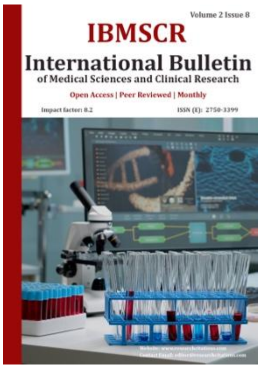

APPLICATION OF NLP ALGORITHMS IN AUTOMATED ASSESSMENT OF ENGLISH WRITING

Ramazonova Sitora
Uzbekistan State World Languages University
ramazonovasitora51@gmail.com

https://doi.org/10.5281/zenodo.17800977

Abstract:

Automated assessment of English writing has gained increasing significance in educational settings due to the growing need for scalable, consistent, and objective evaluation mechanisms. This study investigates the application of Natural Language Processing (NLP) algorithms in automated writing evaluation (AWE), tracing the historical evolution from early surface-feature approaches to modern transformer-based systems that leverage deep contextual embeddings, syntactic parsing, semantic analysis, and discourse modeling. Drawing on a comprehensive review of both foundational systems (e.g., PEG, e-rater) and recent neural architectures, we analyze how these tools assess grammar, vocabulary richness, coherence, cohesion, semantic relevance, and organizational quality. The discussion also highlights key methodological considerations, such as tokenization, parsing, semantic role labeling, coherence modeling, and feedback generation, and reflects on the ethical and pedagogical challenges — including bias, overemphasis on formulaic writing, and fairness in EFL contexts. Finally, the paper explores practical implications for English writing pedagogy in Uzbekistan, arguing that NLP-driven AWE — when combined with human oversight — can offer effective, fair, and pedagogically valuable support for large-scale and formative writing assessment.

Keywords: automated writing evaluation, natural language processing, essay scoring, grammar and cohesion, semantic analysis, educational technology.

Automated assessment of English writing has evolved into one of the most significant innovations in modern educational technology, particularly as institutions increasingly require large-scale, reliable, and cost-effective evaluation of learner-generated texts. The rapid advancement of Natural Language Processing (NLP) has transformed writing assessment from early surface-level feature scoring to today's sophisticated, context-aware systems capable of analyzing grammar, semantics, discourse, and argumentation. This article provides a comprehensive analysis of how NLP algorithms operate within automated writing evaluation (AWE) systems, discusses historical development, outlines current computational techniques, evaluates limitations and ethical challenges, and considers applicability to EFL contexts such as Uzbekistan, where writing proficiency plays a key role in academic and professional advancement. The discussion aims to present a cohesive, research-based narrative of the field while integrating methodological insights, a comparative table of algorithmic approaches, and a forward-looking perspective.

The history of automated writing evaluation can be traced to Page's Project Essay Grade (PEG) introduced in the 1960s. PEG relied primarily on surface-level features such as word count, average sentence length, and variation in sentence structure, using these quantitative variables to predict human scores. Although primitive by modern standards, PEG demonstrated that computational systems could approximate human judgments. In the 1990s and early

IBMSCR | Volume 5, Issue 11, November

2000s, more advanced systems such as ETS's e-rater and Vantage Learning's IntelliMetric emerged. These systems incorporated handcrafted linguistic features, particularly grammar error detection, lexical richness measures, discourse signals, and rudimentary coherence indicators. Their reliance on expert-created rules made them reliable for certain assessment contexts but limited in generalizability. The subsequent introduction of machine learning allowed these systems to move beyond rigid rule sets and learn patterns directly from data. However, the true revolution came with deep learning and transformer-based architectures, which fundamentally changed how machines interpret language.

Modern NLP-driven writing assessment typically begins with text preprocessing. Tokenization, the division of text into words or subword units, is a crucial step because it defines how the system perceives input. Approaches such as Byte Pair Encoding (BPE) and WordPiece enable robust handling of unknown words, misspellings, and morphological variations, all of which are common in student writing. After tokenization, normalization processes ensure consistent capitalization, punctuation, and spacing, reducing noise for downstream algorithms. Part-of-speech tagging then assigns grammatical categories to tokens, allowing the system to analyze morphological agreement, verb tense use, noun phrase complexity, and other syntactic features associated with writing proficiency. Parsing, whether constituency-based or dependency-based, further reveals structural relationships between words and clauses, enabling fine-grained assessments of syntactic sophistication. Writers who use varied clause structures, balanced sentence length, and effective subordination typically score higher, and parsing allows computational systems to capture these nuances.

