Review of Automatic Subjective Answer Grading Software Using Machine Learning

Haniya Abdul Rahman, Arjun Gopi K, Evans Nevil, Gautham Chand
FIT20CS058,FIT20CS039,FIT20CS050,FIT20CS051 BTech, Computer Science - 2020-2024
Federal Institute of Science and Technology
Main Project

I. Abstract

The manual evaluation of handwritten descriptive answers remains a laborious task, characterized by challenges such as incomplete comprehension and data interpretation complexities. While Artificial Intelligence (AI) has been explored for grading computer science responses, its efficacy is hindered by reliance on basic word counts and specific terms, compounded by the scarcity of well-organized datasets. This literature review introduces a machine-learning-based solution that leverages solution patterns and significant keywords to predict answer grades. In our experimentation, we explore the utilization of cosine similarity and semantic search as viable alternatives to traditional methods, with promising results. As the project progresses, it holds promise as an independent tool capable of autonomously evaluating handwritten descriptive answers, mitigating external dependencies. This pioneering approach aims to contribute to the development of more effective and reliable automated evaluation systems for subjective paper responses, addressing the challenges prevalent in educational assessments.

II. A NLP Approach for Automatic Test Evaluation System[1]

A. Authors

Mayank Agarwal Department of Computer Science & Engineering G H Raisoni College of Engineering, Nagpur.

B. Overview

The paper discusses the need for automation in assessment, focusing on a system called AutoEval - Automatic Test Evaluation System. This system utilizes Natural Language Processing (NLP) to address challenges in evaluating theory-based answers, especially in online assessments dominated by option-based questions. The manual evaluation process by educators is time-consuming and resource– intensive, prompting the development of AutoEval. AutoEval aims not only to streamline evaluation but also to standardize paper correction efficiently and reliably. It emphasizes the importance of fair evaluation in shaping student outcomes and recognizing potential biases in manual assessments. The NLP-driven AutoEval system is highlighted for its capabilities in tasks such as identifying grammatical errors, conducting syntactic analysis, assessing semantic similarity, and enabling efficient database storage. The integration of NLP in AutoEval is presented as a potential game-changer, promising increased accuracy, objectivity, and efficiency in the assessmentprocess.

C. Implementation

In answer verification using Natural Language Processing (NLP), the initial step employs the NLTK tool for text pre-processing[2][3]. This involves tasks like stemming and removing stop words to enhance the clarity and structure of the input answer. Following this, tokenization comes into play. It reads the file, breaks sentences into tokens of words, creates a dictionary that captures the vocabulary and constructs a bag-of-words model that represents the input in numerical terms based on word frequency. A pivotal aspect of this method is the utilization of cosine similarity measures. This mathematical technique assesses the angular similarity between the bag-of-words vectors derived from the input answer and reference data. Quantifying the cosine of the angle between these vectors provides a numerical indication of how closely the input aligns with known patterns or correct responses synergy of NLTK, tokenization[2], and cosine similarity measures results in a sophisticated yet efficient NLP-driven approach to answer verification. This method not only automates assessment but also helps understand text similarities better. This leads to more reliable and insightful evaluations of natural languageunderstanding.

D. *Case Study*

This recently developed cosine similarity algorithm presents a groundbreaking solution for the evaluation of theoretical answers, providing a notable reduction in manual efforts and expediting the grading workflow. Operating on a keyword-matching similarity index, the algorithm autonomously examines responses, assigning marks or grades based on the alignment with predefined keywords and key phrases. This algorithmic approach introduces efficiency and consistency into the evaluation process, eliminating the tedious nature of manual assessments. By automating the assessment of theoretical answers, this algorithm ensures a more rapid and standardized grading experience.

E. *Results*

When a student enters the correct information during authentication, he/she will be able to give responses and view the final result. The report then shows the final calculated scores awarded to the student answers. The student can access their results by logging in with their respective register number and their password on the result page. The proposed system takes about 15 seconds to evaluate a response, whereas the current manual assessment takes about 60 seconds. When opposed to manual answer evaluation systems, the proposed systems save 300% of time. As compared to manual systems, the proposed system is around 75-87.5% effective. The proposed system removes all human effort and time required to determine a response. As compared to our proposed system, the amount of money spent on manual response assessment is much higher. Our proposed system is a one-time investment with minimal ongoing maintenance costs.

