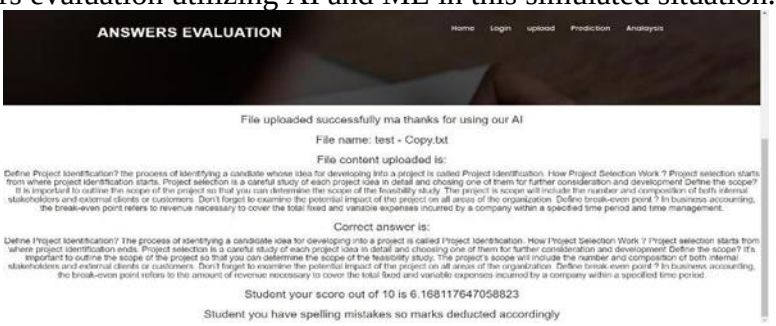

Fig. 6 End Results

VII. Conclusion

This paper gives a clever way to deal with theoretical answers assessment in view of AI and NLP strategies. Two score forecast methods are suggested, which results up to 89% precise results. Various similarities and different edges are examined, and different measures like the watchword's existence and rate planning of sentences are used to deal with the strange instances of semantically free answers. The experimentation results prove that the normal word2vec technique gives better results compared to conventional word implanting strategies as it safeguards the semantics. Besides, WDM performs better compared to Cosine Similitude. In addition, it assists train the AI model quicker. With plenty of preparation, the model can withstand all alone and determine scores without the requirement for any kind of semantics examining. Concerning future enhancements, the word2vec model can be prepared particularly for theoretical solutions and assessment of a specific area, and with huge informational indexes, the quantity of grades or classes in the model can be altogether expanded. Theoretical answers and assessment is an intriguing issue to handle, and later on, we desire to see more proficient techniques of taking care of this issue.

VIII. Future Scope

Personalized Feedback: Tailored suggestions and adaptive learning experiences.
Automated Grading Systems: Accurate, consistent grading reducing educators' workload.
Enhanced Educational Tools: Real-time feedback and adaptive resources from intelligent tutoring systems.
Research and Development: Exploration of new algorithms and improved training datasets.
Scalability and Accessibility: Large-scale evaluation making high-quality education more accessible.
Real-World Applications: Use in job recruitment, customer service, and legal analysis for better decision-making.

IX. References

[1] Kusner, M. J., Sun, Y., Kolkin, N. I., and Weinberger, K. Q. From word embeddings to document distances. In ICML, 2015.
[2] X. Hu and H. Xia, „„„Automated assessment system for subjective questions based on LSI,"" in Proc. 3rd Int. Symp. Intell. Inf. Technol. Secur. Informat. , Apr. 2010, pp. 250–254
[3] Kim, J. E., Chae, J. M., & Jung, S. Y. (2014). Automatic scoring system based on Korean lexico-semantic pattern. *International Journal of Applied Engineering Research*, *9*(22), 14499-14510.
[4] Oghbaie, M., Mohammadi Zanjireh, M. Pairwise document similarity measure based on present term set. J Big Data 5, 52 (2018).
[5] P. Resnik, „„Using information content to evaluate semantic similarity in a taxonomy, in Proc. 14th Int. Joint Conf. Artif. Intell. , IJCAI, Montréal QC, Canada, Aug. 1995, pp. 448– 453

[6] J. Wang and Y. Dong, ‘‘Measurement of text similarity: A survey,’’ Information, vol. 11, no. 9, p. 421, Aug. 2020.
[7] Text similarity analysis for evaluation of descriptive answers, 2021, (V. Bahel and A. Thomas,)arXiv:2105.02935
[8] P. Patil, S. Patil. V. Miniyor, and A. Bandal "Answer evaluation using AI." Int. J. Pure Appl. Math., vol. 118, no. 24. pp. 1-13, 2018
[9] Jabbar, S. Iqbal, M. I. Tamimy, S. Hussain, and A. Akhunzada, „„Empirical evaluation and study of text stemming algorithms, ‟Artif. Intell. Rev., vol. 53, no. 8, pp. 5559–5588, Dec. 2020.
[10] S. Wagh and D. Anand, „„Legal document similarity: A multi criteria decision- making perspective, PeerJ Comput. Sci. , vol. 6, p. e262, Mar. 2020
[11] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell et al., “Language models are few-shot learners,” Advances in neural information processing systems, vol. 33, pp. 1877–1901, 2020
[12] M. I. El Desouki and W. H. Gomaa, “Exploring the recent trends of paraphrase detection,” International Journal of Computer Applications, vol. 975, no. S 8887, 2019
[13] Y. Yang, W.-t. Yih, and C. Meek, “WikiQA: A challenge dataset for open-domain question answering,” in Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. Lisbon, Portugal: Association for Computational Linguistics, Sep. 2015, pp. 2013–2018
[14] Bloehdorn S, Basili R, Cammisa M, Moschitti A (2006) Semantic kernels for text classification based on topological measures of feature similarity. In: Sixth International Conference on Data Mining (ICDM‟06), pp 808–812
[15] Sachin Ravi and Hugo Larochelle. Optimization as a model for few-shot learning. ICLR 2017 (oral), 2016.
[16] Nils R, Iryna G. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks[C]. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. 2019: 3982-3992.
[17] Bag, S., Kumar, S. K., and Tiwari, M. K. (2019). An efficient recommendation generation using relevant jac-card similarity. Information Sciences, 483:53–64.
[18] Chanda Roy, Chitrita Chaudhuri, “Case Based Modelling of Answer Points to Expedite Semi-Automated Evaluation of Subjective Papers”, in Proc. Int. Conf. IEEE 8th International Advance Computing Conference (IACC), 2018, pp. 85-9.