Beyond syntactic features, semantic analysis plays an essential role in determining content relevance, lexical precision, and conceptual depth. Distributional semantic models, initially based on embeddings such as Word2Vec or GloVe, provided the first wave of semantic assessment by mapping words into vector spaces where distance reflects similarity. Modern contextualized models such as BERT, RoBERTa, and GPT-related architectures incorporate multiple layers of context, enabling the system to understand meaning not only at the word level but across sentence and paragraph boundaries. These models allow automated systems to evaluate task fulfillment by comparing the semantic similarity between the prompt and the student's response. They also make it possible to detect irrelevant digressions, topical drift, or superficial content padding, problems that earlier systems often misinterpreted as rich vocabulary or syntactic complexity.

Coherence and discourse organization constitute another essential dimension of writing assessment. Cohesion refers to how sentences link through lexical repetition, pronoun reference, conjunctions, and discourse markers, while coherence refers to the logical progression of ideas. Neural coherence models use sentence embeddings or discourse-aware architectures to determine whether the text flows naturally, whether each paragraph contributes meaningfully to the central argument, and whether transitions guide the reader effectively. In the past decade, research has shown that coherence strongly predicts scoring outcomes, particularly in argumentative and academic writing. Automated systems now evaluate discourse structure by examining introductory clarity, thesis strength, paragraph unity, and conclusion effectiveness, although perfect modeling of rhetorical quality remains an open challenge.

To clarify how different NLP approaches compare in writing assessment, the following table summarizes key algorithmic families, highlighting their strengths and limitations.

Table 1. Comparison of NLP Algorithmic Approaches in Automated Writing Assessment

Approach	Examples	Strengths	Limitations
Rule-based systems	PEG, early e-rater	Transparent, predictable, easy to interpret	Poor generalization, limited linguistic depth
Statistical ML	SVM, regression, random forests	Learns from data, better accuracy	Dependent on handcrafted features
Neural networks	LSTM, CNN	Captures sequential and local patterns, improved fluency modeling	Limited long-range contextual understanding
Transformers	BERT, RoBERTa, GPT	Deep contextual understanding, high accuracy, handles semantics & discourse	Risk of bias, requires large datasets, costly to train

As Table 1 illustrates, transformer-based models currently dominate AES research due to their ability to capture complex linguistic and rhetorical characteristics without explicit feature engineering. Their attention mechanisms enable them to evaluate not only what is written but also how ideas relate across the entire essay.

A critical component of automated assessment is error detection. Systems now identify grammar, usage, vocabulary, mechanics, and stylistic errors with increasing precision. Grammar checkers powered by neural sequence labeling can detect tense inconsistency, subject–verb disagreement, article misuse, incorrect prepositions, and faulty sentence boundaries. Vocabulary-related errors include mis-collocations, inappropriate word choice, and overreliance on simplistic lexicon, all of which signal lower proficiency levels. Mechanic errors such as punctuation and capitalization are also considered, though they are weighted less heavily in academic scoring. For second-language learners, especially in Uzbekistan or broader Central Asian contexts, error patterns may be influenced by native-language transfer, and advanced NLP systems increasingly incorporate multilingual or learner-corpus training to better model L2 patterns.

Fluency assessment further strengthens automated scoring systems by examining sentence rhythm, syntactic balance, length variation, readability indices, and smoothness of transitions. While traditional readability formulas such as Flesch Reading Ease have limitations, they still contribute valuable indicators when combined with neural representations. Writers who maintain a natural flow, avoid monotony, and integrate syntactic variety generally demonstrate greater fluency and thus receive higher scores.

Another transformative advancement is automated feedback generation. Early systems relied on static templates, offering generic advice such as "add more details." In contrast, modern generative models produce personalized, context-sensitive feedback, identifying unclear argumentation, suggesting lexical improvements, highlighting missing evidence, or recommending more coherent paragraph organization. This type of feedback supports iterative revision cycles, enabling learners to refine their writing skills through practice, which is especially important in educational systems where teacher workloads limit the amount of individualized commentary they can provide.

