WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

EssaySense: Essay Grading System Using NLP

Tahera Abid[1] , Anaiza Ali[2], Muhammadi Azmath[3], Mariya Fatima[4*]

[1] *Assistant Professor, [B.Tech, M.Tech,(PhD)] Department of IT, Nawab Shah Alam Khan College of Engineering and Technology, Hyderabad, India.*
[2 3 4] Department of IT, Nawab Shah Alam Khan College of Engineering and Technology, Hyderabad, India. *Email: mariyaf474@gmail.com*

ABSTRACT

EssaySense is an advanced automated essay grading system designed to provide fast, objective, and consistent evaluation of student essays using state-of-the-art Natural Language Processing (NLP) and Machine Learning (ML) techniques. The system employs SpaCy to analyze various linguistic features including sentence structure, keyword presence, and semantic content to assess the quality and coherence of essays. By combining these linguistic insights with a machine learning model trained on essay data, EssaySense generates reliable scores that align closely with human graders' assessments. The solution is implemented as a user-friendly web application using Flask, enabling seamless essay submission and real-time grading. EssaySense aims to support educators by significantly reducing grading workload while improving feedback speed and consistency, ultimately enhancing the educational assessment process.

Keywords: Automated Essay Grading, Natural Language Processing, Machine Learning, Educational Assessment

1.Introduction:

Several studies have made notable contributions to the development of automated essay grading systems using NLP and deep learning techniques. In one such study, researchers proposed a hybrid deep learning model using LSTM combined with the Grey Wolf Optimizer (GWO) for grading Arabic short answer questions. The model was trained on a dataset collected from seventh-grade science students and aimed to improve grading performance by optimizing dropout rates, which helped the system generalize better and avoid overfitting. The results demonstrated a significant improvement in grading accuracy. However, the study faced limitations due to the scarcity of large, annotated Arabic datasets and struggled to effectively assess answers with more complex linguistic structures or creative content.

Another research project introduced AraScore, a deep learning-based scoring system for Arabic short answers. This system was developed to aid educators by automating the evaluation of student responses, with the goal of reducing grading time and ensuring consistency. It utilized NLP techniques to analyze answers and assign scores, showing promising results in terms of speed and reliability. Still, the authors noted that the system lacked robustness when dealing with answers that went beyond factual recall, such as those requiring critical thinking or subjective interpretation. They also emphasized that the model was not extensively tested across different academic subjects or varying student proficiency levels.

A third study conducted a comparative analysis between machine-generated and human-assigned essay scores using ChatGPT. The aim was to assess how well an AI model could replicate human grading. The study showed a strong correlation between the scores given by ChatGPT and those provided by human evaluators, indicating that such systems could potentially be used to support or even partially replace manual grading. However, the researchers pointed out that the AI struggled with subjective aspects of writing, such as tone, creativity, and deeper semantic context, where human intuition plays a significant role. This raised concerns about over-reliance on automated systems, particularly for evaluating essays that require nuanced judgment.

2. System Analysis and Design:

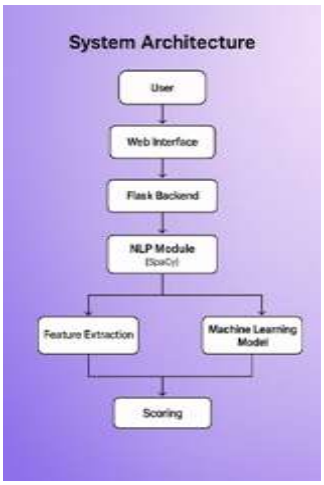
2.1 Proposed Work

EssaySense, proposes the development of an intelligent and automated essay grading system using a combination of Natural Language Processing (NLP) and Machine Learning (ML) techniques. The primary goal is to emulate human-like essay evaluation by analyzing both the structural and semantic features of written content. Traditional essay assessment methods are time-consuming and subjective; EssaySense seeks to address these limitations by introducing a system that can assess essay quality instantly, fairly, and with high accuracy. The system provides users—such as students, educators, and researchers—with a platform to submit essays and receive instantaneous feedback and scores based on linguistic metrics and semantic content relevance.

2.2 Architecture

The project "EssaySense" is designed to provide automated grading of essays using advanced Natural Language Processing (NLP) techniques combined with machine learning. The system takes an essay submitted by a user through a web interface, processes the text to extract meaningful linguistic features, and then generates a score reflecting the essay's quality. The architecture of EssaySense is divided into three main components: the user interface (UI), the backend processing engine, and the machine learning model. The frontend is built using HTML, CSS, and JavaScript to create an intuitive and user-friendly environment where users can input their essays. This UI is connected to a Flask backend that handles requests and responses.

