applied sciences

MDPI

Article

An Automated English Essay Scoring Engine Based on Neutrosophic Ontology for Electronic Education Systems

Saad M. Darwish [1,*], Raad A. Ali [2] and Adel A. Elzoghabi [1]

[1] Department of Information Technology, Institute of Graduate Studies and Research, Alexandria University, 163 Horreya Avenue, El Shatby, P.O. Box 832, Alexandria 21526, Egypt; adel.elzoghby@alexu.edu.eg
[2] Department of Computer, Ministry of Education, Baghdad 10082, Iraq; raad.noor885@gmail.com
* Correspondence: saad.darwish@alexu.edu.eg; Tel.: +20-12-2263-2369

Abstract: Most educators agree that essays are the best way to evaluate students' understanding, guide their studies, and track their growth as learners. Manually grading student essays is a tedious but necessary part of the learning process. Automated Essay Scoring (AES) provides a feasible approach to completing this process. Interest in this area of study has exploded in recent years owing to the difficulty of simultaneously improving the syntactic and semantic scores of an article. Ontology enables us to consider the semantic constraints of the actual world. However, there are several uncertainties and ambiguities that cannot be accounted for by standard ontologies. Numerous AES strategies based on fuzzy ontologies have been proposed in recent years to reduce the possibility of imprecise knowledge presentation. However, no known efforts have been made to utilize ontologies with a higher level of fuzzification in order to enhance the effectiveness of identifying semantic mistakes. This paper presents the first attempt to address this problem by developing a model for efficient grading of English essays using latent semantic analysis (LSA) and neutrosophic ontology. In this regard, the presented work integrates commonly used syntactic and semantic features to score the essay. The integration methodology is implemented through feature-level fusion. This integrated vector is used to check the coherence and cohesion of the essay. Furthermore, the role of neutrosophic ontology is investigated by adding neutrosophic membership functions to the crisp ontology to detect semantic errors and give feedback. Neutrosophic logic allows the explicit inclusion of degrees of truthfulness, falsity, and indeterminacy. According to the comparison with state-of-the-art AES methods, the results show that the proposed model significantly improves the accuracy of scoring the essay semantically and syntactically and is able to provide feedback.

Keywords: automated essay evaluation; neutrosophic ontology; semantic analysis; knowledge representation

Citation: Darwish, S.M.; Ali, R.A.; Elzoghabi, A.A. An Automated English Essay Scoring Engine Based on Neutrosophic Ontology for Electronic Education Systems. *Appl. Sci.* **2023**, *13*, 8601. https://doi.org/10.3390/app13158601

Academic Editor: José Machado

Received: 6 June 2023
Revised: 15 July 2023
Accepted: 22 July 2023
Published: 26 July 2023

1. Introduction

Studies on computer-based assessment (CBA) and automated marking systems (AMS) are booming because educators in a wide range of businesses are becoming more and more interested in both. Automated grading has the potential to improve academic honesty by making assessments more consistent and reliable. It could also reduce teaching expenses, facilitate remote learning, provide students with immediate feedback, and cut response times [1]. An essay is the best way to evaluate a student's progress towards learning goals. This is so because it suggests cognitive abilities such as memory, planning, and analysis. Writing an essay helps students develop their communication skills as well as their ability to analyze and implement information. It is a tedious and time-consuming procedure, but essays and open-ended queries must be graded. Assumptions have been made about the usefulness of machine learning in this area, specifically in the form of automatic essay assessment systems that can assist educators [2].

Essays are graded and evaluated by computers in a procedure known as automatic essay evaluation (AEE). Essay assessment is not the same thing as essay marking. Comparatively, the essay assessment provides students with both a grade and detailed comments on how to improve for next time. AEE is a valuable resource for educators because it not only facilitates the evaluation of student progress but also reduces costs and workload without compromising instruction quality. Numerous text mining applications can profit from using an AEE system to rate and highlight the text's substance, including cover letters, scholarly papers, email classification, and so on. An example of such applications is in the field of text mining in healthcare, where AEE can be used to rate the prescription that the doctor wrote to treat a specific disease according to the reference approved in this case for writing the prescriptions. After that, a text mining algorithm can be used to examine large collections of doctors' perceptions to discover new information or help answer specific questions. To date, most attempts at automating essay scoring (AES) have centered on identifying a suitable machine learning (ML) algorithm and identifying relevant features for teaching the ML algorithm for AES [3,4]. The process of grading essays is depicted in Figure 1.

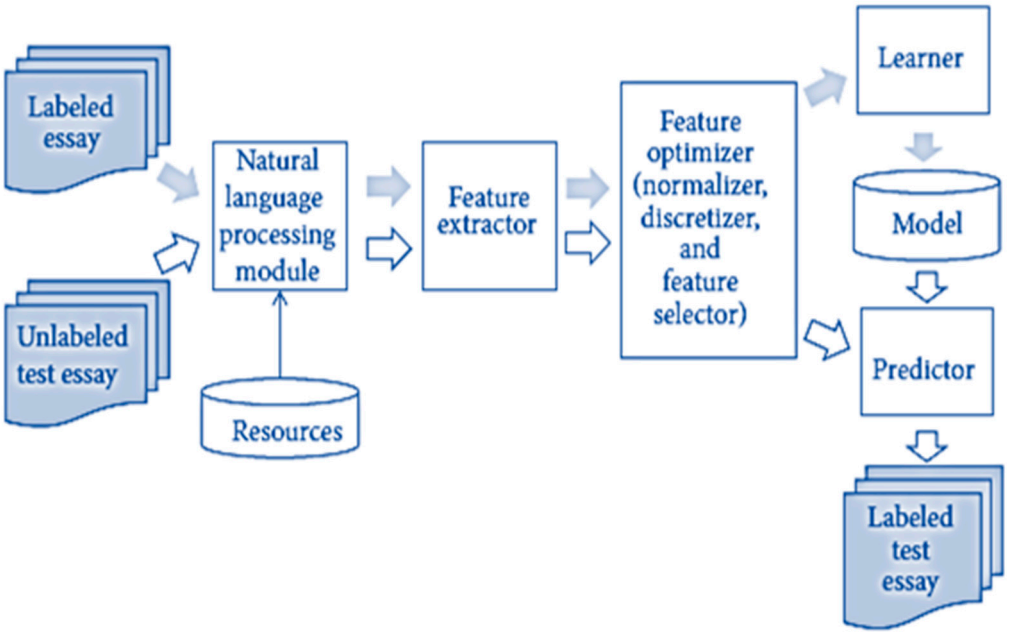

Figure 1. Essay scoring system.

