An Analysis of Automated Essay Scoring Frameworks

Abeer Abdulkarem[1], Anastasia Krivtsun[2]
[1,2]Department of Information Technologies and Management, Platov South-Russian State Polytechnic University (NPI),
Prosveshchenie Str. 132, 346428 Novocherkassk, Russia

Article Information

Article history:

Received: 23, 10, 2024
Revised: 25, 11, 2024
Accepted: 09, 12, 2024
Published: 30,12, 2024

Keywords:
Automated Essay Scoring,
Content Similarity,
Hybrid Framework,
Machine Learning,
Bert

Abstract

Automatic essay scoring (AES) has gained significant popularity in recent years as it provides an efficient and unbiased means of evaluating student writing. Efficient and objective assessment of student writing is essential for educators, as it offers valuable feedback that can aid students in enhancing their writing abilities and achieving academic success. This study offers an extensive examination of several AES frameworks, investigating their performance indicators, underlying algorithms, and suitability for use in a range of educational contexts. The paper examined the three main frameworks of AES, which include content-based, machine learning (ML), and hybrid methods, highlighting the benefits and drawbacks of each. In the end, this analysis hopes to improve automated grading technologies and their integration into educational practices by providing educators, policymakers, and technologists with information about the strengths and weaknesses of AES frameworks through the synthesis of recent research and developments.

Corresponding Author:

Abeer Abdulkarem
Department of Information Technologies and Management,
Platov South-Russian State Polytechnic University (NPI),
Prosveshchenie Str. 132, 346428 Novocherkassk,
Russia.
Email: abeerabdulsalam15@gmail.com

1. INTRODUCTION

Essay grading software is an inventive attempt to minimize the time and effort needed to grade written work while also getting rid of biases and inconsistencies in the evaluation process. The process of evaluating and scoring written prose works is known as AES, and it can be implemented as distributed services or standalone computer software. The main goal of AES is to solve the dependability problems, high prices, and time-consuming nature of manually grading essays. It is crucial to remember that AES is meant to be used as a tool for low-stakes classroom assessments, supporting teachers in their regular essay grading duties rather than to fully replace human assessors [1]. However. Three general frameworks can be used to classify AES systems content based, machine learning, and hybrid [2]. The Framework for Content based Measuring the semantic similarity between a student's writing and a model essay or reference response is the foundation of this system It assesses elements like Content's suitability to the prompt development of Concepts, Essay's coherence and cohesiveness, Domain-specific expertise. Typical methods include of metrics based on knowledge, such as Leacock & Chodorow similarity, Lesk similarity, and shortest path similarity. corpus-based techniques, such as Wikipedia or BNC corpora, and Latent Semantic Analysis (LSA). This content based has limitation it ignoring other crucial elements like language use, organization, and writing quality, they just pay attention to content overlap [3]. Essay grading just requires more than just matching terms or semantic similarity to a reference solution this leads to another issue when writings use relevant keywords repeatedly without developing the topic coherently, they leave themselves open to being "gamed" or tricked and also, they don't take into consideration syntactic components, which are essential for assessing writing ability thus include things like grammar, spelling, sentence structure, etc. As well as detection of off-topic, meaningless, or plagiarized essays is not a reliable function of content similarity algorithms that’s why the hybrid frameworks appear to overcome this limitation [4].

The second framework, ML is a branch of artificial intelligence (AI), focuses on creating algorithms that let computers analyze, interpret, and learn from data. ML systems use patterns and relationships found in data to generate predictions and judgments, as opposed to traditional programming, which involves a programmer explicitly coding rules [5].

Three primary categories of ML exist reinforcement learning, unsupervised learning and supervised learning Supervised learning involves training a model on a labeled dataset, meaning that each training example is paired with an output label. The model has to learn a mapping from inputs to outputs and be able to apply this mapping to previously unknown data [6]. Unsupervised learning works with unlabeled data and looks for underlying structures or hidden patterns in the data. Tasks involving clustering or dimensionality reduction frequently make advantage of this. the process of reinforcement learning, a machine gains the ability to make decisions by operating in a way that maximizes the concept of cumulative reward within its surroundings [7]. Essay scoring could be completely changed by ML, which presents a scalable and effective substitute for current grading techniques. The Educational Testing Service's (ETS) e-rater® essay grading system is a well-known example of ML in action. This system analyzes and scores essays according to different linguistic aspects using natural language processing (NLP) techniques. grammar, usage, mechanics, style, and structure are a few examples of these characteristics. The system has learned the qualities of excellent writing by being trained on a sizable dataset of essays that have been graded by humans [8].

Generally, there are multiple phases involved when using ML models for essay scoring preprocessing the text data, extracting pertinent features, training the model on labeled data, and assessing the model's performance. Tokenization, stemming, and the elimination of stop words are examples of preprocessing. Preprocessing is an essential stage in transforming unstructured text input into a format that is clear and organized so that ML models can use it. Several techniques for preprocessing and cleaning text data involves tokenization, stop word removal, stemming, lemmatization, and so on [9]. The tokenization is involves dividing the essay into smaller pieces known as tokens. These tokens can take the form of words, sub words, or characters, based on the level of detail needed. Stop words are often used to remove words that don't really add anything to the meaning of a text; they can be eliminated to lower the number of dimensions in the data. Stop words contain words like "the," "is," "in," "and," and "to.". Stemming is the process of eliminating prefixes and suffixes to reduce words to their most basic form. Although this approach is quicker and more straightforward, it may not always be accurate. Common stemming algorithms include the Porter Stemmer and Snowball Stemmer, Lemmatization techniques used to reduce words to their lemma, or standard form. Normalization is the process of formatting text to make it easier to read, process, and analyze. This stage involves managing special characters, increasing contractions, eliminating punctuation, and changing the text to lowercase. after preprocessing the text input extracting features AES is the next step. There are different feature extraction strategies from text data from easy to difficult [10].

