Data Engineering and Applications

Review Paper

Automated Essay Scoring via NLP: System Architectures, Feature Engineering, and Evaluation Metrics

Kinga Wojcik[1, *]

[1] Faculty of Computer Science and Information Systems, University of Odz, 90-136 Odz, Poland
*Corresponding author: k.wojcik@uni.lodz.pl

Abstract. With the continuous advancement of natural language processing (NLP) technology, automatic essay scoring systems have emerged as a prominent research focus in the field of education. This paper first introduces the fundamentals of NLP and its applications in text analysis, including part-of-speech tagging, named entity recognition, and text classification. It then elaborates on the architecture of automatic essay scoring systems, covering overall system design, preprocessing modules, feature extraction modules, and scoring model modules. Subsequently, it delves into evaluation methods for essay grading systems, including metric frameworks and experimental designs. This paper aims to provide comprehensive guidance for researchers in related fields, advancing the application of natural language processing in essay grading.

Keywords: *Natural Language Processing; Essay Grading; Text Analysis; Evaluation Methods; System Architecture*

Received on 24 September 2024, Accepted on 29 December 2024, Published on 5 Jan2025

Introduction

As a crucial measure of students' writing proficiency and linguistic expression, essay assessment holds a pivotal position in education [1-3]. While traditional manual grading played a significant role for a period, its shortcomings have become increasingly apparent amid expanding educational scales and growing educational demands [4]. Manual grading is highly subjective, with varying evaluation criteria among teachers leading to inconsistent and non-objective scoring outcomes [5]. Additionally, manual grading is time-consuming and inefficient, requiring teachers to expend significant effort on sentence-by-sentence corrections. This hinders timely feedback delivery to students, adversely affecting teaching effectiveness and student motivation [6].

The emergence of natural language processing (NLP) technology offers a new solution to these challenges [7]. NLP, a vital branch of computer science and artificial intelligence, focuses on enabling computers to understand, generate, and process human language [8]. As NLP technology advances, its applications in education have expanded, giving rise to automated essay grading systems [9]. These systems enable rapid, objective essay grading, providing immediate feedback to both teachers and students, thereby significantly enhancing the efficiency and objectivity of essay evaluation [10]. Through automated scoring, teachers can allocate more time and energy to instructional design and personalized tutoring, while students gain timely insights into their writing proficiency and areas for improvement, allowing for targeted refinement and advancement [11-13].

Currently, numerous scholars worldwide have conducted extensive research in the field of automatic essay scoring, employing diverse methodologies that have propelled rapid advancements in this domain [14]. Early automatic essay scoring systems primarily relied on manually extracted features such as lexical richness, grammatical structure, and sentence length [15]. Reference [16] analyzed a large corpus of exemplary and average essays to identify a set of feature indicators correlated with writing quality, then integrated machine learning algorithms to establish a scoring model. While this approach effectively evaluates basic writing

DOI: 10.64972/dea.2025.v3i1.32

proficiency, it faces limitations in understanding and assessing semantic-level content. In recent years, driven by the rapid advancement of deep learning technologies, applications of deep neural networks (DNNs) [17], convolutional neural networks (CNNs) [18], recurrent neural networks (RNNs) [19], and their variants have increasingly emerged in essay grading [20]. These deep learning models can automatically learn effective feature representations from large text datasets without manual feature extraction, overcoming the shortcomings of traditional methods [21].

However, despite significant progress in the field of automatic essay scoring, several pressing issues and challenges remain. First, essays in different languages and styles pose high demands on the adaptability of scoring models [22]; Second, enhancing the interpretability of scoring models remains a critical challenge [23]. Additionally, optimizing system efficiency and performance when processing large-scale essay datasets is essential to ensure the real-time capability and usability of scoring systems [24].

This paper provides a comprehensive review of natural language processing-driven automatic essay scoring systems, aiming to offer systematic references and guidance for researchers in related fields and to advance the development of this domain. The contributions of this paper are as follows: (1) It systematically reviews the current application status of natural language processing techniques in the field of essay automatic scoring; (2) It provides a detailed analysis of the advantages and disadvantages of various essay automatic scoring system architectures; (3) It thoroughly explores the scientific rigor and effectiveness of evaluation methods for essay automatic scoring systems; (4) It analyzes the challenges faced by essay automatic scoring systems and offers reasonable predictions for future development trends.

Natural Language Processing Technology

Fundamentals of Natural Language Processing

Natural language processing has emerged as a crucial enabler for a wide range of intelligent language applications, bridging the gap between human communication and computational understanding. Its interdisciplinary nature draws upon principles from linguistics, computer science, and cognitive psychology, allowing researchers and practitioners to develop algorithms that interpret, manipulate, and generate natural language data. As educational, commercial, and social contexts increasingly demand interactive and adaptive language technologies, NLP stands at the forefront of efforts to make computers more responsive to human needs and intentions.

Natural Language Processing (NLP) [25-26] is a vital branch of computer science and artificial intelligence, focusing on enabling computers to understand, generate, and process human language. Its foundations encompass linguistic knowledge and text processing techniques. Linguistic knowledge covers lexicology, syntax, and semantics [27], providing theoretical support for computers to understand language structure and meaning. Text processing techniques include steps such as text cleaning, word segmentation, and stem extraction, preparing data for subsequent processing.

The methods and tools developed within the NLP domain are essential for a broad spectrum of downstream tasks, including information extraction, text summarization, sentiment analysis, and dialogue systems. By systematically combining linguistic theory with statistical and computational models, NLP facilitates both the analysis of complex language phenomena and the automation of language-based reasoning. This integration of theory and technology enables the transformation of unstructured text into structured knowledge, laying the groundwork for advanced applications such as automated essay scoring, intelligent tutoring, and adaptive learning platforms. As NLP continues to mature, its influence extends deeper into educational assessment, knowledge management, and human-computer interaction, driving further innovation across disciplines.

The fundamentals of natural language processing primarily encompass lexicology, syntax, and semantics [28], manifested as follows: 1) Lexicology primarily studies word structure, word formation, and inflection. In natural language processing, lexicological knowledge helps computers recognize and understand the basic meanings and usages of vocabulary, laying the foundation for tasks such as part-of-speech tagging and semantic analysis

DOI: 10.64972/dea.2025.v3i1.32

[29]. 2) Syntax investigates sentence structure and composition rules—how words combine into phrases, sentences, and larger textual units. Syntactic analysis is a core NLP task, aiming to determine sentence grammar, identify components like subjects, predicates, and objects, and establish hierarchical relationships between them [30]. 3) Semantics focuses on the meaning level of language, investigating the concepts, relationships, and contextual meanings expressed by words, phrases, and sentences. Semantic analysis aims to enable computers to grasp the true meaning of text, encompassing tasks like word sense disambiguation, semantic role labeling, and semantic relation identification [31].

Applications of Natural Language Processing in Text Analysis

The rapid development of natural language processing has significantly expanded the range of intelligent applications in educational assessment. By leveraging linguistic theory and computational algorithms, NLP enables computers to process and interpret human language at multiple levels, from lexical and syntactic analysis to semantic and pragmatic understanding. Within the context of automatic essay scoring, the integration of NLP techniques allows systems to go beyond surface-level text matching, providing deeper insights into writing quality and content relevance. These technologies empower automated systems to systematically analyze the structure, meaning, and communicative intent of student essays, supporting more objective and comprehensive evaluation processes. As a result, NLP serves as a critical foundation for advancing the automation of complex language assessment tasks and enhancing the overall fairness and consistency of educational evaluation.

Natural Language Processing has diverse applications in text analysis, primarily including part-of-speech tagging [32], named entity recognition [33], text classification [34], sentiment analysis [35], machine translation [36], and text generation [37]. A comparative analysis is presented in Table 1.

Table 1. Role and Development Directions of Natural Language Processing Technologies in Essay Scoring

Technology	Process	Role in Essay Scoring	Future Development Direction
Part-of-speech tagging	Text preprocessing feature extraction model training and tagging	Check grammatical errors evaluate vocabulary usage	Optimization by deep learning methods multilingual POS tagging
Named entity recognition	Preprocessing feature extraction model training and recognition	Assess use of proper nouns enrich content detection	Application of deep learning methods adaptation to multiple domains
Text classification	Preprocessing feature extraction model training and classification	Essay type identification targeted scoring	Fine-tuning deep learning models research on multi-label classification
Sentiment analysis	Preprocessing feature extraction model training and classification	Sentiment expression assessment enhance scoring dimensions	Deep learning for sentiment feature mining fine-grained sentiment analysis
Machine translation	Text analysis model training translation generation and optimization	Multilingual essay processing reference for semantic understanding	Optimization of neural machine translation exploration of multimodal translation
Text generation	Preprocessing model training text generation and optimization	Generate model essay for comparison logical structure assessment	Application of pre-trained language models improvement of generated text quality

The integration of multiple NLP techniques within essay assessment allows for a more holistic and nuanced evaluation of student writing. By leveraging these methods across the scoring pipeline, automated systems are able to capture a broad range of linguistic phenomena, from syntactic structure and terminology usage to sentiment and genre identification. This synergy of diverse NLP approaches not only elevates the precision of automated feedback but also lays the groundwork for adaptive and context-aware scoring mechanisms. Continuous advancements in these foundational technologies will further empower educators to deliver more timely, consistent, and pedagogically relevant assessment outcomes.