The current AES systems regularly encounter difficulties. To begin, evaluating essays is challenging because of linguistic ambiguities and the fact that there is no single "correct" response to any given query. To continue, e-learning material items and e-learning platforms have distinct connection infrastructures. Third, an effective essay evaluation system must gather data about the student's knowledge. Information retrieval, data mining, and natural language processing systems have all benefited from word-based and statistical methodologies, but there is still an urgent need for deeper text understanding. A fourth area where improvement in textual information handling is needed is in the area of ideas, semantic connections, and the surrounding information required for the notion of disambiguation. Finally, the method needs to be trustworthy and practical for educators [5–7].

The current AES approaches take numerous perspectives into account when analyzing the essay. Some of them give the essay a grade based on its quality as described by a set of characteristics that can be roughly categorized as style, content, and semantic attributes [8–11]. The vocabulary, syntax, and techniques (spelling, capitalization, and punctuation) are the main focuses of style. Most systems for evaluating writing styles rely on statistical techniques,

which produce reliable outcomes but struggle to handle morphology when evaluating style attributes. An essay's content characteristics are founded on comparisons to the essay's source text and to other essays that have already been evaluated, and they provide only a hazy description of the essay's semantics. The foundation of semantic characteristics lies in ensuring that the content's meaning is accurate. Attributes from essays can be extracted using several different methods and tools. The most popular approach uses natural language processing (NLP)-based techniques. Latent semantic analysis (LSA) is widely used by content-centric systems; it is a machine learning technique for analyzing the relationships between a corpus of texts and the words within them. Though it excels at finding connections between documents, it ignores the sequence of words. Although these strategies properly ground the essay in its subject area, they do little to address the essay's ambiguous language.

Latent semantic indexing (LSI), a kind of latent semantic analysis (LSA), is the process of examining documents to extract their hidden meanings and ideas [12]. Even though LSA was not designed to be a teaching aid, it was easy to see how its power to assess lexical similarity could be used for essay grading. A high degree of semantic similarity between the essay and the text (syllabus) is desirable given that essays are usually intended to evaluate a student's knowledge and that this knowledge is often acquired through examining a text. Since LSA is not dependent on any particular auxiliary structures, it can be used to extract the meaning and substance of documents expressed in any language. One major drawback of LSA is polysemy. This means it cannot serve as a representation of the term if it can have multiple interpretations [13].

Ontologies provide a common language for describing the meaning of a topic, so information can now be automatically processed to figure out what it means. Communities concerned with lexical similarity have taken a keen interest in ontology in recent years [14]. This is due to its provision of a conceptualization of knowledge that is linked through semantic connections. The use of ontology is not without its restrictions. One of these is the challenge of effectively moving information learned from text to more abstract and philosophical concepts. Second, ontology has trouble dealing with and comprehending ideas whose meaning and connections are unclear. In this way, fuzzy theory can be extremely useful. In contrast to the clear-cut ideas of a topic ontology, a fuzzy ontology includes a wider range of possible interpretations. For the purposes of resolving uncertainty reasoning issues, domain knowledge descriptions are preferable over ontologies of the domain [15–18].

With the help of fuzzy ontology, the search can be broadened to include all relevant concepts, eliminating the need for exact terminology in order to obtain a usable result (since the context of a document need not be identical for the user to learn from it). Fuzzy semantics, the investigation of fuzzy ideas and fuzzy words, is another benefit of fuzzy ontologies because of their adaptability in moving from one ontology to another [14,18].

Fuzzy ontologies are more accurate representations of reality because they allow for gradations of quality in assessments; however, this increased realism comes at the expense of increased indeterminacy, which can be addressed by employing neutrosophic logic [19]. It is possible to include varying degrees of truth, falsity, and indeterminacy in neutrosophic logic [20]. The work presented in this paper utilizes a neutrosophic ontology for representing the knowledge of essays' features, with the end goal of developing a fully automated scoring engine for English writing assessments.

1.1. Problem Statement

The purpose of essay evaluation is to autonomously assign a grade to a student's essay based on several characteristics. These systems analyze the linguistic characteristics of text fragments and take text semantics into consideration in an imprecise manner. Current AES systems are syntax-centric, meaning that they use measures of text similarity but pay no attention to semantics. While much progress has been made, the latter technologies are still not fully automated. At least in the beginning stages, present AES systems call for human input. Constructing an appropriate matrix depiction of word use or frequency also

necessitates a large quantity of data, and the complexity of the calculations involved is high because of the large size of the involved matrices. In addition, current AES systems do not offer error feedback, which is crucial for training purposes.

1.2. Motivation of the Work

Exams based on essays are generally seen as an important part of the learning process. Teachers benefit from having this information because it allows them to better understand their students' progress and challenges. However, when a lot of exams need to be graded at once, the teacher finds himself too busy and unable to give students high-quality feedback in a reasonable amount of time. One of the problems with essay exams is what is generally referred to as bias, which means that different teachers of the same topic can provide wildly varying evaluation ratings for the same applicant. Many methods were tried and found to be very successful in automatically evaluating essays. However, they still face problems. Most of these techniques concentrate only on grammar and spelling mistakes, but they do not pay attention to semantics. Motivated by the above technical problems and the evolution of technological systems, especially natural language processing (NLP), there is a rising concern about the use of semantic AES systems to minimize the time, money, and effort spent by institutions.

1.3. Contribution and Methodology

The novelty of the proposed automated essay evaluation model is that it utilizes a neutrosophic ontology to check for semantic errors in the essay and give feedback to the student. Furthermore, with the aim of enhancing the traditional syntax-based AES, the suggested model utilizes coherent measures to check the global semantics of the essay and track the flow of the text's semantic changes. Integrating neutrosophic reasoning within the ontology building process helps to overcome the problem of concepts' vagueness. Neutrosophy enables us to evaluate and classify objects by considering their truthfulness, falsity, and indeterminateness about their membership in a certain set. The work presented in this paper presents the first effort to embed neutrosophic ontology in AES systems to provide feedback with the score. This model can use more attributes to enhance its scalability.

The rest of this paper is detailed as follows: Section 2 summarizes some preliminaries related to fuzzy ontology and its extension version, neutrosophic ontology that represent the main component of the suggested model. Section 3 provides a literature survey of works related to essay evaluation systems. Section 4 presents the proposed neutrosophic ontology-based AES model. Section 5 reports the evaluation of the proposed model along with the results and the discussion. Finally, Section 6 contains the conclusion of the work and the directions that could be taken in future works.