The Bag of Words (BoW) model views text as an assemblage of word frequencies. Every document is transformed into a vector, with each element denoting the number of words contained within. BoW does not capture the order or context of words. author technique the term frequency-inverse document frequency, or TF-IDF is a better method than BoW it takes into account a word's value in a document in relation to its frequency in all documents. This method lessens the effect of often occurring words and helps find more informative terms and lastly Word Embeddings by expressing words in a continuous vector space, word embeddings, like Word2Vec and GloVe, capture the semantic links between words. These embeddings are effective for feature extraction since they are learned from large corpora and can catch subtleties like synonyms and analogies. And there is more advanced feature engineering like N-grams, Part-of-Speech Tagging (POS), Named Entity Recognition (NER), Topic Modeling and so on [11]. After the feature have been extracted. It is important to minimize the dimensionality of the feature space and choose the most relevant characteristics there is technique for this feature selection methods, this technique work by choose the most useful features, employ strategies like Chi-Square, Mutual Information, and ANOVA and Dimensionality Reduction and Singular Value Decomposition (SVD) and Principal Component Analysis (PCA) might be used. By using these methods, the majority of the data variance is retained while the original features are reduced to a smaller collection of uncorrelated components. Regularization by avoiding overfitting. The process of feature extraction determining and measuring the essay features that are important for grading [12]. Metrics like accuracy, precision, and recall are used to assess the model's performance once it has been trained on a collection of essays with known scores [13].

Although there are benefits to employing ML for essay scoring automatically, there are a number of drawbacks and restrictions to take into account. ML algorithms may unintentionally reinforce biases found in the training set, producing unjust assessments [14]. For instance, the model may generate biased results if the training data contains biases against particular student groups. One further difficulty is in the interpretability of the model's outcomes. Even though ML models can produce results with great accuracy, it might be challenging to comprehend the reasoning behind an essay's grade [15]. This lack of openness can cause issues, particularly in educational environments where teachers and students need to know why certain grades were given. Popular models include support vector machines, random forests, and regression models like linear regression and neural networks have been used in ML [16].

The hybrid framework mixes together ML techniques with content based to create a powerful combination. It makes use of (ML) models trained on the lexical, syntactic, and semantic characteristics of essays in conjunction with content similarity metrics [17]. The main concept is to use content and style data to improve essay score prediction. Preprocessing essays, obtaining pertinent features, vectorizing content, and training predictive models using a combination of content vectors and extracted features are all common steps in a hybrid AES approach [18]. The hybrid framework is thought to be promising since it can enhance AES performance by utilizing the benefits of both ML and content similarity [19]. The rest of this article is organized as follows: Section 2 provides background about Content-based framework, Section 3 describes ML Framework, Section 4 presents Hybrid Framework, in Section 5 we will discuss the challenges and we conclude the paper in Section 6.

2. Content-based framework

With the content similarity framework (CSF), essays are graded or scored according to how closely their scores match those of reference essays, see fig. 1. A set of human-graded reference essays encompassing the whole range of grades or scores on the relevant topics is the gold standard required by the framework. Figure 1 depicts the content similarity framework's process. This system involves a pre-processing stage wherein a subset of selected essays is subjected to tokenization, stop word removal, stemming, and lemmatization in order to minimize noise in the essays. Similarity can be determined by (a) syntactic indications, (b) semantic indicators, or by combining the two. The surface elements of the essay, such as word count, word connectors, part-of-speech, and stemmed words, are referred to as syntactic indicators. In contrast, the meaning of a word, phrase, sentence, and text is shown by the semantics indicators. It is well accepted that the semantic indicators are employed to support the semantic similarity of the entire or partial essays [20]-[24]. Spelling checks, stemming, lemmatization, word segmentation, n-grams [25]-[27], normalization [28]-[29], and word segmentation are frequent syntactic indicators employed in AES [30]. These surface characteristics of the writings have been demonstrated to be helpful in essay grading [31].

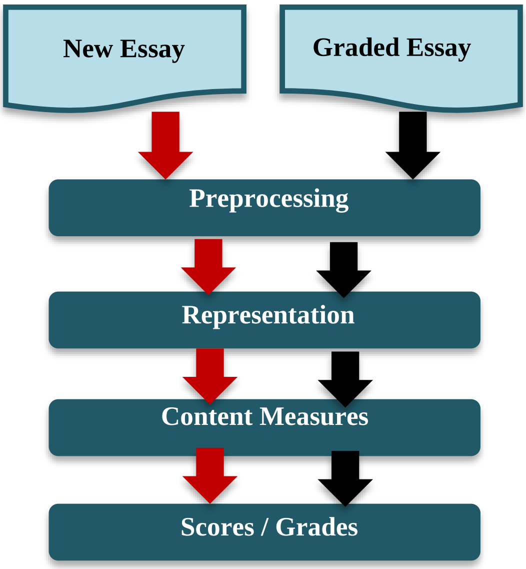

Figure 1: Content similarity framework

The workflow of a content-based AES system involves [32]:

1. pre-processing of essays (stop word removal, lemmatization, POS tagging, tokenization)
2. Essay content vectorization
3. utilizing semantic similarity metrics to contrast the essay vectors to the model answer vectors
4. The essay will be graded according on how closely its content resembles the model.

In this section, we provide an overview of the component the content based and show the workflow for framework and we explore some of algorithms in automatic essay scoring. AES systems perform by compared a student's essay with either a model essay or a reference response. The components of this system is the use of methods like word embeddings word2vec, glove, Bert RoBerta and latent semantic analysis (LSA), the system extracts semantic information from the student's essay [33]. To determine subject relevance, it calculates the semantic similarity between the student essay and the reference material. Lexical Similarity Measurement, that’s mean the lexical similarity between the student's essay and the reference text is ascertained by applying n-gram overlap, cosine similarity and longest common. This makes it easier to determine how closely the words and phrases in the student's essay match those in the reference material [34].

Scoring, an overall content similarity score for the student's essay is produced by combining the semantic and lexical similarity scores. According to the reference answer or rubric, this score indicates how well the essay adheres to the required material. Content-based methods are good at evaluating factual content, but they have limitation for grading more advanced levels writing abilities like structure and coherence [35]. Hybrid approaches integrate content similarity metrics with ML models that evaluate other writing aspects in order to overcome this constraint [36]-[37].