III. Automated Evaluation of Handwritten Answer Script Using Deep Learning Approach[2]

A. *Authors*

(M d. Afzalur Rahaman1 , Hasan M ahmud2 Department of CSE, Hamdard University Bangladesh1,2)

B. *Overview*

The objective of the system is to perform the dual tasks of classifying handwritten text and converting it into a digital format while automating the grading process. In the classification stage, the system utilizes advanced image recognition techniques to categorize and understand handwritten text effectively. This involves the use of machine learning models trained on diverse handwriting styles to ensure robust classification accuracy. Subsequently, the converted digital text is subjected to automated grading mechanisms. Here, the system employs natural language processing (NLP) techniques and potentially machine learning algorithms to assess and assign grades based on predefined criteria. The integration of these functionalities streamlines the overall evaluation process, offering a comprehensive solution for efficiently handling handwritten assessments, converting them into a digital format, and providing automated grading, thereby saving time and enhancing accuracy in educational assessments.

C. *Implementation*

The processed data, having undergone word segmentation and subsequent conversion into a grayscale image alongside the original data, is then prepared for text recognition through an Optical Character Recognition (OCR) model. This model is tailored to accommodate our dataset by resizing all word-segmented images to a uniform dimension of 32128. Leveraging the IAM dataset, which comprises handwritten English sentences, the words are fed into a Convolutional Neural Network (CNN) layer. This CNN layer is configured with a kernel size of (3, 3) and incorporates 64 nodes. Following this initial processing stage, the data undergoes further refinement through the Natural Language Toolkit (NLTK) for data preprocessing. Here, any answers exceeding the designated length are pruned, while shorter responses are padded with zeros to align them with the desired vector length. Subsequently, the preprocessed data is directed into a Long Short-Term Memory (LSTM) model, where it undergoes grading for accurate assessment and evaluation.

D. Case Study

Upon analyzing the models performance with the test set, The accuracy rate approaches 80%. Recognizing the potential for improvement, it is identified that the enhancement of training data is a key factor. Acknowledging the impact of a diverse and expansive dataset, It is aimed to bolster the models learning by incorporating additional varied samples into the training set. This strategic approach to augmenting the training data serves as a proactive measure to further refine the models accuracy and robustness, ensuring its adaptability to a broader range of handwriting styles and scenarios. Through this iterative process, it anticipates a subsequent enhancement in overall model performance, aligning to continually refine and optimize accuracy in the recognition and evaluation of handwritten text.

E. Results

The model's development involved exploring various approaches by adjusting parameters such as deep layers, number of neurons, activation functions, and incorporating bidirectional LSTM layers. Through iterative tuning of these parameters and experimenting with different layer configurations, we aimed to identify the most efficient and optimal model architecture.

Our analysis included rigorous testing of the model's performance using a separate test dataset, where we achieved an accuracy rate approaching 80

It's worth noting that developing a model capable of accurately grading longer texts (e.g., 200-250 words) containing figures and equations requires a more sophisticated level of analysis and study. We are actively pursuing this goal with the aspiration of creating a model that can match the expertise of human graders.

IV. An automated essay evaluation system using natural language processing and sentiment analysis[3]

A. Authors

Vijaya Shetty Sadanand1, Kadagathur Raghavendra Rao Guruvyas1, Pranav Prashantha Patil1 ,Jeevan Janardhan Acharya1, Sharvani Gunakimath Suryakanth2 1Department of Computer Science and Engineering, Nitte Meenakshi Institute of Technology, Bengaluru, India 2Department of Computer Science and Engineering, RV College of Engineering, Bengaluru, India

B. Overview

This study introduces an innovative approach to automate the grading process of essays by employing a Long Short-Term Memory (LSTM) model within Natural Language Processing (NLP) techniques. The system aims to improve the efficiency and effectiveness of essay grading by leveraging NLP for feature extraction and sentiment analysis.

The dataset used for training the model comprises 12,000 essays obtained from the Hewlett Foundation, with each essay being evaluated by two manual graders. Additionally, sentiment analysis is performed using a Twitter dataset containing 10,000 tweets categorized as positive or negative.