2. Preliminaries

2.1. Fuzzy Ontology

An ontology, in the field of computer science, is a systematic description of the information stored in a collection of concepts within a topic and the connections between those concepts. Those who study semantic similarity have taken a keen interest in ontology recently [14]. This is because it offers a coherent, semantically linked idea as a depiction of the information. An ontology is a directed acyclic network with a root node, in which ideas are organized in a hierarchy framework [15]. Concepts with less depth, those closer to the root, have a wider meaning, while those with more depth, those further from the root, are hyponyms with more particular meanings. A domain is represented by an ontology, which has four essential parts: (1) Entities in a domain are represented by concepts; (2) A relationship describes how concepts interact; (3) In the context of the domain, an instance represents a specific illustration of a concept; (4) Axioms are used to indicate a proposition that is always true [16].

Ontologies may be conceptualized as a directed graph in which concepts are connected to one another primarily by taxonomic (is-a) and, in certain situations, non-taxonomic links. Calculating similarity is as easy as finding the shortest route between two ontology nodes using is-a links, which is achieved by mapping input words to ontology concepts using their textual labels [17,18]. In this case, $path(a, b) = l_1, \ldots .l_k$ is the set of all possible links between the taxonomic words a and b. Determine the length of this route by setting $|path(a, b)| = k$. The semantic distance $dis(a, b)$ between a and b is thus defined as when all feasible routes between a and b are considered [21,22]:

$$dis(a, b) = \min_{\forall i} |path_i(a, b)| \quad (1)$$

Formally, an ontology is a 4-tuple $O = (C, P, R, A)$ where (1) C is a collection of domain-specific concepts. Ontologies often treat concepts as classes. (2) P represents certain properties of concepts. In this definition, $p \in P$ is an instance of the ternary relation $p\,(c, v, f)$, where $c \in C$ is an ontology concept, v is a property value linked with c, and f provides constraint on v. Some of the restrictions are: type $f_t \in \{\text{boolean, integer, float, string, symbol, instance, class, } \ldots\}$. Ontology editors often work with a uniform set of data types; the upper and lower bounds of the property's values are specified by the cardinality f_c; the notation range f_r denotes an allowable range of values for the property. (3) $R = \{r | r \subseteq C \times C \times R_t\}$ is a collection of concepts and their corresponding binary semantic relations as specified in C. $R_t = \{\text{one} - \text{to} - \text{one, one} - \text{to} - \text{many, many} - \text{to} - \text{many}\}$ is the type-relationship sets. A set of fundamental relations is defined as {synonym of, kind of, part of, instance of, property of} $\subset$ R. (4) A is a set of axioms. A true truth or reasoning rule is called an axiom. See [15] for more details.

What follows are some of the most common problems with ontology languages and how to prevent them [15]. (1) Due to the lack of information, the quality of the ontology-based model's information retrieval outputs is worse; (2) The development of an ontology language comes at a high cost; (3) Ontologies have a limited number of facts and axioms; (4) The language's vagueness is something it is unable to deal with. Ontology and metadata imperfections should thus be considered. By avoiding imperfections and making assumptions about the quality of ontologies and metadata, we may develop new ontologies by showing how semantic information is used in the domain [19–23].

On description logics, crisp ontologies are built. As a result, they cannot deal with imprecise information easily [15]. To enhance ontologies' capacity for representation and provide exact representation of imprecise data, a fuzzy ontology is utilized. An ontology that employs fuzzy logic to create a more intuitive representation of imprecise and vague knowledge for the purpose of facilitating reasoning about it is called a fuzzy ontology [22]. The idea of fuzzy relations is first introduced in fuzzy ontologies. Degrees of relationship between concepts or individuals may be stated in the interval [0, 1], allowing for a variety of possible degrees of relationship. To construct the imprecision connection, a fuzzy membership function may be used in this manner [14]. The modification is entirely incremental; conversion to a fuzzy ontology adds membership values to the currently existing relations and may also add new entries in the ontology. The ontology membership is normalized with respect to each of the terms in the ontology; that is, the sum of the membership values of each term in the ontology is equal to 1. The fuzzification process of the crisp ontology is shown diagrammatically in Figure 2.

Fuzzy ontology is the hierarchical description of every instance with its fuzzy membership value belonging to a class. A fuzzy domain ontology is a 4-tuple $O_F = (C, P_F, R_F, A_F)$, where (1) A collection of concepts is called C. Each concept has certain properties here that have the values of a fuzzy concept or fuzzy set, which differs from the notion of a crisp ontology. (2) A group of properties is called P_F. A property $p_F \in P_F$ is described as a 5-tuple of the form $p_F(c, v_F, q_F, f, U)$, where $c \in C$ is an ontology concept, v_F is a representation of property values, q_F is a model of linguistic qualifiers that can control or modify the strength of a property value v_F, f is a representation of restriction facets on v_F, and U is the

universe of discourse. Both v_F and q_F are fuzzy concepts at U, but q_F modifies v_F's degree of fuzziness. (3) R_F is a set of inter-concept relations between concepts. $r_F \in R_F$ defined as a 5-tuple of the form $r_F(c_1, c_2, t, S_F, U)$, where $c_1, c_2 \in C$ are ontology concepts, t represents relation type, S_F models relation strengths and is fuzzy concept at U, which can represent the strength of association between concept-pairs $\langle c_1, c_2 \rangle$. (4) A_F is a set of fuzzy rules. In a fuzzy system the set of fuzzy rules is used as knowledge base. Table 1 shows the difference between fuzzy and classical (crisp) ontology [20].

Table 1. Differences between fuzzy and classical ontology.

Aspect	Fuzzy Ontology	Crisp Ontology
Multiply-located terms	Does not occur	Issue for disambiguation
Query expansion	Depends on membership value	Depends on location value
Customization	Modification of membership value is the only requirement	Requires new ontology and/or sharing of ontology
Intermediate locations for grouping	Unnecessary	Construction necessities—may be useful
Storage required	Can be smaller or larger than crisp ontology based on the number of terms in the ontology and the membership values of the relation.	Depends on the number of terms contained in the ontology.