Additionally, based on three syntactic categories—rhetoric, organization, and contents—the Japanese Scoring System (JESS) provides an example of how to facilitate syntactic elements on essay grades (Ishioka & Kameda, 2006). Readability, the proportion of lengthy, distinct words, passive sentences, and neat presentation are used to quantify these qualities. Thus, in order to analyze its content, JESS used both syntactic and semantic clues. concept and subject-specific vocabulary. After then, the essay's score is calculated using the perfect score deduction method. Still, evaluating the essay's quality cannot be based only on its stylistic and syntactic choices. In order to do its content analysis, JESS employed both syntactic and semantic markers. Since various advances have shown that semantic indicators produce more accurate grades or ratings, AES that are just based on syntactic indications are becoming less common. A semantic space in natural language modeling seeks to provide natural language representations that are capable of capturing the context. The Vector Space Model (VSM) is the first known semantic space. It was developed to determine content similarity in essays by analyzing the co-occurrence of words [38].

Latent Semantic Analysis (LSA), which is excellent at determining content analysis, is one of the most widely used syntactic-blind semantics indicators [39]. A distributional model called LSA is employed to extract meaning from a text. "A theory and method for extracting and representing the contextual-usage meaning of words by statistical computations applied to a large corpus of text" was how LSA was described. Essays are translated into a term-document matrix using LSA, and this matrix is then roughly reduced using singular value decomposition (SVD). The goal of the dimension reduction process in LSA is to determine how likely it is that any given word will resemble any other word if it has ever been in another essay in a comparable situation. According to experiment results, AES scoring performance is enhanced when LSA is used in place of syntactic characteristics [40]. In order to determine grades or scores, several content similarity-based AES employed LSA or any of its variants [41]-[46].

However, it has been suggested that Generalized Latent Semantic Analysis (GLSA), an LSA variation that takes word sequence into account, might increase the effectiveness of AESs [47]-[48]. Syntactic and semantic aspects make up the majority of reported current AES developments. Two inputs were used in the study of essay grading using CSF: student answers and key answers [49]. After undergoing pre-processing steps such as noise removal, case conversion, tokenization, stop word removal, negation, conversion, stemming, and synonym conversion, the two essay inputs are represented by a term-document matrix in which each row denotes a term and each column a document. The term's occurrence in each document is represented by a cell in the matrix. A value of 0 indicates that the phrase is not present in the papers. SVD and the cosine similarity measure are the two steps in the LSA process that yield the most popular similarity computation. The term-document matrix is broken down by SVD into $D=U\Sigma V^T$. As $D\approx Uk\Sigma kV^T$ k, the k biggest singular values (dimensionality) are utilized to estimate D. Its goal is to find "latent" notions inside the matrix. After applying SVD to the key answers, each student response is put through the same pre-processing steps and compared to the key answers that are the most similar using the cosine similarity metric to calculate scores. Fig. 2 shows the essay grading process utilizing CSF [50].

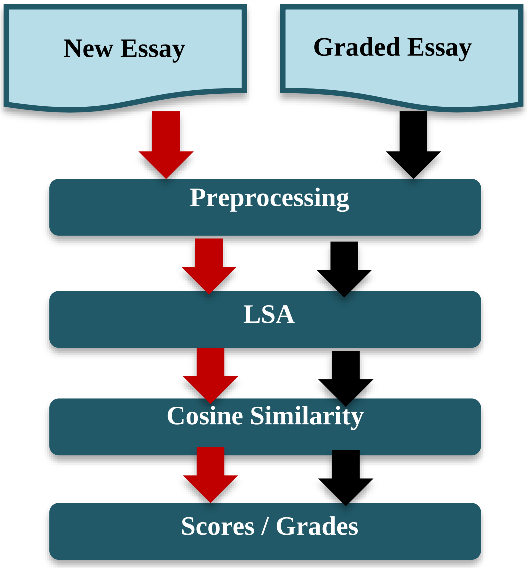

Figure 2: Content similarity framework using CSF [49]

3. Machine Learning (ML) Framework

Essay grading in the Machine Learning Framework (MLF) is handled as a multiclass classification issue, where each grade is represented as a class, as seen in Fig. 3. Computational functions are needed for modeling in order to generalize any article into multiple classes. The machine learning methods employed are primarily from the regression and classification categories since AES has been viewed as the document classification problem. Fig. 3 shows the machine learning framework workflow.

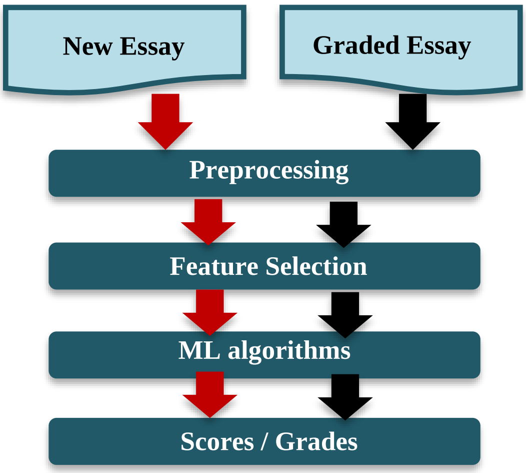

Figure 3: Machine learning [49]

The initial step that prepares the data and eliminates noise is called pre-processing. All writings will go through tokenization, stop word removal, stemming, and lemmatization procedures, much like CSF. Subsequently, the writings undergo processing to preserve noteworthy aspects using the Feature Selection procedure. According to [51], words, syntactic, and dependence aspects were typical MLF traits. In order to find a meaningful feature subspace in minimizing redundant features and complicated computational space and producing the best essay representation, feature selection is a crucial stage in many machine learning applications. In this case, feature selection limited the word count to avoid the dimensionality curse, which might gradually deteriorate classification accuracy. Feature selection or feature deletion are the two general dimensionality reduction methods. The chosen characteristics will be used as inputs to train machine learning models like Support Vector Machine (SVM), k-Nearest Neighbors (kNN), Naive Bayes, and Artificial Neural Network (ANN) after dimensionality reduction. MLF was used in the study described in [52] to determine essay scores. Tokenization, case conversion, and normalization of the essay score within the interval [0,1] are pre-processed for both the student responses and the key answers. The Enhanced AI scoring engine (EASE) is utilized for feature selection, and it may be used to determine length-based representation, POS, word overlap with key responses, and bag of n-grams. Then, the characteristics are put into machine learning algorithms to create student answer scores, which result in a marginal increase versus the baseline. These methods include support vector regression (SVR), Bayesian linear ridge regression (BLRR), and a variation of neural networks. Figure 4 shows the machine learning framework's process as described by [52].