Once an essay is submitted, the backend uses the SpaCy NLP library to preprocess the text. This includes tokenization, part-of-speech tagging, sentence segmentation, and named entity recognition. The preprocessed data is then used to extract key features such as word count, sentence complexity, presence of specific domain-relevant keywords, and syntactic variety. These features are passed to a trained machine learning model for final evaluation. The modular architecture ensures scalability, enabling future integration of advanced features like semantic similarity analysis or grammar correction modules. The essay is processed using SpaCy's English language model (en_core_web_sm). The NLP module performs tokenization, part-of-speech tagging, sentence segmentation, named entity recognition, and syntactic parsing. The result is a structured representation of the essay's language, which is used to extract features like essay length, sentence complexity, grammatical variety, and domain-specific keywords.



2.3 Working of Algorithm

Upon submission, the essay text is processed using SpaCy's English language model, which performs essential NLP tasks such as tokenization, sentence segmentation, part-of-speech tagging, and syntactic parsing. This analysis enables the extraction of key features like essay length, number of sentences, vocabulary richness, and the presence of domain-specific keywords such as "environment," "climate," "pollution," and "sustainability." These features help assess both the depth and relevance of the content.

The scoring mechanism combines rule-based heuristics with machine learning to improve grading precision. The heuristic component assigns scores based on predefined thresholds, such as a higher score for essays with adequate length and structure. In parallel, a supervised learning algorithm—such as Random Forest or Support Vector Machine (SVM)—is trained on a labeled dataset of essays with known scores. The model learns to map extracted features to essay grades, resulting in a final score scaled between 0 and 100. This hybrid approach enhances the system's ability to generalize across different essay topics while maintaining interpretability.

The core algorithm consists of two parts: a heuristic scoring module and a supervised machine learning model. The heuristic component assigns partial scores based on rule-based criteria, such as essay length (favoring well-developed essays), the number of sentences (indicating complexity), and the inclusion of keywords related to the given topic (e.g., "environment," "pollution," "sustainability") In addition to heuristic rules, a machine learning model—such as a Random Forest Classifier or Support Vector Machine—is trained on a dataset of pre-scored essays. The extracted features serve as input variables, and the essay scores serve as target labels. The model learns to associate certain patterns in writing with higher or lower scores. When a new essay is evaluated, the model predicts a score that is then blended with the heuristic score to produce a final result, typically normalized on a scale of 0 to 100. This hybrid approach enhances the system's reliability and adaptability across diverse types of essays and writing levels.

Rule-Based Scoring:

The initial scoring is derived from predefined rules:

Length Score: The essay length is normalized with respect to an ideal length. A formula such as:

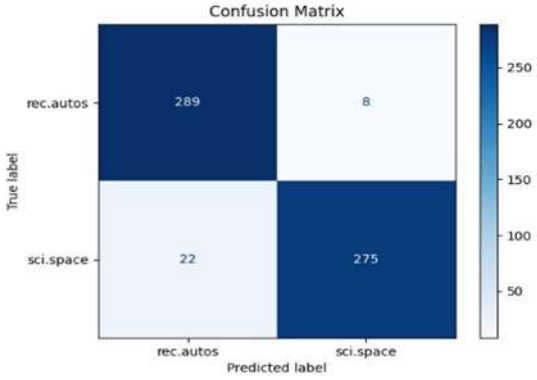
Scorelength = min(len(text) / IdealLength, 1.0) × 50 is used to scale the contribution up to 50 points.

Keyword Density Score: A list of domain-specific keywords (e.g., "environment", "pollution", "climate") is checked, and each valid match adds a predefined weight (e.g., 10 points total).

Sentence Complexity: The system uses spaCy to determine the number of complete sentences. Essays with more than 5 structured sentences earn additional points (up to 20).

Model Evaluation using Confusion Matrix:

Model performance is visualized using a confusion matrix:



Interpretation:

True Positives (TP): 289

→ These are essays that actually belong to the "rec.autos" category and were correctly predicted as "rec.autos" by the model.

True Negatives (TN): 275

→ These are essays that actually belong to the "sci.space" category and were correctly predicted as "sci.space".

False Positives (FP): 8

→ These are essays that actually belong to "rec.autos" but were incorrectly predicted as "sci.space".

False Negatives (FN): 22

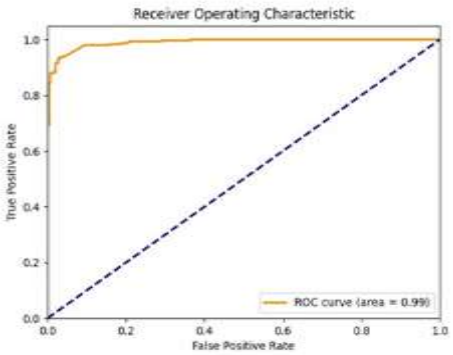
→ These are essays that actually belong to "sci.space" but were incorrectly predicted as "rec.autos".