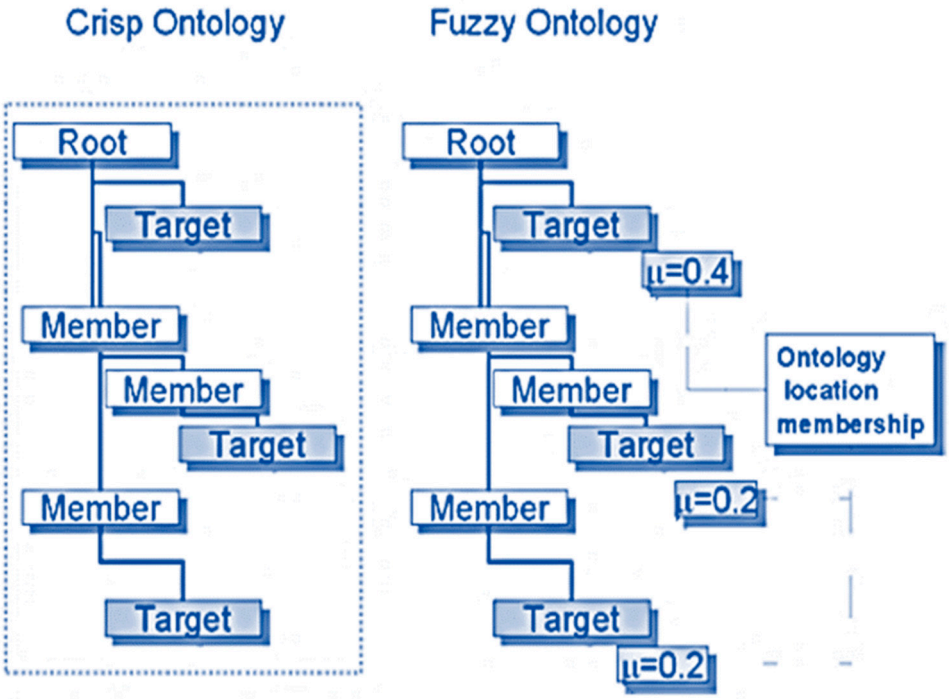

Figure 2. Membership values in a fuzzy ontology.

An extended fuzzy ontology is a 7-tuple $O_F = (c_a, C_F, R, F, Q, O, U)$ where c_a, C_F, R, F, U have same interpretations as defined in fuzzy ontology and Q is the set of the linguistic qualifiers. A qualifier $q \in Q$ and a fuzzy concept $c_F \in C_F$ compose a composition fuzzy concept that can be the value of c_a. O is the set of fuzzy operators at U, which is isomorphic to Q. Most academics analyze applications that incorporate fuzzy ontologies or use the standard techniques for evaluating crisp ontologies while evaluating fuzzy ontologies. Only one study, [24], in which a unique metric is established as follows, is clear about assessing fuzzy ontologies.

$$K = \sum_{(p \neq 0)} p/n - \sum_{pi=0} 0.1/n - \log(n_1/n) \quad (2)$$

where n is the total number of ontology terms, and n_1 is the total number of valuable domain terms that were extracted from selected text documents but are not ontology terms; p stands for the value of probability of (sub) property of ontology element, which indicates the probability that this element locates in the correct position. The learning process is said to be functioning effectively when metrics are increasing, and poorly when they are dropping [22–26].

2.2. Neutrosophic Ontology

In real life, indeterminacy can be seen everywhere. A neutrosophic ontology is a sextuple $O_N = (I, C, T, N, X, indeterminacy)$ where I and C are the sets of instances and classes, respectively. Taxonomy relations among the collection of concepts C are denoted by T. The set of non-taxonomy neutrosophic associative relations is indicated by N. The logical language in which a collection of axioms is stated is called X. The degree of indefiniteness present in the overlapping area is known as indeterminacy [19,20]. Neutrosophic ontology is the integration of traditional and fuzzy ontologies. Neutrosophic ontology is a hierarchical description of every instance belonging to an overlapping region. Every instance will reveal the level of truth, indeterminacy, or falsity, allowing us to assess the probability that it falls into a certain category. The concept of neutrosophic ontology is the same as that of fuzzy ontology, except after defuzzification, the output value is expressed as a triplet, i.e., truthness, indeterminacy, and falsity [19].

Inconsistency and indeterminacy are different situations. For example, a proposition can sometimes be true or false; this is an inconsistency. Sometimes we cannot get accurate results about a problem because of indeterminacy. As such, in an intuitionistic fuzzy set, hesitancy specifies uncertainty, and a neutrosophic set models inconsistency. Namely, in a neutrosophic set, results are accurate but inconsistent; in fuzzy sets, some results have incomplete information. Logically, indeterminacy function (I) is a complement of the member and non-member functions (T, F); that is, indeterminacy function explicitly exists in neutrosophic sets. It should note that indeterminacy function takes a value >0 when available information of phenomenon or problem are incomplete.

3. Related Work

Though the automated scoring systems that are already in place have worked well in the past, they all have the same major flaw. None of these methods made an effort to weight students' responses based on their importance. These methods instead combined NLP-based algorithms, statistics, probabilities, and other data sources such as corpora to arrive at a more accurate approximation of a student's score. Semantic similarity research was widely implemented in essay grading [27–29]. However, they require a massive annotation effort in terms of cost, time, and expertise. In addition, semantic parsing is not fully automated [30]. In [31], the K-Nearest Neighbors (KNN) was used as a categorization method for grading essays. This algorithm uses a dictionary that understands both Thai and English to break down text into individual responses to numerous keywords and phrases. Next, they combined the matching and word similarity methods with the KNN algorithm. KNN performs well when there are few input variables, but it becomes increasingly difficult for the algorithm to anticipate the outcome of a new data point as the number of input variables increases.

When the idea of ontology became widely available, it was put to use in grading essays [32–34]. In [35], the authors evaluate essays using a system that combines semantic matching, ontology creation, information extraction, and the Unified Modeling Language. The method takes advantage of WordNet to calculate the depth of the Least Common Sub-Summer, a metric that gauges the semantic relatedness between material taken from the ungraded essay and some other graded essays. The use of a large collection of already-scored essays helps keep the system's initial training costs low. In addition, the system's ontology-based structure and the use of layers (a presentation layer, a business layer, and a data layer) facilitate the provision of insightful input to students. The primary issue with

the model is that it cannot assess essays whose subjects deviate from those picked, as the taxonomy only includes knowledge about those topics.

Deep learning, defined as multi-layer neural networks that autonomously acquire valuable characteristics from data, was brought to the area of essay grading by the authors in [36]. In this instance, the lower layers learn the receptors for more concrete characteristics, while the upper layers learn more abstract features at a higher level. Using neural network design, the system discovers a dispersed illustration for each word in a collection according to its local context. The automated feature extraction in this system makes it more cost-effective to teach the evaluation system on a large collection of already-graded essays. However, it struggles to isolate the textual features that the network has determined to be most informative. In addition, it does not engage with semantics but rather tests for correct grammar and syntax.

Latent semantic analysis (LSA) has been widely adopted in the area of essay grading in recent years. It was used sometimes with integration with fuzzy logic and sometimes with clustering techniques. When it is integrated with clustering, it achieves some success as it does not need pre-specified classes or categories, so it does not need pre-training. The fuzzy function assigns LSA a fair amount of weight. Even so, it does have some disadvantages. As it stands, LSA is unable to determine whether repeated words are useful. LSA only cares about precise word similarities and ignores how those words are used in context [37]. Another work [38] integrated LSA with the K-nearest neighbor algorithm in essay scoring. The only advantage of this approach was that the correlation degree checked similarity between paragraphs and within paragraphs, so it checked cohesion and coherence. Yet, distance between neighbors is dominated by irrelevant features, and the weights of some words, such as modal verbs, are small because they are distributed evenly among the classes, which can affect the result.