To further clarify how various natural language processing (NLP) technologies are integrated within automatic essay scoring systems, it is beneficial to examine the overall operational workflow. By visualizing the sequence and interactions among processes such as text preprocessing, feature extraction, scoring, and feedback, readers can gain a clearer understanding of the system's modular structure and the role each NLP component plays in the grading pipeline. This systematic overview helps to illustrate how foundational NLP tasks, including part-of-speech tagging, named entity recognition, and text classification, contribute to essay evaluation and support the broader goal of automated, objective, and consistent scoring. The main stages and their interconnections are depicted below in Figure 1.

DOI: 10.64972/dea.2025.v3i1.32

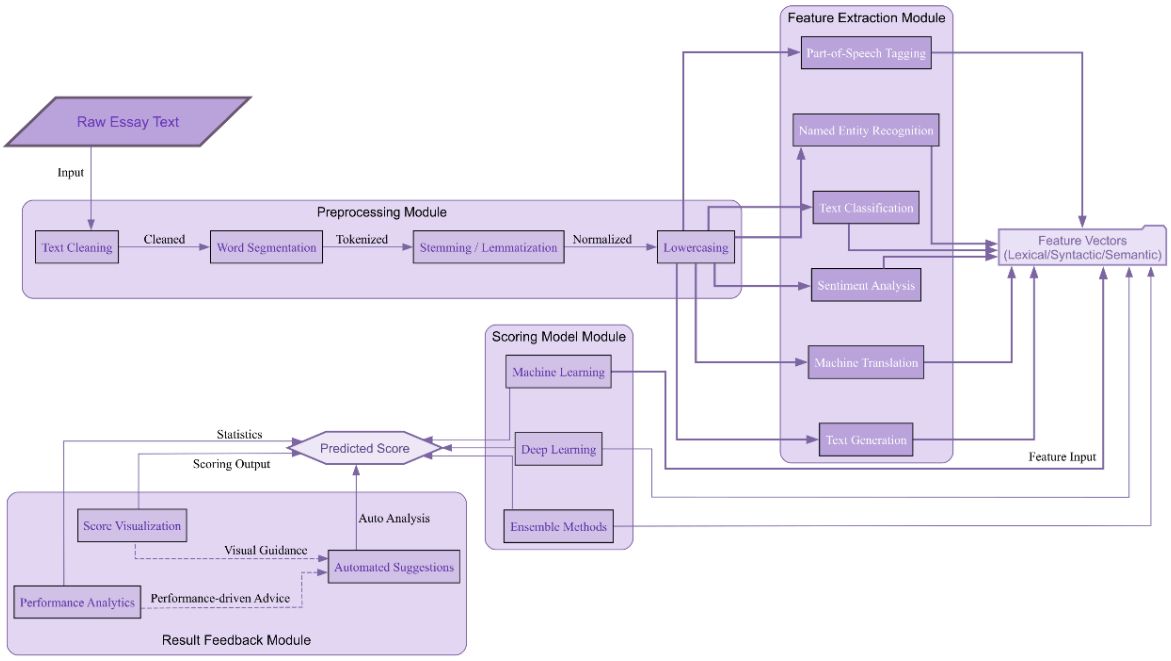

Figure 1. Example workflow of Natural Language Processing applications in automatic essay scoring systems.

The illustrated workflow encapsulates the integration of various NLP technologies within the essay grading pipeline, emphasizing the seamless interaction between preprocessing, feature extraction, scoring, and feedback modules. By visualizing these connections, the figure highlights how each NLP component contributes to the overall reliability and objectivity of automated scoring. This modular approach not only facilitates system scalability and maintenance but also enhances adaptability to diverse educational contexts. Understanding the interplay of these stages is essential for optimizing system performance and ensuring consistent, high-quality assessment outcomes.

Part-of-speech tagging involves annotating each word in a text with its corresponding part of speech, aiding in the analysis of sentence structure and grammatical relationships while providing a basis for grammar checks in essay scoring [32]. Within essay scoring systems, part-of-speech tagging enables the detection of grammatical errors such as improper noun-verb combinations or incorrect article usage [38]. Additionally, it enables analysis of the frequency and diversity of different parts of speech within essays, evaluating students' vocabulary utilization and linguistic richness [39]. Common part-of-speech tagging methods include rule-based approaches [40], statistical methods [41], and deep learning techniques [42]. Rule-based methods rely on predefined grammatical rules and dictionaries, making them suitable for specific languages and domains [40]; Statistical methods learn probabilistic relationships between parts of speech and context by training on annotated corpora, enabling adaptation to diverse text data [41]; Deep learning methods automatically learn contextual information and lexical features within text, achieving high annotation accuracy [42].

In addition to their core role in grammatical analysis, part-of-speech tagging techniques support a variety of downstream natural language processing tasks that enhance the capabilities of automated essay scoring systems. For instance, accurate POS tagging serves as a prerequisite for advanced syntactic parsing, which can further identify phrase boundaries, clause structures, and hierarchical dependencies within sentences. This deeper syntactic information enables systems to assess the complexity and sophistication of student writing, providing a more nuanced evaluation of language proficiency. Furthermore, POS tagging facilitates the extraction of linguistic patterns such as passive constructions, modal usage, and the deployment of cohesive devices, all of which contribute to the overall coherence and style of an essay. By integrating POS-based features with other linguistic indicators, automated scoring models can more effectively capture subtle aspects of writing quality, such as variation in sentence structure and the appropriate use of academic or descriptive language. The robustness and adaptability of POS tagging approaches also make them valuable for cross-linguistic applications,

DOI: 10.64972/dea.2025.v3i1.32

allowing systems to be extended to multiple languages with appropriate adjustments to tagging schemes and grammatical resources. Collectively, these capabilities underscore the foundational importance of part-of-speech tagging in the broader context of automated writing assessment and educational technology.

Named Entity Recognition (NER) identifies entity names within text [33]. In automated essay scoring, it can assess whether students use proper nouns correctly and appropriately, as well as whether such usage enriches essay content. By identifying named entities in essays, evaluators can assess students' correct usage and spelling of these entities, as well as whether relevant entities are appropriately referenced to enhance the essay's persuasiveness and content richness. Common NER methods include rule-based approaches [43], machine learning methods [44], and deep learning methods [45]. Rule-based methods match named entities by defining patterns and dictionaries, suitable for specific domains and languages [43]; Machine learning approaches such as Hidden Markov Models (HMM) [46] and Conditional Random Fields (CRF) [47] learn entity features and contextual relationships through annotated corpora; deep learning methods like Recurrent Neural Networks, Convolutional Neural Networks, and Transformer architectures can automatically extract contextual features from text, enabling end-to-end NER with superior performance [48].

Text classification categorizes text according to specific themes or categories [34]. In automated essay scoring, text classification technology helps systems rapidly and accurately identify essay types, enabling selection of appropriate scoring criteria and models for evaluation. Text classification methods primarily include traditional machine learning approaches [49] and deep learning methods [50]. Traditional machine learning approaches like Naive Bayes [51], Support Vector Machines [52], and Decision Trees [53] typically require manual feature extraction from text, such as bag-of-words models or TF-IDF. Deep learning methods can automatically learn semantic features and contextual relationships within text. By fine-tuning pre-trained language models, they achieve high classification accuracy [54].

Beyond the core function of categorizing essays, text classification also plays a pivotal role in enhancing the adaptability and scalability of automated scoring systems. By accurately recognizing essay genres—such as argumentative, narrative, or expository writing—classification modules can dynamically apply genre-specific rubrics and linguistic features for more precise evaluation. This targeted approach allows the system to account for the unique structural and stylistic conventions associated with different essay types, thereby improving the fairness and validity of score assignments. Furthermore, text classification assists in filtering out off-topic or irrelevant submissions, ensuring that only essays aligned with the prompt are subjected to detailed assessment. The flexibility of classification algorithms makes it feasible to accommodate new or evolving writing tasks with minimal manual intervention, supporting the ongoing evolution of educational assessment. By integrating text classification with other NLP components, automated essay scoring systems are better equipped to deliver nuanced feedback that addresses both content relevance and writing quality, ultimately supporting more effective learning outcomes.

Sentiment analysis examines the emotional orientation embedded in text, such as positive, negative, or neutral sentiments. In automated essay scoring, sentiment analysis helps evaluate the appropriateness and authenticity of students' emotional expressions within essays, as well as their alignment with the essay's theme and requirements [35]. Primary sentiment analysis methods include dictionary-based approaches, machine learning-based approaches [55], and deep learning-based approaches [56]. Dictionary-based methods utilize sentiment lexicon entries and semantic rules to compute text sentiment; machine learning-based methods train classification models like Naive Bayes or Support Vector Machines using annotated sentiment corpora; deep learning-based methods employ neural networks to automatically extract textual sentiment features, such as recurrent neural networks, convolutional neural networks, and pre-trained language models, demonstrating superior performance in handling complex textual sentiment.

Machine translation is a significant application area within natural language processing, aiming to automatically translate one natural language into another [36]. Early machine translation systems primarily relied on rules and dictionaries, translating through manually crafted translation rules and glossaries. However, this approach struggled to handle the complexity and polysemy inherent in language [57]. In recent years, Statistical Machine Translation (SMT) and Neural Machine Translation (NMT) have gradually become mainstream approaches [58]. Statistical Machine Translation analyzes large-scale bilingual corpora to learn statistical patterns between source

DOI: 10.64972/dea.2025.v3i1.32

and target languages, thereby achieving translation [57]. Neural Machine Translation, leveraging deep neural networks, directly maps source language sentences to target language sentences. This approach better captures semantic information and contextual relationships within text, producing more fluent and natural translations. When processing essays in different languages, techniques from machine translation—such as text semantic understanding and cross-lingual knowledge transfer—can be leveraged to enhance the adaptability and scoring accuracy of essay grading systems for multilingual compositions.