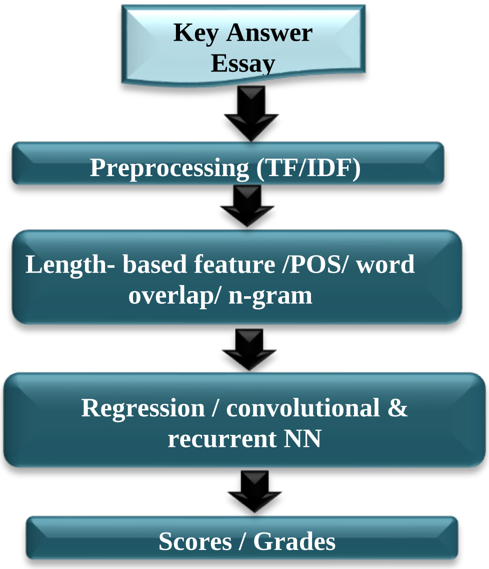
Key Answer Essay
Preprocessing (TF/IDF)
Length- based feature /POS/ word overlap/ n-gram
Regression / convolutional & recurrent NN
Scores / Grades

Figure 4. MLF Framework [53]

AES relies heavily on ML models since they automate the written content evaluation procedure. Large datasets of essays with past human rater scores are used to train these programs. ML models are able to recognize patterns and features that correspond to various score levels by gaining knowledge from these examples. They are able to predict grades for brand-new, unread essays with a high degree of accuracy thanks to this capacity. Furthermore, as new data becomes available, ML models may be updated and improved upon on a regular basis. This flexibility means that the models continue to be applicable and useful when evaluating modern writing subjects and styles [53]. In conclusion, ML models play a crucial role in AES by offering a scalable, reliable, and flexible way to assess written content, which improves the scoring process' overall effectiveness and dependability. As well as the ability of ML models in AES to process enormous amounts of material rapidly and reliably is one of its main benefits. In contrast to human raters, who could become exhausted and exhibit inconsistent scoring, ML models offer a uniform evaluation. This consistency is especially critical in educational environments, where impartial and fair assessment is essential [54]. Various ML model types, are used in AES each with unique advantages and disadvantages, these models all work differently and are best suited for different aspects of the writing assessment procedure. The unique criteria of the scoring assignment, such as the essays' complexity and the availability of training data, determine which of these models should be used in AES [55].

Linear regression are among the most often used in AEG because its easily understood and its operate by creating a linear connection between an essay's characteristics and its grade. Even though they are simple to use and comprehend, they might not be as good at identifying intricate patterns in the data as more complex models [56].

Support vector machines (SVMs) SVMs are accurate models able to managing data with many dimensions. They function by determining the best hyperplane to divide several essay classes. SVMs may perform well even with a small amount of training data, which makes them especially helpful for classification problems. They can, however, be computationally demanding, particularly when dealing with big datasets [57].

Linear Regression because they are easily understood that why frequently utilized. They operate by creating a linear connection between an essay's characteristics and its grade. Even though they are simple to use and comprehend, they might not be as good at identifying intricate patterns in the data as more complex models [58]. Decision Trees by dividing the data into branches according to feature values, decision trees simulate the process of making decisions. They are helpful for comprehending the grading criteria because they are simple to understand and imagine. Decision trees, however, are easily affected by overfitting, particularly when dealing with noisy data [59]. Neural Networks (NN) are the most popular in the AES because neural networks can represent difficult, non-linear relationships, especially deep learning models. They are built by numerous layers of networked nodes that can recognize complex patterns in the data. NN need a lot of computing power and training data in order to operate at high accuracy levels [60]. To ensure the efficacy and dependability of ML models in AES, it is vital to assess their performance on unseen data by using a number of metrics and methods, such as accuracy, precision, recall, and F1-score and cross-validation [61].

The accuracy of a score is determined by dividing the total number of essays by the percentage of accurately anticipated scores. It may miss the subtleties of scoring, particularly if the score distribution is skewed, even though it gives an overall impression of the model's performance Precision and Recall calculates the ratio of true positives to actual positives, whereas precision calculates the ratio of true positive predictions to all projected positives. These measures are very helpful in figuring out how well the model can distinguish between essays that are of a good and a poor quality. F score the F1-score is the harmonic mean of precision and recall, providing a single metric that balances both aspects. It is especially valuable when dealing with imbalanced datasets, as it gives a more comprehensive view of the model's performance [62]. Cross-validation this technique entails dividing the data into several subsets and using various combinations of these subsets to train the model. This method aids in evaluating the model's robustness and generalizability to make sure it functions well on fresh, untested data [63].

Weighted Kappa Quadratic (QWK) a statistical measure, is widely used as the primary evaluation metric for AES systems. It is used to assess the agreement between predicted scores from an AES system and human rater scores. QWK determines the degree of agreement between the two. This shows that greater departures from the expected and actual scores result in a heavier penalty than smaller differences. Ordinal scoring, such as 1-6, is commonly used for essay evaluation in AES. The QWK value is a number between 0 and 1, where 0 denotes no agreement between human raters and the AESsystem.1 denotes perfect agreement. Better performance of the AES system in simulating human scoring is shown by higher QWK values. To handle the issue of prevalence (i.e., when some scores are more prevalent than others), QWK also takes into account the distribution of scores over the rating scale. But additional study has shown that depending only on QWK has disadvantages, including its sensitivity to the rating scale, the risk of the "kappa paradox," and its inability to handle a high volume of raters. Researchers advise combining QWK with other evaluation metrics to provide a thorough analysis of the performance of the AES model [64].

Mean Absolute Error (MAE), the average absolute difference between the predicted scores and the scores provided by a human rater is calculated better performance can be seen by a lower MAE. Accuracy: This refers to the percentage of essays that the automated system and the human raters both assigned the same score. It is computed by dividing the total number of essays by the number of accurately anticipated scores [65]. Pearson's Correlation Coefficient The linear correlation between the anticipated scores and the ratings from human raters is measured by Pearson's correlation coefficient [66].