In [39], N-Gram was integrated with cosine similarity in an AES system. It weighs terms using term frequency-inverse document frequency (TF − IDF). The level of agreement between the teacher's responses and the students' responses is calculated. Exam scores will be determined by the cosine value of this procedure. N-Gram is used to overcome the default of cosine similarity, so word order is taken into consideration. When compared to other methods, such as LSA and similarity by cosine, server performance was shown to be favorable due to reduced CPU consumption and accelerated load times. Despite this, the system remained stuck with the drawbacks of N-gram and cosine similarity. N-gram has also been proven to be better for calculating similarity in short sentences only, and cosine similarity does not take word order into consideration.

Using a fuzzy semantic optimal control retrieval approach to extract useful data from an English information table was proposed by the authors in [40] to improve the AES model's accuracy. This study uses a two-stage training method. First, standard English-to-English translation data were used to build the model. Then, English-to-English translation data, which include fuzzy semantic information, are used to adjust the model for the best control. The model is split into two parts: a sequence-generating network that can forecast the probabilities of individual words and an assessment network that may predict the final results of sentences. In order to increase the reliability of automatic writing scoring systems, the authors in [41] devised a methodology to use deep learning methods in a wireless network environment. The results of the experiments show that the technique is effective in evaluating the writing skills of English as a second language (ESL) students, opening the door to the development of an automatic composition grading system for use in massive online machine exams and self-study websites.

The primary goals of the study reported in [42] are the proposal and training of a long-short-term memory (LSTM) model on a collection of hand evaluated essays with scores. The purpose of sentiment analysis is to establish whether an article is positive, negative, or neutral. A sentiment algorithm is constructed using the Twitter sample dataset, and the student's opinion towards a subject is used to analyze the sentiment. All essays also go through a plagiarism check and a check for grammatical mistakes to ensure originality.

The essay's quality, number of syntactical mistakes, proportion of copying, and general tone all factor into the final score. Students receive constructive criticism in the form of a revised essay.

For automatically scoring spoken Chinese-English translation, the authors of [43] propose using attention LSTM. Semantic terms, phrase drift, and naturalness of speech were prioritized as primary criteria for evaluation. In this article, we employ the synonym discrimination technique to isolate the terms that are essentially the same in all the exam takers' answers. The attention LSTM model is applied at the sentence level in order to analyze the examinees' ability to convey the main idea of the sentence in their translation. Lastly, a person's level of ease in speaking is evaluated according to their tempo, rate, and spread of speech. Each of these three factors contributes to the ultimate translation quality score, which is calculated by adding up their respective weights. The testing results demonstrate that, in comparison to other techniques, the suggested approach yields outcomes that are in excellent accord with those obtained through human scoring.

In [44], researchers investigate human-machine joint grading of subjective assignments using a three-way decision-making process (TA-S3WD). To begin, the authors suggested a TA-S3WD model for human-machine task distribution to set up a formal structure for human-machine joint grading of subjective assignments. Second, they developed a human-machine joint subjective assignment scoring model by combining bidirectional long short-term memory (Bi-LSTM) with the TA-S3WD model. In [45], the authors suggest a novel architecture built on top of recurrent neural networks (RNNs) and convolutional neural networks (CNNs). From the word embedding vectors and the important semantic ideas, the multilayer convolutional layer in their architecture learns and captures the contextual aspects of the word n-gram to create the feature vector at the essay level. To retrieve past and future contextual models, a bi-gated recurrent unit RNN (BGRU) version was used.

In [46], the authors provide a technique for automating the prediction of subjective scores that is both more accurate and more streamlined than previous approaches. Twenty-one language characteristics are evaluated from the assignments, including syntactical, grammatical, emotional, and readability, to account for all the main considerations of a human evaluator. The Automated Student Assessment Prize (ASAP) competition dataset, hosted on Kaggle, is used in this investigation. Quadratic weighted kappa (QWK) is used as the assessment tool; it calculates the degree to which an algorithm's prediction matches a human-graded score. Using the discovered language characteristics, we explore four different machine learning methods. Among the chosen ML methods, a three-layer neural network with feature selection had the highest QWK. Table 2 summarizes the major categories that recent AES systems follow from the perspective of their strengths and limitations.

Though research on automatic essay grading goes back several decades, it still has considerable room for improvement in terms of efficiency and applicability. The analysis indicates that previous research mainly focused on (1) evaluating essays based on their level of linguistic and vocabulary difficulty, (2) ignoring essay consistency (coherence), and (3) failing to give students feedback on their errors. According to our Google Scholar research, little attention has been paid to suggesting novel methods of assessing the grammatical and semantic quality of student essays, as well as checking for internal consistency and providing students with constructive criticism.

Apart from feature extraction and training machine learning models, no system is accessing the essay's completeness. No system provides feedback to the student response and is not retrieving coherence vectors from the essay—another perspective on the constructive, irrelevant, and adversarial student responses still questioning AES systems. All research concentrated on extracting the features using some NLP libraries, training their models, and testing the results. Yet there is no explanation in the essay evaluation system about consistency and completeness. Another limitation is that there is no domain knowledge-based evaluation of essays using Machine Learning models. In the following section, we will dive into the specifics of the proposed AEE model, which is based on a neutrosophic ontology and aims to improve semantically- based evaluation.

Table 2. Summary of an automated essay scoring systems approaches.

Scoring Method	Approach	Strengths	Limitations
String based	Use statistics to measure the similarity between texts to assign a score to the student essays.	1. It is very simple. 2. Ability to incorporate term weights; any kind of term weight can be added. 3. Partial matching is allowed.	1. Assumed independence relationship among the terms. 2. Checks lexical similarity and the not semantics
Corpus based	Use a corpus to get probability or frequency of a word in a corpus.	Preprocessed corpus to reduce computations.	1. Corpus is domain dependent. 2. Some words might get same similarity. 3. Semantic vectors are sparse.
Knowledge based	Use dictionary information such as WordNet to get similarity (for example, path and depth, word relationships, etc.).	Adoptions of human crafted knowledge graphs can increase accuracy.	1. Limited words. 2. Some words can get same similarity if they have the same path and depth.
Ontology-based	Generate the domain concept ontology in essays	Ontology offers domain-specific semantic correspondences where open domain answers would not be assessed effectively	There are several uncertainties and ambiguities that cannot be accounted for by standard ontologies Does not utilize ontologies with a higher level of fuzzification in order to enhance the effectiveness of identifying semantic mistakes