In addition to advancing the accuracy of translation, machine translation technologies have contributed to the broader integration of multilingual processing within educational assessment systems. By enabling the automated analysis of essays written in various languages, these technologies facilitate fairer and more inclusive evaluation practices, accommodating the linguistic diversity of students. Techniques such as word alignment, phrase-based translation, and cross-lingual embeddings allow essay scoring systems to identify equivalencies between expressions in different languages, thereby supporting consistent application of scoring criteria. Furthermore, the use of machine translation aids in the development of shared rubrics and benchmarks, promoting standardization in multilingual assessment environments. As a result, automated essay scoring platforms can provide meaningful and comparable feedback to learners regardless of their native language, supporting global educational objectives and fostering more equitable learning opportunities. This synergy between machine translation and automated assessment highlights the potential for natural language processing to bridge linguistic gaps and enhance the overall effectiveness of educational technology solutions.

Text generation can produce model essays or reference answers for comparison with student submissions, aiding scoring [37]. Applications include automated writing, dialogue systems, and text summarization. In automated essay scoring, text generation technologies can generate high-quality model essays or reference answers, providing standards and grounds for evaluation. Reference [59] demonstrates that comparing student essays with generated models can analyze strengths and weaknesses, such as content completeness, logical structure coherence, and linguistic accuracy.

Based on the above analysis, the advantages and disadvantages of applying natural language processing technologies are summarized in Table 2.

Table 2. Analysis of Advantages and Disadvantages in Applying Natural Language Processing Technologies

Technology	Application Case	Advantages	Limitations
Part-of-speech tagging	Grammar checking tools	Improve accuracy of grammar check provide vocabulary usage suggestions	Tagging errors may cause chain misunderstanding limited in semantic understanding
Named entity recognition	Information retrieval systems	Improve retrieval accuracy extract key entities	Easily confused with similar entities heavily influenced by domain-specific terms
Text classification	Automated news categorization	Efficient large-scale text processing automatic content organization	Classification errors reduce system reliability overly dependent on training data
Sentiment analysis	Customer review sentiment analysis	Gain deep insight into user sentiment trends support product improvement	May misjudge complex sentiment expression difficult to handle metaphor and irony
Machine translation	Multilingual document translation	Break language barriers promote cross-cultural communication	Translation quality varies difficult to handle culture-specific expressions
Text generation	Intelligent writing assistants	Improve writing efficiency inspire creativity	Generated text may lack depth and emotion risk of factual errors

In practical deployment, each NLP technology brings unique value to automated essay scoring while simultaneously presenting specific technical challenges. The application context often determines which strengths can be maximized and which weaknesses require mitigation strategies. For instance, while grammar checking tools can offer actionable insights for language learners, issues like semantic ambiguity and domain adaptation may require supplementary solutions. Understanding the interplay of these factors is essential for developing resilient systems that maintain high standards of reliability and educational fairness across varied use cases.

Based on the application of natural language processing technology, its applicable scenarios and effects are shown in Table 3.

DOI: 10.64972/dea.2025.v3i1.32

Table 3. Applicable Scenarios, Effects, and Challenges of Natural Language Processing Technology

Technology	Applicable Scenario	Effect	Challenge
Part-of-speech tagging	Grammar teaching text analysis	Accurate identification of parts of speech assist language learning	Difficult to handle ambiguous words maintain tagging consistency
Named entity recognition	Information extraction knowledge graph construction	Effective entity recognition support intelligent Q&A	Difficult to handle entity diversity and change cross-domain adaptation
Text classification	Text management public opinion analysis	Fast text classification mining information value	Difficult to handle class imbalance difficult to process new text categories
Sentiment analysis	Brand monitoring user research	Accurate grasp of sentiment trends guide decision-making	Difficult to handle complex multi-turn dialogue sentiment ensure cross-domain accuracy
Machine translation	Cross-language communication cultural transmission	Enable fast translation promote information sharing	Difficult to process professional terminology challenge to optimize translation fluency and accuracy
Text generation	Content creation summary writing	Improve creative efficiency achieve text diversity	Difficult to control quality of generated text avoid plagiarism ensure originality

The suitability of different NLP approaches depends heavily on the characteristics of the educational scenario and the objectives of assessment. By aligning technological capabilities with instructional needs, automated systems can facilitate more effective teaching and learning processes. However, developers must also account for the inherent complexity of natural language, ensuring that solutions remain robust to linguistic diversity, genre variation, and the evolving nature of written communication. Addressing these challenges is critical for achieving sustainable and scalable improvements in automated essay evaluation.

Among the statistical approaches, the Hidden Markov Model (HMM) has been widely used for part-of-speech tagging. The core idea is to find the most probable sequence of tags for a given sentence, which can be mathematically formulated as follows:

$$\arg\max_{T} P(W \mid T)P(T) \qquad (1)$$

where W represents the observed word sequence, and T denotes the possible tag sequence. This probabilistic framework enables the model to consider both the likelihood of the words given the tags and the prior probability of the tag sequence, thus improving tagging accuracy in practical applications.

Architecture of the Automatic Essay Grading System

Overall System Architecture Design

In the realm of educational technology, automated essay scoring systems have evolved to address the growing need for objective and timely assessment of written work. These systems are designed to replicate, as closely as possible, the nuanced judgment of human raters while minimizing subjectivity and inconsistencies. By leveraging advancements in natural language processing and machine learning, automated essay scoring platforms can efficiently process large volumes of student essays, ensuring consistent evaluation criteria across diverse subjects and grade levels. The development of such systems is underpinned by the recognition that effective essay assessment extends beyond mere grammar and spelling checks, requiring a comprehensive analysis of content relevance, logical coherence, and linguistic sophistication. To meet these demands, designers of automated grading platforms have adopted a modular approach, allowing for flexible integration of new analytical techniques and easier adaptation to varying educational standards.

The overall architecture design of the automatic essay grading system aims to build an efficient, accurate, and scalable grading platform. The core objective is to achieve efficient operation and convenient maintenance through modular design, which decomposes complex grading tasks into multiple relatively independent yet closely collaborating submodules [59]. A typical overall architecture of an automatic essay scoring system generally comprises preprocessing, feature extraction, scoring model, and result feedback modules [60]. The functional analysis of each module is presented in Table 4.

DOI: 10.64972/dea.2025.v3i1.32

Table 4. Module Functions of the Automatic Essay Scoring System

Module	Main Function	Input	Output
Preprocessing module	Standardize raw essay text including text cleaning word segmentation stem extraction	Raw essay text	Cleaned text word segmentation result stemming result
Feature extraction module	Extract valuable information from preprocessed text such as lexical grammatical semantic features	Cleaned text word segmentation result stemming result	Lexical feature vector grammatical feature vector semantic feature vector
Scoring model module	Use machine learning or deep learning algorithms to predict essay scores based on extracted features	Lexical feature vector grammatical feature vector semantic feature vector	Essay score
Result feedback module	Present scoring results and related analysis suggestions in an intuitive and understandable way	Essay score	Visualized scoring results analysis suggestions

A modular and systematic design underpins the reliability and scalability of automated essay scoring platforms. Each functional component contributes specialized expertise to the overall process, ensuring that data integrity is preserved from initial input through to the generation of final feedback. This architecture not only streamlines operational efficiency but also supports future enhancements, such as the incorporation of new language models or the adaptation to emerging educational standards. Ultimately, such a structure is crucial for delivering consistent, transparent, and actionable assessment at scale.

A comprehensive understanding of the automatic essay grading system requires not only a discussion of its individual modules but also an appreciation of how these modules interact as an integrated whole. By presenting the overall system architecture, it becomes evident how preprocessing, feature extraction, scoring, and result feedback modules collaborate to create an efficient and robust grading platform. Such an architectural overview facilitates insight into the logical flow of information and the dependencies among components, providing a framework for further analysis and optimization. The following diagram in Figure 2 offers a visual representation of the typical structure and functional relationships within an automatic essay scoring system.

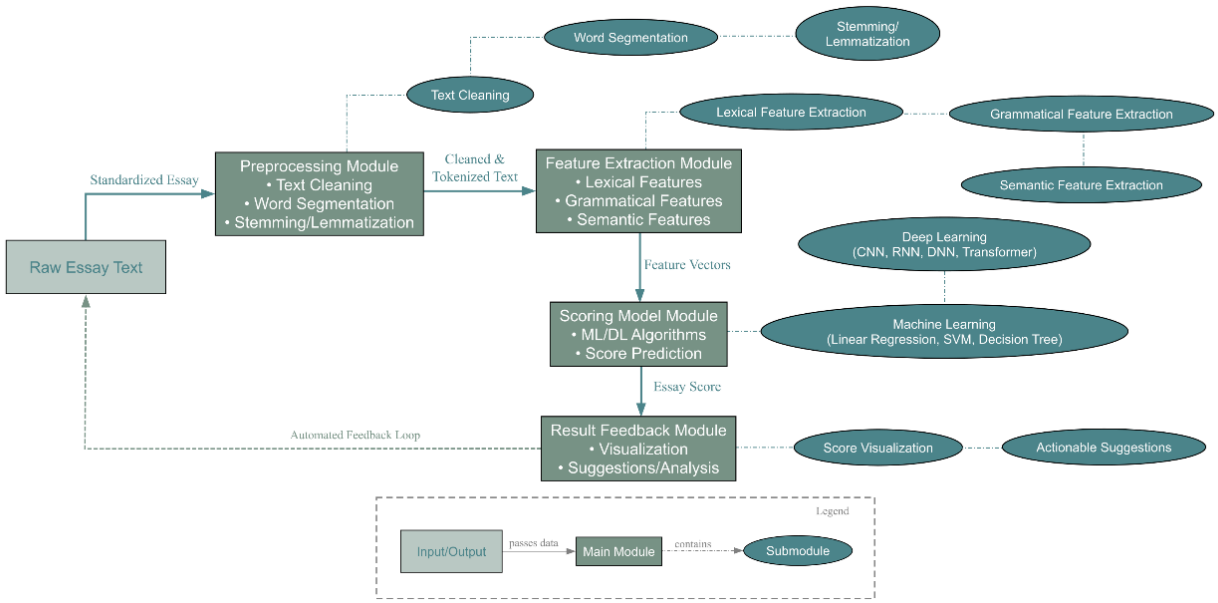

Figure 2. Overall architecture diagram of an automatic essay scoring system, illustrating main modules and their interactions.