A stronger correlation is indicated by a greater value (closer to 1 or -1). Root Mean Squared Error (RMSE) The square root of the average squared difference between the expected and actual scores is determined using the Root Mean Squared Error (RMSE) method While MAE, Pearson's correlation, and RMSE measure the variance or inaccuracy between predicted and actual ratings, QWK and accuracy measures assess the agreement directly [67].

4. Hybrid Framework (HF)

The newly developed hybrid framework (HF) aggregates style and substance to determine essay scores or grades by combining the strengths of machine learning and content similarity. The machine learning algorithms in the framework are used to generalize syntactic features (indices, topics, and domain-specific keywords), where CSF is used to retrieve the closest key answer score in the semantic space. This is different from the general MLF, where machine learning algorithms are directly used to derive the grades or score. The hybrid framework's process flow is comparable to that of the ML framework, with the exception that essay scores and grades are determined using both content similarity measures and machine learning techniques. A basic workflow of the hybrid framework is shown in Fig. 5.

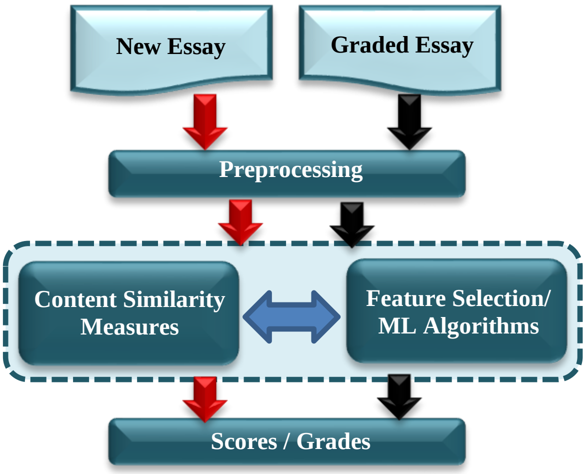

Figure 5: Hybrid Framework

Recent research using the hybrid framework has included the following: vectorization of essay scores using artificial neural networks and LSA [68]; two-step process classification using SVM and LSA Framework [69]-[70]and linear regression on specific features derived from an ontology and using LSA to measure content similarity [71]. Two-step grading was employed as the HF in the experiments published in [67]: SVM was used to categorize the essay's topic, and LSA was utilized to calculate the similarity between the student's responses and the key answers. The goal is to route the essay to the associated key answers essay and to exclude other topic essays using a pre-trained SVM model on the themes of the key answers. The goal of the LSA is to create a semantic space for key answer essays. To do this, each student essay score falls between 0 and 100, and the score for each essay is determined by comparing its key response score to the closest one using the Frobenius norm similarity metric. Comparing the research to human raters, significant accuracy (>95%) is reported. Figure 6 shows the hybrid framework's process as described in [67].

By evaluate essays' writing quality and content accuracy by combining methods from ML and content similarity This is how they normally operate the system takes lexical and semantic characteristics out of the essay in content similarity. Latent Semantic Analysis (LSA) is one method used to compute semantic similarity. Sentence embeddings BERT, RoBERTa and word embeddings (Word2Vec,GloVe) Lexical/string similarity is quantified using techniques such as: N-gram overlap Similarity of cosines LCS stands for longest common sequence. And ML algorithm extract aspects of the essay, such as its length, language use, syntax, discourse structure, etc. .these characteristics are given into ML models such as regression, random forests, and neural networks In order to give a more thorough evaluation of the essay, the hybrid technique makes use of the advantages of both ML (which captures writing quality elements like coherence, organization, etc.) and content similarity (which guarantees subject relevance) [72].

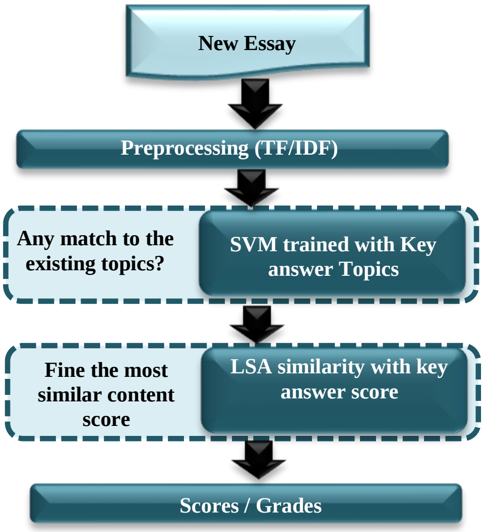
New Essay
Preprocessing (TF/IDF)
Any match to the existing topics?
SVM trained with Key answer Topics
Fine the most similar content score
LSA similarity with key answer score
Scores / Grades

Figure 5. Hybrid Framework [67]

A typical hybrid AES system's workflow consists of the following steps [73]:

1- essay pre-processing (stop word removal, tokenization, POS tagging, and lemmatization.
2- Extraction of features (word count, essay length, misspelled words).
3- Vectorization of essay material
4- Extracting syntactic characteristics through feature selection.
5- Using the syntactic characteristics and content vectors to train a prediction model

5. Discussion & Analysis

AES frameworks are developed to streamline the grading process, reduce biases, and enhance consistency in evaluation. Three prominent AES frameworks—Content-based, ML, and Hybrid—have been widely adopted. Below is a comparison highlighting the differences, strengths, and limitations of these frameworks, see table 1.