4. The Proposed Neutrosophic Ontology Based AEE Model

The following input and output criteria have been defined in order to model the automated essay evaluation problem as a supervised machine-learning problem: An example of input is a student's answer to a question. Which label is ultimately given (the classification task) depends on how closely the student's response matches the corresponding model answer. It is possible to give a mathematical formulation of the AEE issue as follows: Many machine learning (ML) categorization methods work by giving each data instance i a value that represents how well it fits into category k [47].

$$Score\ (x_i,\ k) = \beta_K.x_i \tag{3}$$

x_i is the feature vector for data instance i and β_K is the weight vector for category k. $Score(x_i,\ k)$ is the value calculated after classifying instance i into category k. During this stage, a classifier is trained using features drawn from the training set's many data instances. By doing so, we can determine the ideal weights β_K for each category k that aid in the prediction of category k for each data instance in the test phase [1,2,4,6]. In this type of feature analysis, the essay answer is analyzed in terms of its structure where the coherence, writing style, and spelling mistakes are being considered.

The unsupervised learning paradigm relates to the challenge of categorizing data without using a predetermined example or training set, but rather using a distance or similarity function. In the unsupervised AES, each answer is compared with a reference answer or other student answers to produce a similarity score, which is then utilized as the answer's final score. This kind of pairwise similarity may be employed alone or in conjunction with a clustering approach that seeks to arrange similar responses into numerous groups (grade A, grade B, etc.). In the supervised learning paradigm, on the other hand, a prior or example dataset is created to train the classification algorithm. In this respect, previous students' answers, as well as reference answers, are presented with their real instructor score, with the goal of training the machine learning algorithm to predict the score of incoming, testing, or unknown answers. The purpose of this learning process is to predict a numeric value rather than a predefined class label (i.e., machine learning classification).

In this section, a novel model is presented that, rather than focusing on grammar alone, considers the essay's totality (its vocabulary, semantic, and content characteristics).

Semantic assessment is the main emphasis, and students receive meaningful feedback as a result. Semantic characteristics are incorporated to ensure that the essay as a whole makes sense based on the connections made between ideas, rather than merely between adjacent words. The model additionally checks for coherence in the essay by identifying entities, relationships, and cross-concepts. It also makes an effort to employ commonsense knowledge ontologies (in our case, the neutrosophic ontology), allowing them to function across a variety of domains. Figure 3 is the high-level schematic representation of the proposed model. The sections that follow will go into depth on each stage of the suggested AEE paradigm.

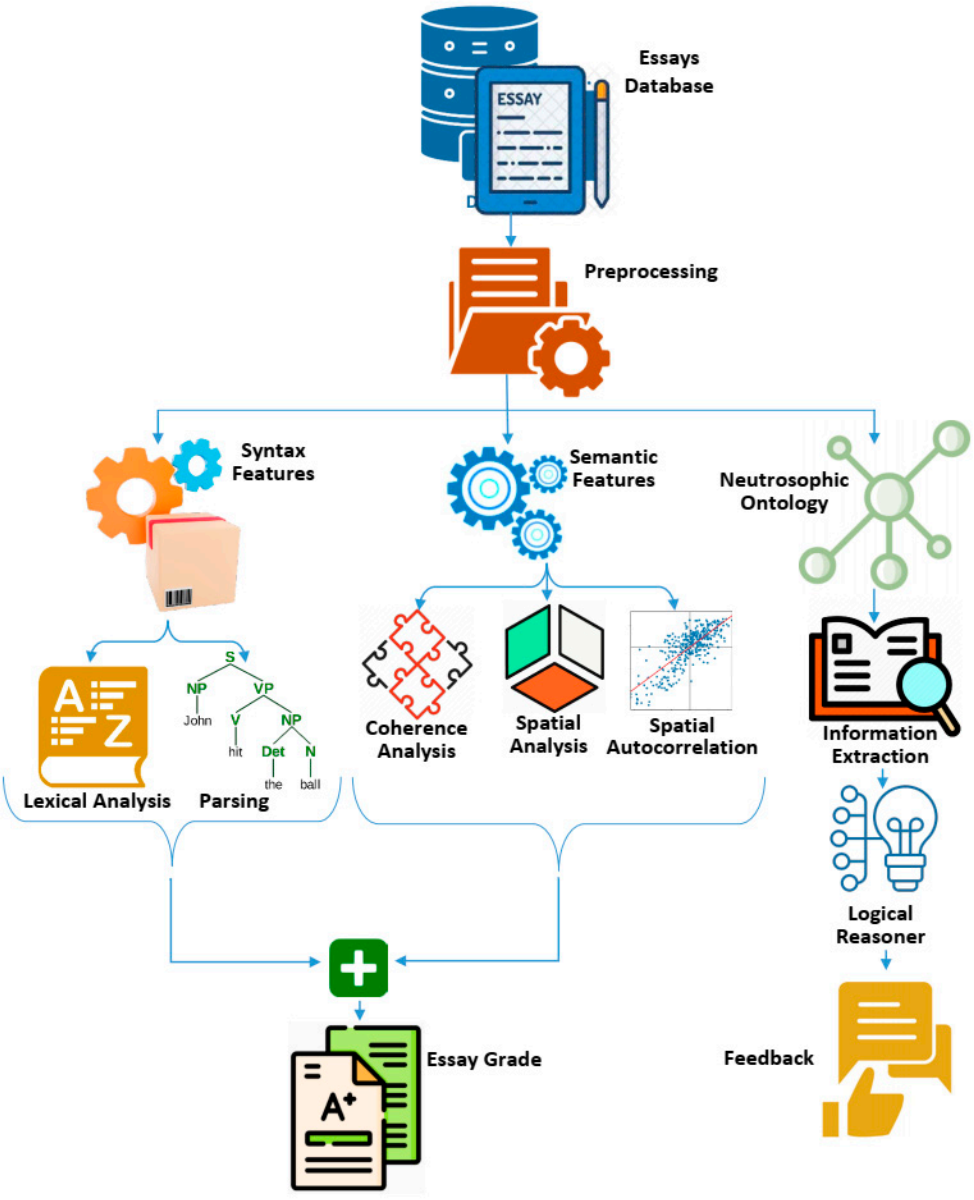

Figure 3. High-level schematic representation of the proposed model.

4.1. Step 1: Preprocessing Phase

The model starts by reading an article and segmenting it into sentences during the preprocessing stage. Tokenization, part-of-speech tagging, detecting and labelling stop words and determiners, converting to lower case, and stemming are some of the preprocessing procedures it performs after making a copy of each sentence. This is necessary since the model cannot handle the essay as a whole [47].