This architectural diagram presents a modular perspective on system design, where each component operates both independently and collaboratively to achieve comprehensive essay evaluation. The structured flow of information—from initial text input to final feedback—demonstrates how data transformations underpin accurate scoring and insightful analysis. Such a layout enables the system to be flexible, supporting enhancements and integrations with emerging technologies. By mapping the dependencies and data flow,

DOI: 10.64972/dea.2025.v3i1.32

researchers and developers can more effectively identify bottlenecks and opportunities for improvement within the system.

Preprocessing Module

The preprocessing module serves as the foundational component of the automatic essay scoring system. Its function is to standardize and format raw essay texts, eliminating noise information and unifying text formats to create favorable conditions for subsequent feature extraction and analysis [61].

A robust preprocessing stage not only improves the clarity and uniformity of the input data but also reduces the risk of downstream errors in the analysis pipeline. By systematically addressing inconsistencies in the raw text—such as irregular spacing, mixed encodings, or non-standard punctuation—preprocessing ensures that all essays are evaluated on a level playing field. This normalization process is especially important when handling large and diverse essay corpora, where variations in input format can introduce bias or hinder the generalizability of the scoring model. As a result, careful design and implementation of the preprocessing module are crucial for maintaining the integrity and reliability of the entire automated assessment workflow.

Text cleaning is the primary task of the preprocessing module. It preserves pure text content by removing irrelevant characters, symbols, numbers, and redundant spaces [62]. Text cleaning operations are typically implemented using string processing techniques such as regular expressions, enabling rapid and accurate identification and processing of noisy segments within the text. Additionally, text vocabulary undergoes case unification processing, converting all letters to lowercase. This prevents issues like duplicate word counts caused by case differences, enhancing the accuracy and efficiency of subsequent text processing [63].

In practical applications, text cleaning must strike a careful balance between thoroughness and preservation of meaningful content. Overly aggressive cleaning may inadvertently remove valuable linguistic cues, such as stylistic markers or intentional use of punctuation, which could be relevant for evaluating writing quality. Therefore, it is essential to calibrate text cleaning rules to retain features that contribute to the assessment objectives, such as sentence boundaries and paragraph markers. Furthermore, adapting cleaning strategies to different genres or educational contexts can help optimize the preprocessing outcomes and support more nuanced downstream analysis.

Word segmentation is another critical task within the preprocessing module. It divides continuous text into discrete lexical units or phrases, forming the foundation for the vast majority of natural language processing tasks [64]. The quality of segmentation results directly impacts the accuracy of subsequent operations like part-of-speech tagging and named entity recognition, making the selection and optimization of segmentation algorithms crucial.

Given the diversity of natural language, word segmentation must accommodate a broad range of linguistic phenomena, including compound words, contractions, and multi-word expressions. Particularly in languages with ambiguous word boundaries, segmentation algorithms must be sensitive to contextual cues and syntactic patterns to avoid misinterpretation of meaning. Fine-tuning segmentation approaches to align with the characteristics of the target language and the specific requirements of essay scoring can significantly enhance the precision of subsequent feature extraction processes. Ultimately, effective segmentation forms the bedrock upon which reliable linguistic analysis and automated evaluation are built.

Stemming and stem restoration operations aim to reduce lexical diversity by converting different morphological forms of words back to their base forms, thereby enhancing feature stability and consistency [65]. Stemming extracts the core part of a word by removing prefixes and suffixes. Morphological normalization restores words to their standard dictionary forms. These operations not only reduce lexical dimensionality and mitigate data sparsity issues but also enable the system to focus on core lexical semantics, thereby enhancing the generalization capability and accuracy of scoring models.

A comparative analysis of preprocessing operations is presented in Table 5.

DOI: 10.64972/dea.2025.v3i1.32

Table 5. Comparative Analysis of Preprocessing Operations

Operation	Advantages	Disadvantages
Text cleaning	Remove irrelevant characters symbols numbers retain pure text content	Over-cleaning may lead to loss of useful information
Word segmentation	Split text into lexical units facilitate feature extraction	Segmentation errors affect subsequent analysis poor adaptation to domain-specific terms
Stemming	Reduce vocabulary diversity improve feature stability	Over-stemming may change meaning not applicable to some words
Lemmatization	Restore words to standard form unify word representation	Requires high-quality lemmatization rules or models slower processing speed

Selecting appropriate preprocessing techniques is a foundational step in the construction of high-quality text analytics systems. Effective preprocessing enhances the clarity and uniformity of input data, directly impacting the accuracy of downstream analysis and modeling. Striking the right balance between data simplification and the preservation of meaningful linguistic information remains a central concern, especially as systems are extended to handle increasingly diverse and unstructured essay corpora.

Feature Extraction Module

The feature extraction module plays a crucial role in automatic essay scoring systems, with its core task being to extract features from preprocessed text that reflect essay quality. The comprehensiveness and accuracy of feature extraction directly impact the performance of scoring models and the final scoring results [66]. Feature extraction encompasses lexical, syntactic, and semantic features, among others [56], as detailed in Table 6.

Table 6. Classification and Comparison of Feature Extraction Methods

Method	Applicable Scenario	Advantages	Disadvantages
Lexical feature extraction	All types of essay scoring	Directly reflect vocabulary usage simple calculation	Ignores semantic and grammatical information shallow text understanding
Grammatical feature extraction	Essay scoring with high grammar requirements	Reflect language structure and expression norms dig deeper into grammar info	Sensitive to grammatical errors requires high-quality grammar analysis tools
Semantic feature extraction	Essay scoring requiring deep semantic understanding	Capture deep meaning and logical relationships improve scoring accuracy	High computational complexity heavily dependent on model quality and training data

Feature engineering remains a pivotal aspect of automatic essay scoring, as the choice and quality of extracted features determine the extent to which a system can accurately reflect human judgment. A well-designed feature extraction strategy captures not only surface-level language attributes but also deeper structural and semantic relationships within the text. By combining different types of features, automated systems can better approximate the holistic criteria typically employed by human raters, enhancing both scoring validity and diagnostic feedback.

Understanding the effectiveness of an automatic essay scoring system relies heavily on the quality and diversity of features extracted from student essays. By comparing different approaches to feature extraction—lexical, grammatical, and semantic—researchers and practitioners can better appreciate the strengths and limitations of each method in capturing various dimensions of writing quality. This comparison highlights the importance of selecting appropriate features that align with the assessment objectives and the capabilities of the chosen scoring model. To provide a concise summary of these methods and facilitate direct comparison, Figure 3 presents a visual overview of the primary feature extraction techniques commonly applied in automatic essay evaluation.

DOI: 10.64972/dea.2025.v3i1.32

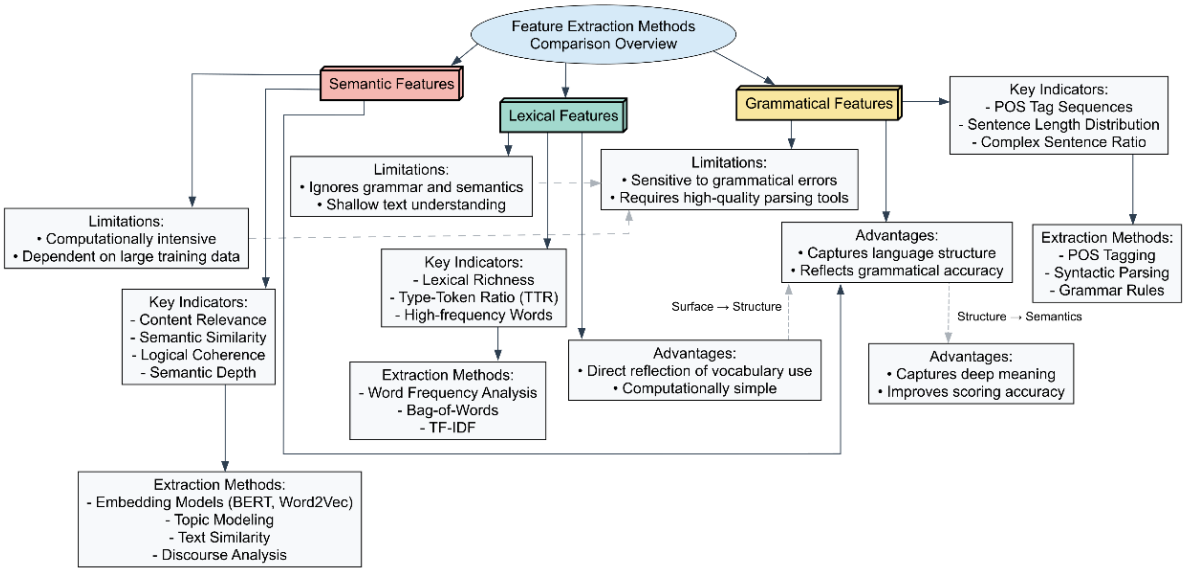

Figure 3. Comparison of feature extraction methods: lexical, grammatical, and semantic features.