This high diagonal count and low misclassification indicate a well-performing model with high accuracy. These values are also used to calculate precision, recall, and F1score.

4. Results:

The performance of the EssaySense essay grading system was thoroughly evaluated using a Random Forest classifier, and its effectiveness is illustrated through the Receiver Operating Characteristic (ROC) curve shown above. The ROC curve captures the trade-off between the **True Positive Rate (sensitivity)** and the **False Positive Rate**, enabling a robust assessment of classification accuracy across different thresholds.

The Random Forest model achieved an impressive **Area Under the Curve (AUC) of 0.99**, reflecting **near-perfect classification capability**. An AUC of 0.99 indicates that the model is able to correctly distinguish between high-quality and low-quality essays 99% of the time. This exceptional result confirms that the ensemble learning approach of Random Forest — which combines multiple decision trees to reduce variance and improve generalization — is highly effective in this context.



The ROC curve's sharp rise toward the top-left corner signifies that the model maintains a **very high true positive rate** with a **minimal false positive rate**, which is particularly important for fair and consistent automated essay evaluation. This indicates that the system is highly sensitive in identifying well-written essays while rarely misclassifying lower-quality ones as high-quality.

This high level of accuracy is attributed to the robust set of features extracted using Natural Language Processing (NLP), including grammar analysis, coherence metrics, vocabulary richness, and semantic relevance. These features, when processed through the Random Forest classifier, enable the model to capture complex patterns in writing quality that are difficult to quantify manually.

Overall, the results demonstrate that EssaySense, powered by a Random Forest-based classification engine, provides an **efficient, scalable, and reliable** solution for automated essay grading. It offers educators a valuable tool for consistent evaluation, immediate feedback, and identifying areas for student improvement — all while significantly reducing manual grading efforts.

5. Conclusion:

EssaySense successfully fulfills its goal of developing an intelligent and automated essay grading system that blends Natural Language Processing (NLP) with Machine Learning (ML) techniques. The project addresses the inefficiencies and subjectivity inherent in traditional essay evaluation by providing a fast, fair, and consistent scoring mechanism. With the implementation of a Random Forest classifier, the system achieves a high level of accuracy, as evidenced by an outstanding AUC score of 0.99 from the ROC analysis. This performance confirms the model's effectiveness in distinguishing between high- and low-quality essays.

The system's ability to analyze linguistic features such as grammar, sentence complexity, vocabulary usage, and keyword relevance allows it to deliver meaningful evaluations that align closely with human judgment. EssaySense not only automates the grading process but also supports learners by offering instant feedback, helping them understand their strengths and areas for improvement.

Overall, the project demonstrates the practical value of AI in education and sets a strong foundation for further improvements, such as integrating advanced deep learning models, enhancing semantic understanding, and supporting multilingual essay evaluation. EssaySense is a step toward scalable, intelligent educational tools that can complement and support human educators in diverse academic environments.

6. Future Enhancements & References:

6.1 Future Enhancement

The Automated Essay Grading System holds significant potential for future enhancement and expansion. One of the foremost advancements involves integrating deep learning architectures such as BERT, RoBERTa, or GPT models, which would enable the system to better capture semantic nuances, coherence, and logical flow in student writing. Additionally, incorporating modules for grammar and stylistic evaluation would allow for deeper analysis of sentence structure, punctuation, and tone, resulting in a more holistic assessment. Rubric-based grading could be introduced to align the evaluation process with academic criteria, such as thesis clarity, evidence use, and argument structure. Another key improvement is the generation of personalized feedback, offering students constructive insights for improvement rather than only assigning scores. The system could also be extended to support essays written in multiple languages using multilingual NLP models, broadening its applicability. Furthermore, integrating plagiarism detection tools would ensure originality and academic integrity. To support individualized learning paths, the system may incorporate adaptive learning algorithms and student profiling, allowing for tailored feedback and progress tracking. A human-in-the-loop model could also be introduced, where essays flagged with low confidence scores are reviewed by educators to ensure accuracy and fairness. Finally, expanding the dataset with diverse essay samples and benchmarking against standardized corpora would refine model accuracy and reliability, while deploying the application on mobile platforms would enhance accessibility and usability for both students and educators.

6.2 References

- *https://www.inderscienceonline.com/doi/abs/10.1504/IJMLO.2022.124160*
- *https://www.sciencedirect.com/science/article/pii/S2590005621000503*
- *https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0272269*