The LSTM model architecture consists of two LSTM layers, a dropout layer, and a densely connected layer with rectified linear unit (ReLU) activation function. The model is trained using the mean squared error loss function and RMSprop optimizer. Grammar, spelling, and syntactical errors are detected using language tools, while plagiarism detection involves comparing the essay with web sources.

The grading process involves pre-processing the essay, detecting errors, and converting the text into a word vector format. Sentiment analysis and plagiarism detection results are then incorporated into the grading criteria, ensuring a comprehensive assessment of each essay.

C. Implementation

The system's approach to essay grading is comprehensive and multifaceted, drawing upon various natural language processing techniques to ensure a fair and accurate evaluation process. By analyzing the content, structure, and syntax of essays, it aims to provide unbiased assessments. Utilizing sentiment analysis, the system evaluates the overall tone of the essays, distinguishing between positive and negative sentiments expressed within the text. Additionally, the integration of plagiarism detection mechanisms enables the identification of any instances of copied content, offering transparency by providing detailed reports including source URLs and the percentage of plagiarized material. Central to its functionality is an LSTM network, meticulously trained on a diverse dataset of 12,000 essays using a K-fold mechanism for cross-validation, ensuring robustness and reliability. Augmenting its capabilities, a nave Bayes classifier assists in discerning the emotional nuances present in the essays. Furthermore, the system's ability to identify and rectify spelling and grammar errors empowers writers with insightful feedback for refinement. Through this holistic approach, the system endeavours to uphold standards of integrity and objectivity in the grading process while fostering continuous improvement inwritingskills.

D. Case Study

The paper contributes significantly to automated essay evaluation by introducing a novel system that integrates Natural Language Processing and sentiment analysis. This approach offers a comprehensive evaluation by addressing both structural and subjective aspects. The inclusion of sentiment analysis adds a valuable layer, categorizing essays as positive or negative. Plagiarism detection surpasses conventional methods by providing detailed information such as URLs and a percentage breakdown of matched material. Additionally, the robust training of the LSTM network on a dataset of 12,000 essays ensures adaptability and effectiveness. The systems ability to identify spelling and grammar errors enhances its completeness, setting a new standard in automated gradingmethodologies.

E. Results

The model presented here combines NLP and machine learning to streamline grading processes in educational settings. Utilizing an LSTM network, trained on 12,000 essays graded by two manual graders, achieved a high average QWK score of 0.911. Additionally, a Nave Bayes classifier accurately determines the sentiment of student essays with 99.4

The system identifies syntactic errors and provides detailed feedback to students, while a plagiarism detector flags any copied content from web sources. Factors including errors, plagiarism, and essay quality contribute to final grades. Continual training using newly submitted essays enhances system accuracy.

Future improvements may involve evaluating English fluency based on language style and leveraging feedback for further enhancements.

V. Subjective Answers Evaluation Using Machine Learning and Natural Language Processing[4]

A. Authors

FARRUKH BASHIR1, HAMZA ARSHAD1, ABDUL REHMAN JAVED2*, NATALIA KRYVINSKA3*, SHAHAB S. BAND4.

B. Overview

Assessing subjective papers manually is both challenging and labour-intensive. Key obstacles include insufficient comprehension of the data and the complexities of integrating Artificial Intelligence (AI) for analysis. While numerous efforts have been made to automate the scoring of students' answers using computer science, most rely on conventional counting methods or specific word usage patterns. Moreover, there's a notable scarcity of meticulously curated datasets in this domain. This paper proposes an innovative approach that harnesses a variety of machine learning and natural language processing techniques, alongside tools such as Wordnet, Word2vec, word movers distance (WMD), cosine similarity, multinomial naive Bayes (MNB), and term frequency-inverse document frequency (TF-IDF), to automate the evaluation of descriptive answers. By leveraging solution statements and keywords, answers are evaluated, and a machine learning model is trained to predict grades. Results indicate that WMD outperforms cosine similarity overall, and with adequate training, the machine learning model could function independently. Experimental findings demonstrate an accuracy rate of 88% without the MNB model, which is further improved by 1.3% with its inclusion.