The comparative visualization in this figure underscores the importance of selecting appropriate features to capture different dimensions of writing quality. Lexical features provide a straightforward measure of vocabulary usage, while grammatical and semantic features delve deeper into language structure and meaning. By juxtaposing these approaches, the figure clarifies the trade-offs involved in feature selection and the potential impact on scoring accuracy. This comparison not only guides model development but also informs the design of more equitable and nuanced assessment criteria, ultimately supporting fairer and more comprehensive essay evaluation.

Lexical characteristics serve as one of the fundamental dimensions for assessing essay quality, encompassing metrics such as lexical richness, lexical diversity, and the usage of high-frequency vocabulary. Lexical richness can be measured by calculating the ratio of distinct words to the total word count in an essay, reflecting the breadth and accumulation of a student's vocabulary [67]. Lexical diversity, meanwhile, focuses on the evenness of word distribution. Statistically analyzing the frequency and distribution of high-frequency words within essays reveals students' mastery and application of core vocabulary, identifies potential overreliance on fixed expressions, and provides crucial insights for evaluating linguistic expression.

In addition to these quantitative measures, qualitative analysis of lexical usage further enriches the evaluation of essay quality. By examining the appropriateness and contextual relevance of word choices, automated systems can distinguish between mere vocabulary variety and the effective deployment of language. For instance, the use of collocations, idiomatic expressions, and academic vocabulary often signals a more advanced command of language, as compared to repetitive or formulaic phrasing. Moreover, the presence of synonyms and paraphrasing throughout an essay demonstrates a student's flexibility in expression and ability to avoid redundancy. Beyond individual word selection, the integration of transitional phrases and cohesive devices also contributes to lexical sophistication, supporting the logical flow and coherence of the essay. Collectively, these lexical attributes not only reflect a student's linguistic competence but also provide a nuanced perspective on their ability to convey complex ideas with clarity and precision. As a result, comprehensive assessment of lexical features plays a vital role in capturing the multidimensional nature of writing proficiency within automated essay scoring frameworks.

Grammatical feature extraction primarily analyzes syntactic structures in essays, including part-of-speech tagging sequences, sentence length distribution, and the proportion of complex sentences [68]. Part-of-speech tagging sequences reveal whether sentence structures are grammatically correct and logical. Analyzing these annotated patterns in essays can identify common grammatical errors. Sentence length distribution reflects the

DOI: 10.64972/dea.2025.v3i1.32

richness of sentence structures and expressive fluency. Generally, a balanced mix of long and short sentences enhances an essay's expressiveness and readability. The proportion of complex sentences reflects students' mastery and application of advanced grammatical structures, serving as a key indicator for evaluating the linguistic sophistication and logical coherence of compositions.

Semantic feature extraction leverages semantic analysis techniques from natural language processing to delve into content relevance, logical coherence, and semantic depth [69-70]. Content relevance is assessed by calculating semantic similarity between the composition's theme and a predefined topic. Logical coherence analysis focuses on the natural flow of transitions between paragraphs and connections between sentences, often achieved by evaluating the use of conjunctions and semantic relevance. Semantic depth reflects the intellectual richness and knowledge depth expressed in the essay. It can be preliminarily quantified by analyzing lexical semantic complexity and the hierarchical relationships between concepts, providing robust support for comprehensively assessing the semantic quality of essays. To quantify lexical richness in student essays, one of the most commonly used metrics is the Type-Token Ratio (TTR). This measure reflects the diversity of vocabulary utilized in the text and is calculated as:

$$\mathrm{TTR} = \frac{N_{\mathrm{types}}}{N_{\mathrm{tokens}}} \qquad (2)$$

where N_{types} is the number of unique words (types) and N_{tokens} is the total number of words (tokens) in the essay. A higher TTR indicates greater lexical diversity, which is generally associated with higher writing quality.

Scoring Model Module

The scoring model module serves as the core component of the automated essay scoring system, with its primary function being to predict essay scores based on extracted features [71]. This module typically employs machine learning or deep learning algorithms for training. By learning the mapping relationship between features and scores from large datasets of annotated essays, it constructs models capable of accurately predicting essay scores.

In practical implementations, the selection of an appropriate scoring model greatly influences the accuracy and fairness of automated essay evaluation. Traditional machine learning methods, such as linear regression, support vector machines, and decision trees, have been widely adopted due to their interpretability and relatively low computational requirements. These algorithms typically rely on carefully engineered features that capture lexical, syntactic, and semantic aspects of student writing. By analyzing patterns within annotated datasets, these models can generalize scoring criteria across diverse essay topics and proficiency levels. To further enhance model robustness, ensemble approaches that combine multiple algorithms have also been explored, leveraging the complementary strengths of different classifiers. Importantly, the effectiveness of any scoring model depends not only on algorithmic design but also on the quality and representativeness of the training corpus. As such, curating balanced and well-annotated essay datasets remains a critical step in ensuring that the automated system delivers consistent, objective, and pedagogically valuable feedback to both educators and learners.

Traditional machine learning algorithms are widely applied in essay automatic scoring, including linear regression [55], support vector machines [72], and decision trees [73]. Linear regression models assume a linear relationship between features and scores, determining model parameters by minimizing the difference between predicted and actual scores. Support vector machines classify essays into different score categories by finding a hyperplane in the feature space, making them suitable for handling nonlinear relationships. Decision tree algorithms progressively assign essays to different score categories based on feature values through a series of decision rules, offering high interpretability. While these traditional algorithms achieve satisfactory scoring results with limited data and thorough feature engineering, their ability to handle complex semantic features and large-scale data is relatively constrained.

For many early automatic essay scoring systems, linear regression has served as a fundamental approach to predict essay scores based on extracted features. The general form of a linear regression scoring model is given by:

DOI: 10.64972/dea.2025.v3i1.32

$$\hat{y} = w_0 + w_1 x_1 + w_2 x_2 + \cdots + w_n x_n \quad (3)$$

where $\hat{y}$ denotes the predicted score, $x_1, x_2, \ldots, x_n$ are the feature values, and $w_0, w_1, \ldots, w_n$ are the model coefficients learned from the training data. This linear mapping allows for interpretable relationships between input features and output scores.

Evaluation Methods for Automatic Essay Scoring Systems

Evaluation Metric System

To ensure the reliability and validity of automatic essay scoring systems, it is essential to conduct systematic and rigorous performance evaluations. These assessments not only verify the agreement between automated scores and human raters but also help identify potential biases and limitations within the scoring algorithms. A robust evaluation framework typically encompasses both quantitative and qualitative analyses, allowing researchers to examine the accuracy, consistency, and fairness of scoring outcomes across diverse essay prompts, genres, and proficiency levels. Furthermore, comprehensive evaluation enables the identification of cases where the system may underperform, such as essays with unconventional structures or creative language use, thereby informing iterative improvements in model design. By adopting a multifaceted approach to evaluation, stakeholders can better understand the strengths and weaknesses of automated grading systems and guide their responsible deployment in educational settings.

Assessing the performance of automatic essay scoring systems requires establishing a scientific and reasonable metric system. Common evaluation metrics include accuracy, recall, and F1 score [74]. Accuracy measures the proportion of correctly predicted essays out of the total number predicted; recall indicates the proportion of correctly predicted essays out of the actual number of essays that should have been correctly predicted; the F1 score is the harmonic mean of accuracy and recall, comprehensively reflecting system performance. In addition to these classification evaluation metrics, indicators such as mean squared error and root mean squared error can also be used to assess the degree of discrepancy between scoring results and human evaluations [75].

To objectively evaluate the performance of automatic essay scoring systems, several standard metrics are widely adopted, including accuracy, recall, and the F1 score. These can be mathematically defined as follows:

$$\begin{aligned} \text{Accuracy} &= \frac{TP+TN}{TP+TN+FP+FN} \\ \text{Recall} &= \frac{TP}{TP+FN} \\ \text{F1} &= 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \end{aligned} \quad (4)$$

where TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives, respectively. These metrics provide a comprehensive view of the model's classification effectiveness.

Evaluating the performance of automatic essay scoring systems involves the application of various quantitative metrics to assess the accuracy and reliability of scoring outputs. By employing standard measures such as accuracy, recall, and F1 score, researchers can systematically compare different models and approaches, thus ensuring the objectivity and scientific rigor of the evaluation process. A clear understanding of these evaluation metrics, as well as their mathematical definitions and interpretative value, is essential for both model development and result interpretation. To aid comprehension and support effective comparison, Figure 4 offers a visual summary of commonly used evaluation metrics in the context of automatic essay scoring.