4.2. Step 2: Calculate Syntax Grade

The grammar, syntax, and lexical extent of the essay are all evaluated in two stages to arrive at the syntactic mark [5–7].

4.2.1. Lexical Analysis

According to [48], the best and most popular technique to quantify lexical richness is via a type-to-token ratio. This ratio is calculated by dividing the number of unique lexical items (tokens) by the total number of lexical items in the essay. To figure out how many lexical words are in an essay, each lexical item has been counted once. This means that the singular and plural forms of a single lexeme are counted as two items. Different versions of the same word were only counted once when figuring out how many different words there were. Errors were excluded from the calculation of the second lexical difference measure. The following factors contribute to the total lexical richness:

$$\text{Lexical Variation (With Error)}: \quad LV_1 = \sum Lds / \sum Ls \tag{4}$$

$$\text{Lexical Variation (Without Error)}: \quad LV_2 = \sum (Lds - \sum LEs) / \sum Ls \tag{5}$$

$$\text{Percentage of Lexical Error}: \quad LE = \sum LEs / \sum Ls \tag{6}$$

$$\text{Lexical Density}: \quad Ld = \sum L / \sum Tw \tag{7}$$

where Lds is the total of the items in each segment, Ls is the sum of lexical items (without stop words) per segment, LEs is the sum of lexical errors per segment, Tw is the total number of words in the essay (including stop words), and L is the total number of items.

4.2.2. Parsing

A grammatical review is performed here. It breaks down an article word by word, looking for patterns in the language. In the end, the decoding process yields a parse tree that looks like Figure 4, with the sentence at its base and intermediary nodes (noun phrases, verb phrases, etc.) having offspring and thus being called non-terminals [5,6]. A parsing tree application within the NLTK (Natural Language Toolkit) from Stanford was utilized to parse the sentence. If the node is broken at the higher nodes, the student will lose marks from 0.5 to 1. If the tree is broken at lower nodes, the student will lose marks from 0.1 to 0.4. The mark will be estimated according to the depth of the node. The result of the lexical analysis and parsing tree operations is a vector that contains five elements that are the average of the corresponding elements of each part of the essay.

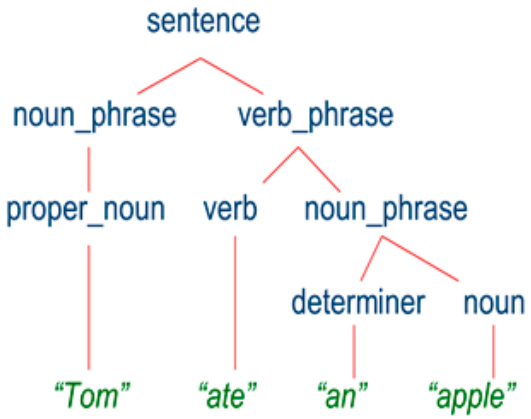

Figure 4. Parse tree.

4.3. Step 3: Calculating Semantic Grade

Semantics are important in the evaluation of essays, as students may write sentences that are irrelevant to the subject of the essay. Semantics is the study of the meaning of words and sentences. Using sematic grading, the essay scoring system provides a systematic checking of essay based on the similarity of its meaning to the model as implied on the content. This system assesses answers to subjective questions by finding a matching ratio for the keywords in instructor and student answers. Coherence characteristics, which verify the essay's uniformity, are based on the assumption that the essay's lexical content evolves naturally as the reader progresses through the text [6,7]. In order to break down essays into their various components, the algorithm first moves a frame (window) over the text in increments of 40 words. The standard number of words in an essay is used to determine how big a "window" should be.

4.3.1. Semantic Space

The essay must be translated into a language understood by computers. One way to quantify a word's significance in a corpus is via its term frequency-inverse document frequency (*TF-IDF*) representation [8]. The regularity with which a word (term) t appears in a document d is the basis for the term frequency $TF(t, d)$. When looking for how uncommon a word is throughout a collection Ds, one can use the inverted document frequency $IDF(t, Ds)$ [8].

$$TF - IDF(t, d, Ds) = TF(t, d) \cdot IDF(t, Ds) \tag{8}$$

$$TF - IDF(t, d, Ds) = \frac{|\{t \in d\}|}{|\{w \in d\}|} \cdot log \frac{|\{d \in Ds\}|}{|\{d \in Ds : t \in d\}|} \tag{9}$$

where w represents the relative weight of the words in each segment of the essay in relation to the word frequency of the essay as a whole. Simply put, the frequency of a term in a text determines its weight. Coherent essays are presumed to have a high degree of proximity between points in high-dimensional semantic space that are characterized by the $TF - IDF$ vectors of essay parts [9,10,49,50].

4.3.2. Coherence Analysis

Coherence is about making everything flow smoothly. The reader can see that everything is logically arranged and connected, and relevance to the central focus of the essay is maintained throughout. Coherence characteristics, which verify the essay's uniformity, are based on the assumption that the essay's lexical content evolves naturally as the reader progresses through the text. Cohesion is one of the key aspects of coherence that relates to the linking of ideas within a sentence, the linking of sentences (the ties between sentences) within a paragraph, and the linking between paragraphs. Coherence measures use a semantic space representation of an essay's parts to calculate the distance between those parts. Four features, split evenly between those measured by cosine similarity and those measured by Euclidean distance, are used to evaluate the suggested model's level of cohesiveness [12].

1. Average distance between neighboring points: Cosine similarity measures the degree to which words within an essay are conceptually close to one another, with higher values indicating closer proximity in meaning.

$$cos\theta = A \cdot B / \|A\| . \|B\| \tag{10}$$

   A and B are points (tokens) in the semantic space.

2. Two points' average distance: It is essential to evaluate the persistence of an idea throughout the essay.

$$\sqrt{\sum_{1=1}^{n}(A_i - B_i)^2} \tag{11}$$

n is features number in the vector.

3. Two points' maximum difference: This metric is employed to assess the spatial scope of the studied idea by measuring the radius of the region encompassed by points.

$$M = \sqrt{\sum_{i=1}^{n} C^2}, \quad C = \|A - B\| \tag{12}$$

C is the two points' difference, M is the two points' maximum difference.

4. Clark and Evans' closest neighbor distances at each point: This helps measure spatial connections. It quantifies the degree to which an observed distribution deviates from its nearest peer distribution.

$$CE = \frac{2\sqrt{N}\sum_{i=1}^{N} ds_i}{N} \tag{13}$$

where ds is the distance between a given point and its closest neighbor, and N is the number of points.

5. Distribution of distances to neighbors based on their frequency: It checks the proportion of content variations from the main idea.

$$C(ds) = \frac{|\ ds\ \leq\ \overline{ds}\ |}{N} \tag{14}$$

$\overline{ds}$ is the average neighbor distance. Herein each essay's part will produce a vector that contains five elements. The final product of the coherence analysis will be a vector of five elements that represent the average of the corresponding elements of each part [49,50].