Despite these advances, automated writing assessment faces several significant challenges. One persistent issue is distinguishing between genuine errors and creative experimentation. Writers sometimes intentionally break conventions for rhetorical effect; however, algorithms optimized for normative patterns may misclassify such deviations as mistakes. Another issue is prompt padding: some students artificially inflate essay length or use overly sophisticated vocabulary to manipulate the scoring system. Although modern semantic models mitigate this problem, no system is immune. Bias also remains a concern because models trained on limited datasets may favor specific discourse styles, cultural norms, or dialects. For example, essays written by Central Asian EFL learners may differ stylistically from essays written by native English speakers, yet scoring models rarely account for such variation unless explicitly trained to do so. Transparency is equally important: educators and test developers need explanations for how scores are generated, especially in high-stakes contexts such as university entrance exams or professional certification.

Looking to the future, the integration of automated assessment into EFL contexts like Uzbekistan presents both opportunities and responsibilities. Educational institutions increasingly adopt digital platforms, and AWE tools can support placement testing, classroom assessments, and proficiency exams. They enable frequent low-stakes writing practice, generate immediate feedback, and lighten instructor workload. However, hybrid scoring approaches combining human and machine ratings remain essential to preserve fairness, especially for creative or high-stakes tasks. Emerging research on explainable AI may also help reveal why a model assigns a specific score, increasing trust among educators and students. Furthermore, advances in discourse analysis, multimodal assessment, and cross-lingual modeling promise to enhance the sophistication of writing evaluation. Future systems may analyze not only final essays but also drafting behaviors, revision patterns, and planning notes, enabling assessment of writing processes rather than just products.

NLP algorithms have radically transformed automated assessment of English writing, progressing from surface-level metrics to deep contextual models capable of evaluating meaning, coherence, grammar, vocabulary, and argumentation with impressive accuracy. While challenges remain regarding bias, transparency, creativity, and cultural variation, the growing integration of NLP in educational settings demonstrates substantial potential to improve writing instruction and assessment. When implemented responsibly, automated systems can complement human expertise, enhance teaching efficiency, and support learners in developing stronger writing skills. As NLP research continues advancing, automated writing assessment is likely to become even more accurate, fair, and pedagogically aligned with the complex demands of modern language education.

Automated assessment of English writing — especially when powered by advanced NLP — has showed both promising effectiveness and significant constraints. Empirical studies comparing automated scoring with human raters suggest that well-designed systems can approximate human judgments with decent reliability. For instance, a recent study evaluating an NLP-based automated essay grading (AEG) system on a variety of essay types (argumentative, narrative, descriptive, etc.) found a moderate to substantial agreement between automated scores and human raters across all essay types. This indicates that NLP-based systems are flexible enough to handle different writing genres — not only academic essays but also narrative or persuasive writing — which broadens their applicability in educational contexts.

Moreover, more refined, transformer-based or hybrid systems show even stronger performance. For example, a hybrid model combining shallow linguistic features, discourse patterns, and neural context embeddings has been shown to more accurately assess essay quality compared to models relying only on shallow features or purely on embeddings. Another advanced model, TransGAT — which integrates transformer-based embeddings with graph neural networks to model syntactic dependencies — reported very high agreement (quadratic weighted kappa ~ 0.854) on analytic scoring across dimensions such as coherence, vocabulary, and grammar. These results illustrate that when modern NLP systems are carefully designed, automated scoring can reach high levels of reliability, even on fine-grained measures beyond holistic grades.

In addition to scoring, automated systems increasingly provide corrective feedback. A recent research article proposed integrated systems combining automated writing evaluation (AWE) and grammatical error correction (GEC), offering not only a score but also suggestions on grammar, style, and coherence — thus supporting formative learning and helping learners iteratively improve. Meta-analytic evidence also supports the pedagogical benefits of AWE: a meta-analysis of 1993–2021 research found that AWE usage is associated with a moderate to large gain in writing performance among EFL/ESL learners (effect size $g \approx 0.59$ between groups, $g \approx 0.98$ within groups), particularly improving vocabulary usage, and benefits are stronger when AWE is used over medium to long durations and among learners of intermediate proficiency.This suggests that AWE isn't only a tool for grading — it can actively contribute to learner development if integrated appropriately in teaching.