DOI: 10.64972/dea.2025.v3i1.32

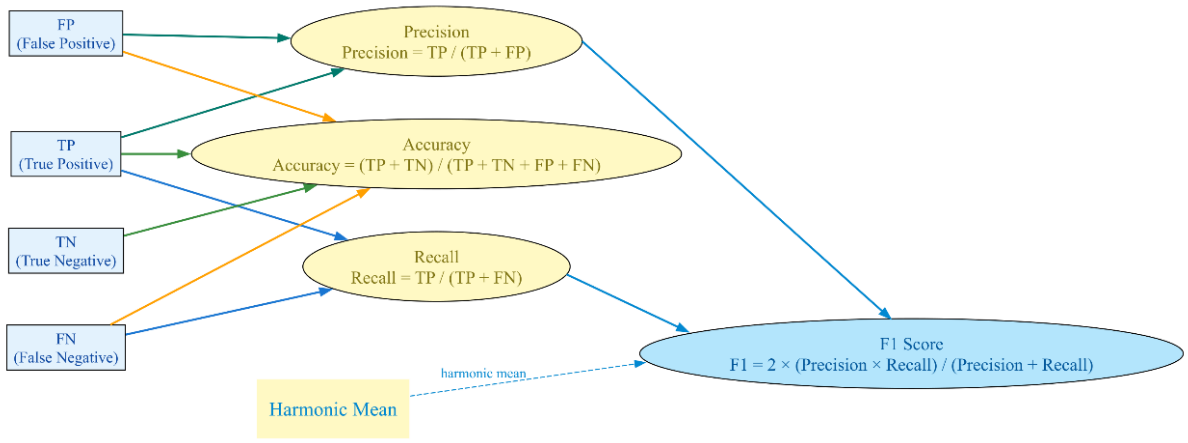

Figure 4. Visualization of evaluation metrics (accuracy, recall, F1 score) used in automatic essay scoring.

The presented metrics offer a standardized framework for assessing the effectiveness of essay scoring models, enabling objective comparison across different systems and methodologies. By incorporating measures such as accuracy, recall, and F1 score, the evaluation process becomes more transparent and scientifically rigorous. This figure emphasizes the necessity of multidimensional evaluation, as each metric captures a distinct aspect of system performance. Clear understanding and application of these metrics facilitate iterative improvements, guiding the development of more robust and reliable automated assessment tools.

Experimental Design for Evaluation

Before embarking on the evaluation of automatic essay scoring systems, it is crucial to establish a robust experimental framework that ensures both methodological rigor and replicability. The experimental setup should be aligned with the research objectives, accounting for the diversity of essay topics, writing styles, and language proficiency levels encountered in real-world educational settings. Standardizing the evaluation process not only facilitates fair comparison between different scoring models but also helps to mitigate potential biases arising from dataset imbalances or annotation inconsistencies. Moreover, a well-structured experimental protocol provides transparency in reporting, enabling other researchers to validate and extend prior findings. Careful consideration of these foundational elements lays the groundwork for meaningful performance assessment and supports the ongoing development of reliable automated essay scoring solutions.

When conducting evaluation experiments, it is necessary to select an appropriate experimental dataset and partition it into training, validation, and test sets. The training set is used to train the scoring model, the validation set for adjusting model hyperparameters, and the test set for the final performance evaluation. Experimental design typically follows these steps:

1) Experimental dataset selection and partitioning. Select a representative essay dataset covering diverse topics, styles, and proficiency levels, with corresponding human ratings for each essay;

2) Data Preprocessing. Perform text cleaning, word segmentation, stemming, and other preprocessing operations on all essays to standardize formats and remove noise. Normalize human ratings to meet model training requirements;

3) Feature Extraction. Extract lexical, syntactic, and semantic features from the preprocessed essays. Literature [60] has explored features such as lexical richness, lexical diversity, part-of-speech tagging sequences, sentence length distribution, and text similarity;

4) Model Training and Validation. Train the scoring model using the training set, optimizing model parameters based on the selected algorithm. Evaluate model performance on the validation set, adjust hyperparameters according to evaluation metrics, and select the optimal model;

DOI: 10.64972/dea.2025.v3i1.32

5) Model Testing and Evaluation. Assess the performance of the final selected model using the test set, calculate various evaluation metrics, and analyze the model's generalization capability and practical application effectiveness;

6) Result Analysis and Optimization. Based on the evaluation results, analyze the model's strengths and weaknesses to identify areas for improvement. Address identified issues by further optimizing the model, such as adjusting feature extraction methods, refining the model architecture, or augmenting the training data.

The empirical validation of automatic essay scoring systems necessitates a well-structured experimental design that encompasses data preparation, feature engineering, model training, validation, and performance analysis. By delineating each stage of the evaluation process, researchers can ensure methodological transparency and reproducibility, which are critical for advancing the field and building trust in automated assessment technologies. A detailed depiction of the experimental workflow helps clarify how data flows through the system and how each component contributes to the final evaluation outcomes. To provide a clear and comprehensive overview of these procedures, Figure 5 illustrates the standard experimental workflow employed in the evaluation of automatic essay scoring systems.

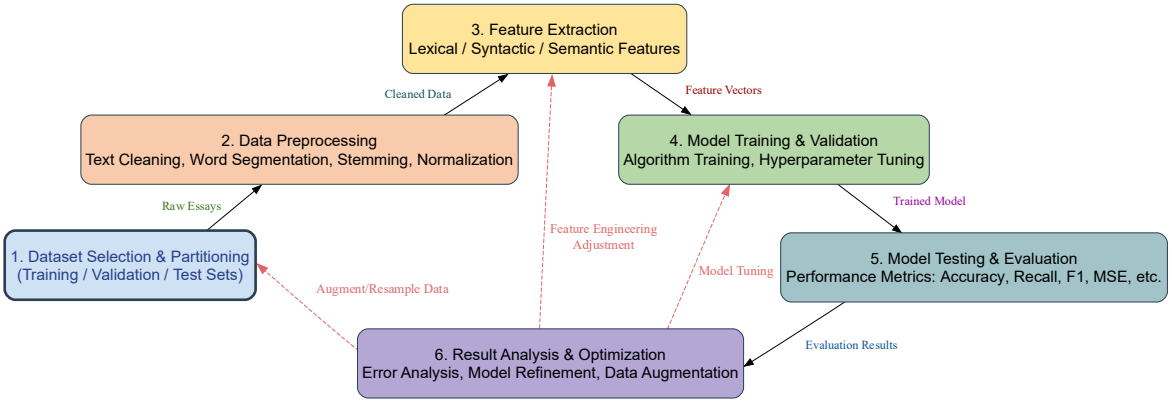

Figure 5. Experimental workflow for evaluating an automatic essay scoring system.

The depicted experimental workflow provides a comprehensive overview of the sequential stages involved in system validation, from data preparation to performance analysis. By systematically outlining each step, the figure ensures methodological clarity and reproducibility, which are fundamental for advancing research in automated essay scoring. This structured approach allows for the identification of variables affecting model performance and supports continuous optimization. Moreover, the workflow fosters a more transparent evaluation environment, promoting confidence in research findings and the real-world applicability of essay scoring systems.

Conclusion

This paper provides a comprehensive review of natural language processing-driven automatic essay scoring systems. It first introduces the fundamentals of natural language processing and its applications in text analysis, then elaborates on the architecture of automatic essay scoring systems, including preprocessing modules, feature extraction modules, and scoring model modules, while exploring evaluation methods for such systems. Research findings indicate that NLP technologies provide robust support for essay grading, enabling rapid and objective evaluation that enhances the efficiency and objectivity of essay assessment. However, challenges persist in this field, such as adaptability to diverse languages and writing styles, interpretability of scoring models, and optimization of system efficiency and performance. Future research should focus on enhancing model interpretability, optimizing system performance, and exploring cross-language and cross-domain applications to advance the broader adoption and development of automatic essay scoring systems in educational settings.

Furthermore, the integration of NLP-driven essay scoring systems into educational environments offers substantial benefits beyond efficiency and objectivity. These systems facilitate more consistent feedback for students, helping to identify strengths and areas for improvement in their writing. By leveraging a variety of linguistic features—ranging from lexical richness and grammatical accuracy to semantic coherence—automated scoring platforms can provide multi-dimensional assessments that mirror human evaluators' holistic perspectives. Importantly, such systems support teachers by reducing grading workloads, allowing greater focus on individualized instruction and pedagogical innovation. Despite existing limitations, the increasing sophistication of NLP methodologies holds promise for fostering fairer and more transparent evaluation standards, while also paving the way for adaptive learning technologies that cater to diverse learner needs. As the field evolves, fostering collaboration between computational researchers, educators, and linguists will be vital to ensure that automatic essay scoring solutions remain pedagogically sound, ethically responsible, and attuned to the dynamic landscape of educational assessment.