C. Implementation

Cosine similarity and word movers distance were utilized for assessing similarity. The approach involved a two-step procedure: an initial evaluation based on solutions and keywords, succeeded by training a machine learning model for grading. A specialized dataset, curated by field experts, was employed for this purpose. Machine learning techniques facilitated automatic evaluation, with diverse metrics employed to compare models. The study demonstrated the superiority of word movers distance over cosine similarity, achieving an 88% accuracy rate without multinomial naive Bayes (MNB) and reducing errors by 1.3 per cent with MNB. This approach offers the potential to significantly diminish manual workload, thereby affording educators more time for teaching preparation.

D. Case Study

The study presented in the paper makes a substantial contribution to the field by providing empirical evidence that the word2vec approach consistently outperforms traditional word embedding techniques in terms of semantic preservation. Through comprehensive analysis and experimentation, the research showcases the advancement achieved with word2vec, highlighting its superior ability to accurately represent language semantics across various contexts and datasets. This insight holds significant implications for a wide range of applications within natural language processing, offering valuable guidance and direction to researchers and practitioners seeking to leverage cutting-edge methods for language representation and analysis. By demonstrating the efficacy of word2vec in capturing nuanced semantic relationships, the study not only enhances our understanding of computational linguistics but also facilitates the development of more robust and effective language processing models and systems. As such, this research serves as a foundational pillar for advancing the state-of-the-art in language understanding and lays the groundwork for future innovation in the field.

E. Result

That the word movers distance (WMD) exhibits superior performance compared to cosine similarity when automatically evaluating descriptive answers. Moreover, the machine learning model achieves an impressive accuracy rate of 88% even without the incorporation of the multinomial naive Bayes (MNB) model. With the inclusion of the MNB model, the error rate is further reduced by 1.3 per cent, indicating the effectiveness of this additional technique. These findings suggest that the proposed approach has the potential to significantly reduce the manual labour associated with routine tasks, thereby freeing up valuable time for educators to focus on teaching and curriculum preparation. By streamlining the assessment process, educators can devote more energy to enhancing the quality of instruction and developing innovative teaching strategies, ultimately leading to an enriched learning experience for students.

VI. A Keyword Based Technique to Evaluate Broad Question Answer Script[5]

A. Author

Tamim Al Mahmud,190179526@aston.ac.uk, Aston University, Birmingham, United Kingdom, Md Gulzar Hussain,gulzar.ace@gmail.com, Green University of Bangladesh, Dhaka, Bangladesh, Sumaiya Kabir,sumaiya@cse.green.edu.bd,Green University of Bangladesh, Dhaka, Bangladesh

B. Overview

An innovative solution for electronically evaluating subjective answer scripts, aiming to enhance the efficiency and accuracy of educational assessment processes. By integrating various functionalities, the system effectively examines and assesses written answer scripts. Key features include keyword identification, comparison with parsed keywords from open and closed domains, and detection of grammatical and spelling errors. Through experimentation with 100 student answer scripts, the proposed system achieved a precision score of 0.91, indicating its effectiveness in electronic evaluation. Overall, this solution offers a promising approach to streamline and improve the assessment of subjective responses in educational settings.

C. Implementation

In the proposed system for answer script evaluation, linguistic analysis plays a pivotal role, and three key linguistic analysis tools are employed: JLanguageTool, Perfect Tense API, and Grammar Bot API. JLanguageTool serves as a robust language-checking tool, capable of identifying grammatical errors, stylistic issues, and language-related inconsistencies within the text. Perfect Tense API supplements this by offering advanced grammar and spelling checks, ensuring a comprehensive linguistic analysis. Additionally, the Grammar Bot API contributes by providing further insights into grammatical structures, aiding in the identification of potential errors and language enhancements.

The answer analysis component utilizes the MediaWiki Action API, a web service designed to provide access to various features of a wiki. This API facilitates the extraction of valuable information from the answer scripts, allowing for a deeper understanding of the content and its alignment with expected criteria. The integration of the MediaWiki Action API ensures a dynamic and adaptable approach to answer analysis, allowing for a more nuanced evaluation.