Table 1: Comparison Table: Key Features of AES Frameworks

Feature	Content-based Framework	Machine Learning Framework	Hybrid Framework
Core Principle	Measures similarity between student essays and reference texts based on semantic and lexical analysis.	Utilizes supervised/unsupervised ML models to classify essays or predict scores based on extracted features.	Combines ML techniques with content similarity metrics to capture both writing quality and content relevance.
Strengths	- Good for evaluating factual accuracy. - Straightforward implementation.	- Handles complex writing styles. - Capable of generalizing patterns from large datasets.	- Integrates strengths of both approaches. - Balances content relevance with writing quality analysis.
Limitations	- Ignores linguistic aspects like grammar, organization, and coherence. - Can be gamed using repetitive keywords.	- Requires large labeled datasets. - May unintentionally reinforce biases in training data.	- Computationally expensive due to the combination of frameworks. - Increased complexity in implementation.
Example Techniques	- Latent Semantic Analysis (LSA) - Cosine Similarity - Word Embeddings (Word2Vec, GloVe)	- Support Vector Machines (SVM) - Neural Networks (NN) - Random Forest	- BERT for semantic embeddings - Neural Networks for syntactic features - Combination of LSA and ML for evaluation
Evaluation Metrics	- Semantic similarity scores.	- Accuracy, Precision, Recall, F1-Score, QWK.	- Combined semantic and syntactic evaluation.
Notable Example	JESS (Japanese Scoring System) for semantic and syntactic indicators.	ETS e-rater® using NLP features for grammar, usage, and mechanics.	Two-step grading approach using SVM and LSA for thematic and content similarity.
Best Use Case	Low-stakes evaluations requiring factual correctness.	High-stakes evaluations with complex patterns in writing styles.	Comprehensive assessments involving both factual and linguistic quality analysis.

5.1 Content-based Framework:

- Suitable for scenarios where factual accuracy and relevance are paramount.
- Fails to evaluate linguistic nuances or writing structure effectively.

5.2 Machine Learning Framework:

- Highly scalable and flexible for diverse scoring requirements.
- Requires significant computational resources and labeled datasets.

5.3 Hybrid Framework:

- Best suited for holistic evaluation, combining the strengths of both approaches.
- The complexity of implementation and resource requirements can be a challenge.

6. Conclusion

The paper explores various Automatic essay scoring frameworks. Our work focuses on three key frameworks: content similarity frameworks, machine learning frameworks, and hybrid frameworks, each with its own set of benefits and challenges. From basic content matching to advanced predictive modeling, each of these solutions offers several advantages. Finally, including content and linguistic aspects into essay grading frameworks represents a step toward more complete and trustworthy automated essay scoring systems. This dual-focus strategy enhances assessment accuracy while simultaneously providing students with deeper feedback, perhaps improving their learning and writing abilities. As the discipline evolves, additional study and improvement of these methodologies will be critical to obtaining more complex and effective essay grading systems.

References

[1] Ramesh, Dadi, and Suresh Kumar Sanampudi. "An automated essay scoring systems: a systematic literature review." Artificial Intelligence Review 55.3 (2022): 2495-2527.

[2] Lin, Chenxi, et al. "Toward large-scale mapping of tree crops with high-resolution satellite imagery and deep learning algorithms: A case study of olive orchards in Morocco." Remote Sensing 13.9 (2021): 1740.

[3] Spiro, Rand J., and Jihn-Chang Jehng. "Cognitive flexibility and hypertext: Theory and technology for the nonlinear and multidimensional traversal of complex subject matter." Cognition, education, and multimedia. Routledge, 2012. 163-205.

[4] Fyfe, Paul. "How to cheat on your final paper: Assigning AI for student writing." AI & SOCIETY 38.4 (2023): 1395-1405.

[5] Zhang, Du, and Jeffrey JP Tsai. "Machine learning and software engineering." Software Quality Journal 11 (2003): 87-119.

[6] Hadsell, Raia, Sumit Chopra, and Yann LeCun. "Dimensionality reduction by learning an invariant mapping." 2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06). Vol. 2. IEEE, 2006.

[7] Mahadevan, Sridhar. "Average reward reinforcement learning: Foundations, algorithms, and empirical results." Machine learning 22.1 (1996): 159-195.

[8] Chen, Hongbo, and Ben He. "Automated essay scoring by maximizing human-machine agreement." Proceedings of the 2013 conference on empirical methods in natural language processing. 2013.

[9] Kathuria, Abhinav, Anu Gupta, and R. K. Singla. "A review of tools and techniques for preprocessing of textual data." Computational Methods and Data Engineering: Proceedings of ICMDE 2020, Volume 1 (2021): 407-422.

[10] Dhini, Bachriah Fatwa, et al. "Automatic essay scoring for discussion forum in online learning based on semantic and keyword similarities." Asian Association of Open Universities Journal 18.3 (2023): 262-278.

[11] Santoso, Joan, et al. "Named entity recognition for extracting concept in ontology building on Indonesian language using end-to-end bidirectional long short term memory." Expert Systems with Applications 176 (2021): 114856.

[12] Li, Feng, et al. "Automatic essay scoring method based on multi-scale features." Applied Sciences 13.11 (2023): 6775.

[13] Taghipour, K., & Ng, H. T. (2016). A neural approach to automated essay scoring. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (pp. 1882-1891). Association for Computational Linguistics. https://doi.org/10.18653/v1/d16-1193

[14] Packin, Nizan Geslevich, and Yafit Lev-Aretz. "Learning algorithms and discrimination." Research handbook on the law of artificial intelligence. Edward Elgar Publishing, 2018. 88-113.

[15] Kumar, Vivekanandan, and David Boulanger. "Explainable automated essay scoring: Deep learning really has pedagogical value." Frontiers in education. Vol. 5. Frontiers Media SA, 2020.

[16] Rodriguez-Galiano, Victor, et al. "Machine learning predictive models for mineral prospectivity: An evaluation of neural networks, random forest, regression trees and support vector machines." Ore Geology Reviews 71 (2015): 804-818.

[17] Dessi, Danilo. "Knowledge extraction from textual resources through semantic web tools and advanced machine learning algorithms for applications in various domains." (2020).

[18] Tashu, Tsegaye Misikir, Chandresh Kumar Maurya, and Tomas Horvath. "Deep Learning Architecture for Automatic Essay Scoring." arXiv preprint arXiv:2206.08232 (2022).

[19] Shin, Jinnie, and Mark J. Gierl. "More efficient processes for creating automated essay scoring frameworks: A demonstration of two algorithms." Language Testing 38.2 (2021): 247-272.

[20] Islam, M. M., & Hoque, A. L. (2013). Automated Bangla essay scoring system: ABESS. In 2013 International Conference on Informatics, Electronics and Vision (ICIEV) (pp. 1-5). IEEE Conference Publication. https://doi.org/10.1109/iciev.2013.6572694.