References

[1] Yamamoto, M., Umemura, N., & Kawano, H. (2020). Proposal of Japanese Vocabulary Difficulty Level Dictionaries for Automated Essay Scoring Support System Using Rubric. Journal of the Operations Research Society of China, 8(4), 601–617. DOI:10.1007/s40305-019-00270-z

[2] Ahmad, R. (2019). E-learning Automated Essay Scoring System Menggunakan Metode Searching Text Similarity Matching Text. Jurnal Penelitian Enjiniring, 22(1), 38–43. DOI:10.25042/jpe.052018.07

[3] Darwish, S. M., Ali, R. A., & Elzoghabi, A. A. (2023). An Automated English Essay Scoring Engine Based on Neutrosophic Ontology for Electronic Education Systems. Applied Sciences (Switzerland), 13(15). DOI:10.3390/app13158601

[4] Conijn, R., Kahr, P., & Snijders, C. (2023). The Effects of Explanations in Automated Essay Scoring Systems on Student Trust and Motivation. Journal of Learning Analytics, 10(1), 37–53. DOI:10.18608/jla.2023.7801

[5] Machicao, J. C. (2019). Higher Education Challenge Characterization to Implement Automated Essay Scoring Model for Universities with a Current Traditional Learning Evaluation System. Advances in Intelligent Systems and Computing, 918, 835–844. DOI:10.1007/978-3-030-11890-7_78

[6] Uto, M., Xie, Y., & Ueno, M. (2020). Neural Automated Essay Scoring Incorporating Handcrafted Features. COLING 2020 - 28th International Conference on Computational Linguistics, Proceedings of the Conference, 6077–6088. DOI:10.18653/v1/2020.coling-main.535

[7] Uto, M. (2021). A review of deep-neural automated essay scoring models. Behaviormetrika, 48(2), 459–484. DOI:10.1007/s41237-021-00142-y

[8] Susilawati, E., Lubis, H., Kesuma, S., Pratama, K., & Khaira, I. (2022). The Mediating Role of Moral Self-Regulations between Automated Essay Scoring Adoption, Students' Character and Academic Integrity among Indonesian Higher Education Sector. Eurasian Journal of Educational Research, 2022(102), 54–71. DOI:10.14689/ejer.2022.102.004

[9] Wilson, J., & Rodrigues, J. (2020). Classification accuracy and efficiency of writing screening using automated essay scoring. Journal of School Psychology, 82, 123–140. DOI:10.1016/j.jsp.2020.08.008

[10] Bai, J. Y. H., Zawacki-Richter, O., Bozkurt, A., Lee, K., Fanguy, M., Cefa Sari, B., & Marin, V. I. (2022). Automated Essay Scoring (AES) Systems: Opportunities and Challenges for Open and Distance Education. Tenth Pan-Commonwealth Forum on Open Learning. DOI:10.56059/pcf10.8339

[11] Yun, J. (2023). Meta-Analysis of Inter-Rater Agreement and Discrepancy Between Human and Automated English Essay Scoring. English Teaching(South Korea), 78(3), 105–124. DOI:10.15858/engtea.78.3.202309.105

[12] Qian, L., Zhao, Y., & Cheng, Y. (2020). Evaluating China's Automated Essay Scoring System iWrite. Journal of Educational Computing Research, 58(4), 771–790. DOI:10.1177/0735633119881472

[13] Hussein, M. A., Hassan, H. A., & Nassef, M. (2020). A trait-based deep learning automated essay scoring system with adaptive feedback. International Journal of Advanced Computer Science and Applications, 11(5), 287–293. DOI:10.14569/IJACSA.2020.0110538

[14] Shermis, M. D. (2022). Anchoring Validity Evidence for Automated Essay Scoring. Journal of Educational Measurement, 59(3), 314–337. DOI:10.1111/jedm.12336

[15] Almusharraf, N., & Alotaibi, H. (2023). An error-analysis study from an EFL writing context: Human and Automated Essay Scoring Approaches. Technology, Knowledge and Learning, 28(3), 1015–1031. DOI:10.1007/s10758-022-09592-z

[16] Ferrara, S., & Qunbar, S. (2022). Validity Arguments for AI-Based Automated Scores: Essay Scoring as an Illustration. Journal of Educational Measurement, 59(3), 288–313. DOI:10.1111/jedm.12333

[17] Rysová, K., Rysová, M., Novák, M., Mírovský, J., & Hajičová, E. (2019). EVALD – a Pioneer Application for Automated Essay Scoring in Czech. The Prague Bulletin of Mathematical Linguistics, 113(1), 9–30. DOI:10.2478/pralin-2019-0004

[18] Gaheen, M. M., ElEraky, R. M., & Ewees, A. A. (2021). Automated students arabic essay scoring using trained neural network by e-jaya optimization to support personalized system of instruction. Education and Information Technologies, 26(1), 1165–1181. DOI:10.1007/s10639-020-10300-6

[19] Susilawati, E., Lubis, H., Kesuma, S., Pratama, I., & Khaira, I. (2023). Exploring the antecedents of Student Academic Integrity: The Impact of Using Digital Technology Automated Short Essay Scoring (ASES) Assessment Models in Learning. Eurasian Journal of Educational Research, 2023(103), 145–164. DOI:10.14689/ejer.2023.103.009

[20] Kusuma, J. S., Halim, K., Pranoto, E. J. P., Kanigoro, B., & Irwansyah, E. (2022). Automated Essay Scoring Using Machine Learning. 2022 4th International Conference on Cybernetics and Intelligent System, ICORIS 2022, 1–5. DOI:10.1109/ICORIS56080.2022.10031338

[21] Li, H., & Dai, T. (2020). Explore Deep Learning for Chinese Essay Automated Scoring. Journal of Physics: Conference Series, 1631(1), 12036. DOI:10.1088/1742-6596/1631/1/012036

[22] Osakwe, K. A., Ola, K., & Omotosho, P. (2021). Contactless Academia – The Case for Automated Essay Scoring (AES) System in COVID 19 Pandemic. Current Journal of Applied Science and Technology, 4, 17–29. DOI:10.9734/cjast/2021/v40i431292

[23] Yuan, S., He, T., Huang, H., Hou, R., & Wang, M. (2020). Automated chinese essay scoring based on deep learning. Computers, Materials and Continua, 65(1), 817–833. DOI:10.32604/cmc.2020.010471

[24] UYSAL, İ., & DOĞAN, N. (2021). Automated Essay Scoring Effect on Test Equating Errors in Mixed-format Test. International Journal of Assessment Tools in Education, 8(2), 222–238. DOI:10.21449/ijate.815961

[25] Ikram, A., & Castle, B. (2020). Automated Essay Scoring (AES); A Semantic Analysis Inspired Machine Learning Approach: An automated essay scoring system using semantic analysis and machine learning is presented in this research. ACM International Conference Proceeding Series, 147–151. DOI:10.1145/3436756.3437036

[26] Chamidah, N., Yulianti, E., & Budi, I. (2023). Evaluating the Impact of Sentence Tokenization on Indonesian Automated Essay Scoring Using Pretrained Sentence Embeddings. Revue d'Intelligence Artificielle, 37(5), 1101–1108. DOI:10.18280/ria.370502

[27] Alobed, M., Altrad, A. M. M., Bakar, Z. B. A., & Zamin, N. (2021). Automated Arabic Essay Scoring Based on Hybrid Stemming With Wordnet. Malaysian Journal of Computer Science, 2021(Special Issue 2), 55–67. DOI:10.22452/mjcs.sp2021no2.4

[28] Mizumoto, A., & Eguchi, M. (2023). Exploring the potential of using an AI language model for automated essay scoring. Research Methods in Applied Linguistics, 2(2). DOI:10.1016/j.rmal.2023.100050

[29] Lee, A. V. Y., Luco, A. C., & Tan, S. C. (2023). A Human-Centric Automated Essay Scoring and Feedback System for the Development of Ethical Reasoning. Educational Technology and Society, 26(1), 147–159. DOI:10.30191/ETS.202301_26(1).0011

[30] Park, K. Y., & Lee, Y. S. (2022). Deep Learning Algorithm Exploration for Automated Korean essay Scoring. Korean Society for Educational Evaluation, 35(3), 465–488. DOI:10.31158/jeev.2022.35.3.465

[31] Rupp, A. A., Casabianca, J. M., Krüger, M., Keller, S., & Köller, O. (2019). Automated Essay Scoring at Scale: A Case Study in Switzerland and Germany. ETS Research Report Series, 2019(1), 1–23. DOI:10.1002/ets2.12249

[32] Li, X., Chen, M., & Nie, J. Y. (2020). SEDNN: Shared and enhanced deep neural network model for cross-prompt automated essay scoring. Knowledge-Based Systems, 210. DOI:10.1016/j.knosys.2020.106491

[33] Yang, R., Cao, J., Wen, Z., Wu, Y., & He, X. (2020). Enhancing automated essay scoring performance via fine-tuning pre-trained language models with combination of regression and ranking. Findings of the Association for Computational Linguistics Findings of ACL: EMNLP 2020, 1560–1569. DOI:10.18653/v1/2020.findings-emnlp.141

[34] Beseiso, M., & Alzahrani, S. (2020). An empirical analysis of BERT embedding for automated essay scoring. International Journal of Advanced Computer Science and Applications, 11(10), 204–210. DOI:10.14569/IJACSA.2020.0111027

[35] Sharma, A., Katlaa, R., Kaur, G., & Jayagopi, D. B. (2023). Full-page handwriting recognition and automated essay scoring for in-the-wild essays. Multimedia Tools and Applications, 82(23), 35253–35276. DOI:10.1007/s11042-023-14558-z

[36] Hussein, M. A., Hassan, H., & Nassef, M. (2019). Automated language essay scoring systems: A literature review. PeerJ Computer Science, 2019(8), e208. DOI:10.7717/peerj-cs.208

[37] Jong, Y. J., Kim, Y. J., & Ri, O. C. (2022). Improving Performance of Automated Essay Scoring by Using Back-Translation Essays and Adjusted Scores. Mathematical Problems in Engineering, 2022. DOI:10.1155/2022/6906587

[38] Shin, J., & Gierl, M. J. (2021). More efficient processes for creating automated essay scoring frameworks: A demonstration of two algorithms. Language Testing, 38(2), 247–272. DOI:10.1177/0265532220937830