Nonetheless, there remain key limitations and open challenges. A systematic review of AES research highlights that many systems — even those using advanced features — still struggle with content relevance, deep semantic adequacy, coherence & completeness, and domain knowledge; the majority of existing solutions implicitly rely on prompt-irrelevant heuristics or surface features. Another significant concern is creativity, argument originality, and critical thinking: automated systems often fail to appreciate unusual, novel, or culturally-influenced writing styles, metaphors, humor or subtle rhetorical devices. Studies of newer generative or large-language-model (LLM)-based scorers (e.g., ChatGPT) show mixed results. Some research finds acceptable agreement with human raters under certain conditions, especially for formative or low-stakes writing tasks. But others report significant divergence, especially in EFL contexts, and question their reliability for high-stakes assessment. Moreover, when models rely on shallow or generic features, they are vulnerable to “gaming” strategies:

for example, padding essays with long sentences or complex vocabulary without meaningful content — which may artificially inflate scores.

Given these findings, the role of automated writing assessment in educational practice should be seen as complementary, not replacement. Hybrid systems combining human and machine evaluation — or AI-assisted feedback followed by human review — provide a balanced pathway: machines handle volume, consistency, and basic feedback; humans handle nuance, creativity, and high-order judgment. Studies recommend this “human-in-the-loop” approach, particularly in high-stakes or creative writing contexts.

For EFL contexts such as Uzbekistan, where class sizes can be large and human scoring resources limited, integrating AWE systems — especially advanced transformer/hybrid ones — can support large-scale writing assessment and formative feedback. But to ensure fairness and educational value, institutions should: use diverse corpora (including EFL learner texts), monitor for bias, combine automated scoring with teacher oversight, and prioritise feedback generation over purely numeric scores.

In light of the above evidence and limitations, the following conclusions and recommendations arise:

1. Automated scoring using advanced NLP (transformers, hybrid models) is sufficiently reliable for many educational contexts. When properly trained and validated, these systems can approximate human grading on grammar, coherence, vocabulary, and structure.
2. Automated feedback (error correction, stylistic suggestions) is one of the most valuable advantages of AWE — it enables iterative learning and regular writing practice, which is especially beneficial for language learners.
3. Automated systems should not replace human raters — especially for high-stakes assessments or creative/argumentative writing. Rather, use a hybrid human-machine model to combine strengths.
4. Future AES/AWE research should focus on semantic relevance, domain knowledge, rhetorical and discourse quality, and fairness across different learner backgrounds. This includes building diverse datasets (including non-native, EFL contexts), developing multilingual or cross-cultural models, and integrating explainable AI to interpret scoring decisions.
5. In educational practice (e.g., in Uzbekistan), AWE should be used as a tool to supplement teaching — to provide frequent writing practice, instant feedback, and formative assessment — but not as the only evaluation method.

In summary, the application of NLP algorithms in automated assessment of English writing has matured significantly, offering scalable, consistent, and pedagogically useful tools. While obstacles remain — particularly around semantic depth, creativity, fairness, and high-order writing qualities — the trajectory of recent research suggests steady progress. With careful implementation, hybrid scoring practices, and ongoing research, NLP-driven writing assessment can become a powerful ally in language education rather than a mere technological gimmick.

References:

1.Attali, Y., & Burstein, J. Automated Essay Scoring with e-rater V.2. ETS Research Report. 2006.

2.Dikli, S. An Overview of Automated Scoring of Essays. Journal of Technology, Learning and Assessment. 2006.
3.Dong, F., Zhang, Y., & Yang, J. Attention-Based Neural Networks for Essay Scoring. COLING. 2017.
4.Page, E. B. The Imminence of Grading Essays by Computer. Phi Delta Kappan. 1966.
5.Rejapov, I., Mannonov, A., Saydaliyeva, G., & Gaybullaev, O. Automated Essay Evaluation Systems for Scientific Writing in Engineering Education. Archives for Technical Sciences. 2025.
6.Shermis, M. D., & Burstein, J. (Eds.). Handbook of Automated Essay Evaluation. Routledge. 2013.
7.Taghipour, K., & Ng, H. T. A Neural Approach to Automated Essay Scoring. EMNLP. 2016.
8.Vaswani, A., Shazeer, N., et al. Attention Is All You Need. NeurIPS. 2017.
9.Wang, I. X., Wu, X., Coates, E., Zeng, M., Kuang, J., Liu, S., & Qiu, M. Neural Automated Writing Evaluation with Corrective Feedback. arXiv. 2024.
10.Zesch, T., & Horbach, A. Neural Approaches to Modeling Coherence in Student Writing. TACL. 2021.