[21] Omar, N., & Mezher, R. (2016). A hybrid method of syntactic feature and latent semantic analysis for automatic Arabic essay scoring. Journal of Applied Sciences, 16(5), 209-215. https://doi.org/10.3923/ jas.2016.209.215.

[22] Sendra, M., Sutrisno, R., Harianata, J., Suhartono, D., & Asmani, A. B. (2016). Enhanced latent semantic analysis by considering mistyped words in automated essay scoring. In 2016 International Conference on Informatics and Computing (ICIC) (pp. 304-308). IEEE Conference Publication. https://doi.org/10.1109/ IAC.2016.7905734

[23] Landauer, T. K., Laham, D., & Foltz, P. W. (2000). The intelligent essay assessor. IEEE Intelligent Systems, 15, 27-31.

[24] Lim, Chun Then, et al. "A comprehensive review of automated essay scoring (AES) research and development." Pertanika Journal of Science & Technology 29.3 (2021): 1875-1899.

[25] Li, Jin. "Assessing the accuracy of predictive models for numerical data: Not r nor r2, why not? Then what?." PloS one 12.8 (2017): e0183250.

[26] Chen, Z., & Zhou, Y. (2019). Research on automatic essay scoring of composition based on CNN and OR. In 2019 2nd International Conference on Artificial Intelligence and Big Data (ICAIBD) (pp. 13-18). IEEE Conference Publication. https://doi.org/10.1109/icaibd.2019.8837007

[27] Xu, Y., Ke, D., & Su, K. (2017). Contextualized latent semantic indexing: A new approach to automated Chinese essay scoring. Journal of Intelligent Systems, 26(2), 263-285. https://doi.org/10.1515/jisys-2015-0048

[28] Taghipour, Kaveh, and Hwee Tou Ng. "A neural approach to automated essay scoring." Proceedings of the 2016 conference on empirical methods in natural language processing. 2016.

[29] Ratna, A. A. P., Kaltsum, A., Santiar, L., Khairunissa, H., Ibrahim, I., & Purnamasari, P. D. (2019a). Term frequency-inverse document frequency answer categorization with support vector machine on automatic short essay grading system with latent semantic analysis for japanese language. In 2019 International Conference on Electrical Engineering and Computer Science (ICECOS) (pp. 293-298). IEEE Conference Publication. https://doi.org/10.1109/ICECOS47637.2019.8984530

[30] Loraksa, C., & Peachavanish, R. (2007). Automatic Thai-language essay scoring using neural network and latent semantic analysis. In First Asia International Conference on Modelling & Simulation (AMS'07) (pp. 400-402). IEEE Conference Publication. https://doi.org/10.1109/ams.2007.19

[31] Ong, D. A., Razon, A. R., Guevara, R. C., & Prospero C. Naval, J. (2011, November 24-25). Empirical comparison of concept indexing and latent semantic indexing on the content analysis of Filipino essays. In Proceedings of the 8th National Natural Language Processing Research Symposium (pp. 40-45). De La Salle University, Manila

[32] Subathra, G., and A. Antonidoss. "A multiobjective based Encryption scheme for content-based addressing in Blockchain." 2021 5th International Conference on Electronics, Communication and Aerospace Technology (ICECA). IEEE, 2021.

[33] Naseem, Usman, et al. "A comprehensive survey on word representation models: From classical to state-of-the-art word representation language models." Transactions on Asian and Low-Resource Language Information Processing 20.5 (2021): 1-35.

[34] Bailey, Stephen. Academic writing: A handbook for international students. Routledge, 2014.

[35] Bae, Jungok, Peter M. Bentler, and Yae-Sheik Lee. "On the role of content in writing assessment." Language Assessment Quarterly 13.4 (2016): 302-328.

[36] Iehab A. K., Mohanad A. A., NEW APPROACH TO PREDICTION OF MEMORY LEAK IN HPC HIGH-PERFORMANCE COMPUTING BY USING MPI (MESSAGE PASSING INTERFACE). (2024). Iraqi Journal for Applied Science, 1(1), 1-8. https://doi.org/10.69923/IJAS.2024.010101

[37] Raj, Nisha S., and V. G. Renumol. "A systematic literature review on adaptive content recommenders in personalized learning environments from 2015 to 2020." Journal of Computers in Education 9.1 (2022): 113-148.

[38] Al-Jouie, M., & Azmi, A. (2017). Automated evaluation of school children essays in Arabic. Procedia Computer Science, 117, 19-22. https://doi.org/10.1016/j.procs.2017.10.089

[39] Israa A., Nuha S. M., Saja S. M., Optimizing Skin Disease Diagnosis using Metaheuristic Algorithms: A Comparative Study. (2024). Iraqi Journal for Applied Science, 1(1), 72-80. https://doi.org/10.69923/IJAS.2024.010108

[40] Mariam l., Nihad S. k.,Preview the predictive performance of the STR, ENN, and STR-ENN hybrid models. (2024). Iraqi Journal for Applied Science, 1(1), 9-23. https://doi.org/10.69923/IJAS.2024.010102

[41] Awaida, S. A., Shargabi, B. A., & Rousan, T. A. (2019). Automated Arabic essays grading system based on F-score and Arabic wordnet. Jordanian Journal of Computers and Information Technology (JJCIT), 5(3), 170-180. https://doi.org/10.5455/jjcit.71-1559909066

[42] Amalia, A., Gunawan, D., Fithri, Y., & Aulia, I. (2019). Automated Bahasa Indonesia essay evaluation with latent semantic analysis. Journal of Physics: Conference Series, 1235, Article 012100. https://doi. org/10.1088/1742-6596/1235/1/012100

[43] Alghamdi, M., Alkanhal, M., Al-Badrashiny, M., Al-Qabbany, A., Areshey, A., & Alharbi, A. (2014). A hybrid automatic scoring system for Arabic essays. AI Communications, 27(2), 103-111. https://doi.org/10.3233/ aic-130586

[44] Contreras, J. O., Hilles, S., & Abubakar, Z. B. (2018). Automated essay scoring with ontology based on text mining and nltk tools. In 2018 International Conference on Smart Computing and Electronic Enterprise (ICSCEE) (pp. 1-6). IEEE Conference Publication. https://doi.org/10.1109/icscee.2018.8538399

[45] ALI K., Kamaljit L., A Hybrid System Based on Three Levels to Hide Information using JPEG Color Images. (2024). Iraqi Journal for Applied Science, 1(2), 30-45. https://doi.org/10.69923/8rjqxq24

[46] Shehab, A., Faroun, M., & Rashad, M. (2018). An automatic Arabic essay grading system based on text similarity Algorithms. International Journal of Advanced Computer Science and Applications, 9(3), 263-268. https://doi.org/10.14569/IJACSA.2018.090337

[47] Darwish, S. M., & Mohamed, S. K. (2020). Automated essay evaluation based on fusion of fuzzy ontology and latent semantic analysis. In *The International Conference on Advanced Machine Learning Technologies and Applications (AMLTA2019) 4* (pp. 566-575). Springer International Publishing.