[39] Stephen, T. C., Gierl, M. C., & King, S. (2021). Automated essay scoring (AES) of constructed responses in nursing examinations: An evaluation. Nurse Education in Practice, 54(1), 103085. DOI:10.1016/j.nepr.2021.103085

[40] Lim, C. T., Bong, C. H., Wong, W. S., & Lee, N. K. (2021). A comprehensive review of automated essay scoring (Aes) research and development. Pertanika Journal of Science and Technology, 29(3), 1875–1899. DOI:10.47836/pjst.29.3.27

[41] Kumar, V., & Boulanger, D. (2020). Explainable Automated Essay Scoring: Deep Learning Really Has Pedagogical Value. Frontiers in Education, 5. DOI:10.3389/feduc.2020.572367

[42] Ramesh, D., & Sanampudi, S. K. (2022). An automated essay scoring systems: a systematic literature review. Artificial Intelligence Review, 55(3), 2495–2527. DOI:10.1007/s10462-021-10068-2

[43] Beseiso, M., Alzubi, O. A., & Rashaideh, H. (2021). A novel automated essay scoring approach for reliable higher educational assessments. Journal of Computing in Higher Education, 33(3), 727–746. DOI:10.1007/s12528-021-09283-1

[44] Xue, S., Zhang, J., Zhou, J., & Ren, F. (2022). Robust Automated Essay Scoring by Using Attentive Capsule. Proceedings of 2022 8th IEEE International Conference on Cloud Computing and Intelligence Systems, CCIS 2022, 595–599. DOI:10.1109/CCIS57298.2022.10016365

[45] Contreras, J. O., Hilles, S. M. S., & Abubaker, Z. B. (2019). Automated essay scoring using ontology generator and natural language processing with question generator based on blooms taxonomy's cognitive level. International Journal of Engineering and Advanced Technology, 9(1), 2448–2457. DOI:10.35940/ijeat.A9974.109119

[46] Li, F., Xi, X., Cui, Z., Li, D., & Zeng, W. (2023). Automatic Essay Scoring Method Based on Multi-Scale Features. Applied Sciences (Switzerland), 13(11). DOI:10.3390/app13116775

[47] Tan, J. S., & Tan, I. K. T. (2022). Ablation Study on Feature Group Importance for Automated Essay Scoring. Asia-Pacific Journal of Information Technology and Multimedia, 11(1), 90–101. DOI:10.17576/apjitm-2022-1101-08

[48] Jeon, Y. (2023). Exploring the Efficacy of Automated Essay Scoring Systems in the Context of Korean EFL High School. The Korea English Language Testing Association, 18(1), 63–95. DOI:10.37244/ela.2023.18.1.63

[49] Huang, Z., Liu, H., Wu, J., & Lv, C. (2023). Conditional Predictive Behavior Planning With Inverse Reinforcement Learning for Human-Like Autonomous Driving. IEEE Transactions on Intelligent Transportation Systems, 24(7), 7244–7258. DOI:10.1109/TITS.2023.3254579

[50] Carmon, C. M., Morgan, B., Hu, X., & Graesser, A. C. (2023). Automated Assessment of Initial Answers to Questions in Conversational Intelligent Tutoring Systems: Are Contextual Embedding Models Really Better? Electronics (Switzerland), 12(17), 18. DOI:10.3390/electronics12173654

[51] Pei, F., Zhang, J., Yuan, M., He, F., & Yan, B. (2023). The Evaluation Technology of Manufacturer Intelligence Regarding the Selection of the Decision Support System of Smart Manufacturing Technologies: Analysis of China–South Africa Relations. Processes, 11(7). DOI:10.3390/pr11072185

[52] Chen, Y. (2023). Analyzing the design of intelligent English translation and teaching model in colleges using data mining. Soft Computing, 27(19), 14497–14513. DOI:10.1007/s00500-023-09096-7

[53] Wang, Q., Wan, Y., & Feng, F. (2023). Human–machine collaborative scoring of subjective assignments based on sequential three-way decisions. Expert Systems with Applications, 216, 119466-. DOI:10.1016/j.eswa.2022.119466

[54] Narayanan, S., Ramakrishnan, R., Durairaj, E., & Das, A. (2023). Artificial Intelligence Revolutionizing the Field of Medical Education. Cureus. DOI:10.7759/cureus.49604

[55] Subashini, K., & Narmatha, V. (2023). OptiPhishDetect: Optimized Phishing Detection through Learning based GCN with Scoring Model. International Journal of Intelligent Engineering and Systems, 16(6), 863–873. DOI:10.22266/ijies2023.1231.71

[56] Zhang, J. (2022). Data-Driven Teaching Model Design of College English Translation Using Intelligent Processing Technology. Wireless Communications and Mobile Computing, 2022. DOI:10.1155/2022/6559772

[57] Pang, Y. (2022). Evaluation and Promotion of a Multidimensional Information Intelligent Speech System in Dialect Teaching. Journal of Sensors, 2022. DOI:10.1155/2022/1692080

[58] Qin, F. (2022). College English Intelligent Writing Score System Based on Big Data Analysis and Deep Learning Algorithm. Journal of Database Management, 33(5). DOI:10.4018/jdm.314561

[59] Zhu, Q. (2023). Enhancing vulnerability scoring for information security in intelligent computers. International Journal of Intelligent Networks, 4, 253–260. DOI:10.1016/j.ijin.2023.09.002

[60] Gao, M., Ahipasaoglu, S., & Schuster, K. (2023). Automated poetry scoring using BERT with multi-scale poetry representation. International Journal of Intelligent Systems Technologies and Applications, 21(3), 250–261. DOI:10.1504/IJISTA.2023.133694

[61] Sujatha, T., Blessing, W. N. R., & Palarimath, S. (2023). Mining Competitors and Finding Winning Plans Using Feature Scoring and Ranking-Based CMiner++ Algorithm: Finding Top-K Competitors. International Journal of Intelligent Information Technologies, 19(1). DOI:10.4018/IJIIT.318670

[62] Ahmed, S., Arif, M., Kabir, M., Khan, K., & Khan, Y. D. (2022). PredAoDP: Accurate identification of antioxidant proteins by fusing different descriptors based on evolutionary information with support vector machine: Identification of Antioxidant proteins. Chemometrics and Intelligent Laboratory Systems, 228. DOI:10.1016/j.chemolab.2022.104623

[63] Jamei, M., Maroufpoor, S., Aminpour, Y., Karbasi, M., Malik, A., & Karimi, B. (2022). Developing hybrid data-intelligent method using Boruta-random forest optimizer for simulation of nitrate distribution pattern. Agricultural Water Management, 270. DOI:10.1016/j.agwat.2022.107715

[64] Wang, D., Liu, H., & Li, Y. (2022). Intelligent weight generation algorithm based on binary isolation tree. Engineering Applications of Artificial Intelligence, 109, 104604-. DOI:10.1016/j.engappai.2021.104604

[65] Wang, Y., & Na, K. S. (2022). Innovative Research on English Teaching Model Based on Artificial Intelligence and Wireless Communication. International Journal of Reliability, Quality and Safety Engineering, 29(5). DOI:10.1142/S0218539322400071

[66] Chen, R., Ju, C. H., & Shen.tu, F. (2022). A Credit Scoring Ensemble Framework using Adaboost and Multi-layer Ensemble Classification. ACM International Conference Proceeding Series, 72–79. DOI:10.1145/3549179.3549199

[67] Zhang, J., Du, X., & Bi, R. (2022). Intelligent Recognition System of Sports Athletes' Wrong Actions Based on AI+IoT. Wireless Communications and Mobile Computing, 2022. DOI:10.1155/2022/3455224

[68] Pang, S., Yang, G., Li, Y., He, Y., & Shen, Z. (2022). An Intelligent Evaluation System of Air Traffic Control Training Simulator under Special Situation. ACM International Conference Proceeding Series, 34–38. DOI:10.1145/3568364.3568370

[69] My, B. T. T., & Ta, B. Q. (2023). An interpretable decision tree ensemble model for imbalanced credit scoring datasets. Journal of Intelligent and Fuzzy Systems, 45(6), 10853–10864. DOI:10.3233/JIFS-230825

[70] Zhang, X. (2022). Intelligent Recommendation Algorithm of Multimedia English Distance Education Resources Based on User Model. Journal of Mathematics, 2022. DOI:10.1155/2022/2012700

[71] Hu, N., & Bi, Y. (2022). Multimodal Intelligent Acoustic Sensor-Assisted English Pronunciation Signal Acquisition and Phonetic Calibration. Journal of Sensors, 2022. DOI:10.1155/2022/3383685

[72] Sethi, A. (2024). Artificial Intelligence in Health Professions Education. Journal of Shalamar Medical & Dental College - JSHMDC, 5(1), 1–3. DOI:10.53685/jshmdc.v5i1.227

[73] Wang, Z., Chen, X., & Zhao, Y. (2022). Design and Implementation of Intelligent Decision Support System for THS in Large-Scale Events. Lecture Notes in Operations Research, Part F3781, 142–150. DOI:10.1007/978-981-16-8656-6_13

[74] Zhu, Q. (2022). Reform and Practice of Public English Examination Mode in Colleges and Universities Using Big Data Analysis and Speech Recognition. Mobile Information Systems, 2022. DOI:10.1155/2022/7225495

[75] Duan, W. (2022). Research on Scoring of Business English Oral Training Based on Deep Neural Network. Scientific Programming, 2022. DOI:10.1155/2022/9193454