After performing linguistic analysis and answer analysis, the system compares the scores generated from these analyses. This comparative evaluation enables a holistic assessment of the answer script, incorporating linguistic nuances, grammatical accuracy, and contextual relevance. The scores provide valuable insights into the strengths and weaknesses of the response, guiding the overall grading process. This multifaceted approach not only enhances the precision of the evaluation but also contributes to a more comprehensive understanding of the quality of the answers, emphasizing the importance of linguistic and contextual analysis in the assessment of written responses

D. Case Study

The proposed system aims to tackle the challenges of evaluating a large number- of student answer scripts by introducing a hassle-free and cost-effective approach. Conducting exams for numerous students is traditionally costly and time-consuming. However, this system minimizes both time and cost by leveraging automation in the evaluation process, significantly reducing the need for extensive human involvement and physical resources. Moreover, it serves as a proactive measure to decrease human grader errors, ensuring a consistent and objective assessment through automated mechanisms. By streamlining the assessment process, the proposed system not only addresses logistical constraints but also enhances the overall efficiency, fairness, and reliability of large-scale exam evaluations in educational settings

E. Result

The automated exam evaluation system demonstrates impressive performance metrics, as evidenced by its precision score of 0.91. This score reflects its high accuracy in correctly identifying and scoring relevant content within answer scripts. Additionally, with a recall score of 0.81, the system effectively captures a substantial proportion of relevant information from the answer scripts, indicating its ability to comprehensively assess student responses. The balanced F-score of 0.87 further underscores the system's robust performance, considering both precision and recall. This balanced measure highlights the system's capability to provide accurate and comprehensive evaluations of descriptive answers, contributing to fair and reliable assessment outcomes in educational settings.

VII. CONCLUSIONS

The development of a machine-learning-based solution for the automated evaluation of handwritten descriptive answers addresses the challenges inherent in manual grading processes. By leveraging techniques such as cosine similarity and semantic search, the system demonstrates promising results in accurately predicting answer grades.

Furthermore, the incorporation of Natural Language Processing and sentiment analysis enriches the evaluation process by providing a comprehensive assessment of both structural and subjective aspects of the responses. The system's ability to detect plagiarism, spelling, and grammar errors enhances its effectiveness and completeness, setting a new standard in automated grading methodologies.

Moreover, the proposed system aims to streamline the assessment process, minimizing both time and cost by reducing the need for extensive human involvement and physical resources. This proactive approach not only addresses logistical constraints but also ensures a consistent and objective assessment through automated mechanisms, thereby enhancing the overall efficiency, fairness, and reliability of large-scale exam evaluations in educational settings.

REFERENCES

[1]

[2] M. Agarwal, R. Kalia, V. Bahel and A. Thomas, ”AutoEval: A NLP Approach for Automatic Test Evaluation System,” 2021 IEEE 4th International Conference on Computing, Power and Communication Technologies (GUCON), Kuala Lumpur, Malaysia, 2021, pp. 1-6, doi: 10.1109/ GUCON50781.2021.9573769.

[3] Sadanand, Vijaya Shetty, et al. ”An automated essay evaluation system using natural language processing and sentiment analysis.” International Journal of Electrical & Computer Engineering (2088-8708) 12.6 (2022).

[4] Tulu, Cagatay Neftali, Ozge Ozkaya, and Umut Orhan. ”Automatic short answer grading with semspace sense vectors and malstm.” IEEE Access 9 (2021): 19270-19280.

[5] Choi, Donghyun, et al. ”Adaptive batch scheduling for open-domain question answering.” IEEE Access 9 (2021): 112097-112103.

[6] Dhandapani, Aarthi, and Viswanathan Vadivel. ”Question answering system over semantic web.” IEEE Access 9 (2021): 46900-46910.

[7] Rahman, Md Motiur, and Ferdusee Akter. ”An Automated Approach for Answer Script Evaluation Using Natural Language Processing.” IJC-SET 9 (2019): 39-47.

[8] Senapati, Deepak, et al. ”Descriptive Indic Answer Script Evaluation Using Deep Learning.” Innovations in Electrical and Electronic Engineering: Proceedings of ICEEE 2021. Singapore: Springer Singapore, 2021. 901-912.

[9] Al Mahmud, Tamim, et al. ”A keyword-based technique to evaluate broad question answer script.” Proceedings of the 2020 9th International Conference on Software and Computer Applications. 2020.