[48] Recchia, G., & Jones, M. N. (2009). More data trumps smarter algorithms: Comparing pointwise mutual information with latent semantic analysis. *Behavior research methods*, *41*(3), 647-656.

[49] 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 & Technology*, *29*(3), 1875-1899.

[50] Bashir, M. F., Arshad, H., Javed, A. R., Kryvinska, N., & Band, S. S. (2021). Subjective answers evaluation using machine learning and natural language processing. *IEEE Access*, *9*, 158972-158983.

[51] Zheng, M. (2024). *Using Data Preprocessing Techniques and Machine Learning Algorithms to Explore Predictors of Word Difficulty in English Language Assessment* (Doctoral dissertation, The University of Iowa).

[52] Rathnayakc, B. S. S., & Ganegoda, G. U. (2018, April). Heart diseases prediction with data mining and neural network techniques. In *2018 3rd International Conference for Convergence in Technology (I2CT)* (pp. 1-6). IEEE.

[53] Suskie, Linda. Assessing student learning: A common sense guide. John Wiley & Sons, 2018.

[54] Hilbert, Sven, et al. "Machine learning for the educational sciences." Review of Education 9.3 (2021): e3310.

[55] Hussein, Mohamed Abdellatif, Hesham Hassan, and Mohammad Nassef. "Automated language essay scoring systems: A literature review." PeerJ Computer Science 5 (2019): e208.

[56] Odeh al-awaida, Saeda Esmaile. "Automated Arabic essay grading system based on support vector machine and text similarity algorithm." Doctoral dissertation (2019).

[57] Sundaram, Arun C. A Comparison of Machine Learning Techniques on Automated Essay Grading and Sentiment Analysis. MS thesis. The Ohio State University, 2015.

[58] Harrison, Allan G., and David F. Treagust. "Learning about atoms, molecules, and chemical bonds: A case study of multiple-model use in grade 11 chemistry." Science education 84.3 (2000): 352-381.

[59] Maimon, Oded Z., and Lior Rokach. Data mining with decision trees: theory and applications. Vol. 81. World scientific, 2014.

[60] Gholami, Amir, et al. "A survey of quantization methods for efficient neural network inference." Low-Power Computer Vision. Chapman and Hall/CRC, 2022. 291-326.

[61] Vaiyapuri, Thavavel, and Adel Binbusayyis. "Application of deep autoencoder as an one-class classifier for unsupervised network intrusion detection: a comparative evaluation." PeerJ Computer Science 6 (2020): e327.

[62] Ishioka, T., & Kameda, M. (2006). Automated Japanese essay scoring system based on articles written by experts. In Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (pp. 233-240). Association for Computational Linguistics. https://doi.org/10.3115/1220175.1220205

[63] Meyer, Travis A., et al. "A user's guide to machine learning for polymeric biomaterials." ACS Polymers Au 3.2 (2022): 141-157.

[64] Doewes, Afrizal, Nughthoh Kurdhi, and Akrati Saxena. "Evaluating quadratic weighted kappa as the standard performance metric for automated essay scoring." 16th International Conference on Educational Data Mining, EDM 2023. International Educational Data Mining Society (IEDMS), 2023.

[65] Ayub, Mubbashir, et al. "A Jaccard base similarity measure to improve performance of CF based recommender systems." 2018 International Conference on Information Networking (ICOIN). IEEE, 2018.

[66] Yu, Han, and Alan D. Hutson. "A robust Spearman correlation coefficient permutation test." Communications in Statistics-Theory and Methods 53.6 (2024): 2141-2153.

[67] Çano, Erion, and Maurizio Morisio. "Hybrid recommender systems: A systematic literature review." Intelligent data analysis 21.6 (2017): 1487-1524.

[68] Mohammed, S. H., & Al-augby, S. (2020). Lsa & lda topic modeling classification: Comparison study on e-books. *Indonesian Journal of Electrical Engineering and Computer Science*, *19*(1), 353-362.

[69] Tian, X., Pavur, R., Han, H., & Zhang, L. (2023). A machine learning-based human resources recruitment system for business process management: using LSA, BERT and SVM. *Business Process Management Journal*, *29*(1), 202-222.

[70] Chen, Y. T., Chen, C. H., Wu, S., & Lo, C. C. (2018). A two-step approach for classifying music genre on the strength of AHP weighted musical features. *Mathematics*, *7*(1), 19.

[71] Arora, V. (2023). A Framework for Cloud-Based EHR Security Using Hybrid Cryptographic Methods of AES and ECC.

[72] Birla, Nayna, Manoj Kumar Jain, and Avinash Panwar. "Automated assessment of subjective assignments: A hybrid approach." Expert Systems with Applications 203 (2022): 117315.

[73] Lei, Jian, Quanwang Wu, and Jin Xu. "Privacy and security-aware workflow scheduling in a hybrid cloud." Future Generation Computer Systems 131 (2022): 269-278.

BIOGRAPHIES OF AUTHORS

	Abeer Abdulkarem: PhD student at Department of Information Technologies and Management, Platov South-Russian State Polytechnic University (NPI), Prosveshchenie Str. 132, 346428 Novocherkassk, Russia. Email: abeerabdulsalam15@gmail.com
	Scopus
	Anastasia Krivtsun: Associate professor at Department of Information Technologies and Management, Platov South-Russian State Polytechnic University (NPI), Prosveshchenie Str. 132, 346428 Novocherkassk, Russia. Email: anastasia.srstu@gmail.com
	Scopus