4.4. Step 4: Spatial Data Analysis

In order to extract implicit knowledge, such as statistical qualities, this collection of attributes specifies the spatial aspects of the data. Descriptive spatial statistics are used to classify the essays. The employed attributes quantify the spatial central tendency and spatial dispersion. These attributes are defined based on the points in the semantic space [49,51,52]:

4.4.1. Standard Distance (Equal to the Standard Deviation in Space)

It is employed in the determination of the total range of variation in a given point distribution.

$$S_D = \sqrt{\frac{\sum_{k=1}^{n}\sum_{i=1}^{N}\left(D_i^k - \overline{D_c^k}\right)^2}{N}} \tag{15}$$

Point i's kth coordinate component is denoted by $D_i^k,\ k = 1, \ldots, n,\ i = 1, \ldots, N$; the mean center's kth coordinate component is denoted by $\overline{D_c^k}$; n is the number of dimensions (the sum of the derived words, verbs, adjectives, and adverbs); and N is the number of points. Both the standard distance and the standard deviation are extremely sensitive to extreme numbers. Because squared lengths to the center are used to calculate this measure, the usual points have a disproportionate impact on its size, allowing for the identification of deviating (incoherent) parts of an essay.

4.4.2. Relative Distance

It is a way to quantify the spatial dispersion that exists between two points.

$$R_D = \frac{S_D}{d_{max}} \tag{16}$$

where S_D the standard is distance and d_{max} is a maximum distance of any point from the average of the points. This allows for the direct analysis of point-pattern dispersion across regions of various sizes. Herein, each essay's part will produce a vector that contains two elements. The final product of the spatial analysis will be a vector of two elements that represent the average of the corresponding elements of each part.

4.4.3. Spatial Autocorrelation

The spatial autocorrelation of essays is used to assess the overall and local semantic consistency of the essay content, thereby determining the structural quality of the essay. The spatial autocorrelation tool measures the correlational similarity between adjacent sentences based on both feature locations and feature values simultaneously. Given a set of features and an associated attribute, it evaluates whether the pattern expressed is clustered, dispersed, or random. One can use spatial autocorrelation measures to characterize whether data tends to be discrete (negative spatial autocorrelation) or clustered together in space (positive spatial autocorrelation). They help us determine the degree to which the essay material is semantically coherent on a global and local scale. Positive spatial correlations in an essay suggest that its sections are coherent and that information is presented in a logical progression, as discussed in the author's previous work [53]. Moran's (I), Geary's (C), and Getis's (G) are all unique spatial autocorrelation metrics. These three metrics, computed with the Python spatial analysis library, have been modified for use in high-dimensional semantic space [54]:

1. Moran's (I): To assess the clustering pattern as a whole, this is crucial information. Originally developed for a 2-dimensional space, the work here averages the distinguishing measure over dimensions so that it may be used to a high-dimensional semantic space:

$$I = \frac{N}{S} \cdot \frac{1}{n} \sum_{k=1}^{n} \left[\frac{\sum_{i=1}^{N} \sum_{j=1}^{N} w_{ij} \left(D_i^k - \overline{D_c^k} \right) \left(D_j^k - \overline{D_c^k} \right)}{\sum_{i=1}^{cc} \left(D_i^k - \overline{D_c^k} \right)^2} \right] \tag{17}$$

S is the total of all weights w_{ij}. Every pair of points i and j is assigned a weight w_{ij}, where $w_{ij} = 1$ if i and j are neighbors and $w_{ij} = 0$ otherwise. There is a 1 to +1 spread for I. If I is positive, then there is spatial autocorrelation, which indicates that nearby points tend to cluster together. Absolute spatial randomness is represented by values close to zero.

2. Geary's (C): In this case, the interaction is not the cross-product of standard deviations from the mean, but rather the differences in intensity between different reflection points. Once more, the adjusted measure is calculated in a high-dimensional semantic space and is averaged over all dimensions (the average of extracted nouns, adjectives, verbs, and adverbs):

$$C = \frac{(N-1)}{2} \cdot \frac{1}{n} \sum_{k=1}^{n} \left[\frac{\sum_{i=1}^{N} \sum_{j=1}^{N} w_{ij} \left(D_i^k - D_j^k \right)^2}{\sum_{i=1}^{N} \sum_{j=1}^{N} w_{ij} \left(D_i^k - \overline{D_c^k} \right)^2} \right] \tag{18}$$

3. Getis–Ord's (G): A more local scale can be used to study point patterns. Getis–Ord's (G) checks overall density or lack of density of all pairs of values ($D_i D_j$) such that i and j are within distance d of each other. It is also adjusted to be use it in a high-dimensional space:

$$G(AD) = \frac{1}{n} \sum_{k=1}^{n} \left[\frac{\sum_{i=1}^{N} \sum_{j=1}^{N} w_{ij}(d) D_i^k D_j^k}{\sum_{i=1}^{cc} \sum_{j=1}^{cc} D_i^k D_j^k} \right] \tag{19}$$

AD is the average distance between two points in the semantic space. A weighting function $w_{ij}(d)$ is used to assign binary weights to every pair of points, where $w_{ij}(d) = 1$, if i and j are within distance AD and $w_{ij}(d) = 0$, otherwise. Herein, each essay's part will produce a vector that contains three elements. The final product of the spatial analysis will be a vector of three elements that represent the average of the corresponding elements of each part.

4.5. Step 5: Calculating Essay Grade

Most AES tools rely on a score prediction made with machine learning [54–58]. The most used algorithm is the linear regression. It is used for its ease of interpretability. More specifically, given an input feature vector $x \in \mathbb{R}^m$, an output is predicted $\hat{y} \in \mathbb{R}^m$ using a linear model with a weight of β:

$$\hat{y} = \beta_0 + \S^{\mathsf{T}}\beta \tag{20}$$

To learn values for the parameters $\theta = \langle \beta_0, \beta \rangle$we minimize the sum of squared errors for a training set containing a pairs of essays and scores $\langle x_i, y_i \rangle$ where $x_i \in \mathbb{R}^m$and $y_i \in \mathbb{R}^m$:

$$\hat{\theta} = \underset{\theta=(\beta_0,\beta)}{\operatorname{argmin}} \left(\frac{1}{2m} \sum_{i=1}^{m} (y_i - (\beta_0 + \S_i^{\mathsf{T}}\beta))^2 + \alpha P(\beta) \right) \tag{21}$$

m is the number of essays, α is a threshold parameter, $P(\beta)$ is the penalty term for the weights that is calculated as.

$$P(\beta) = \sum_{j=1}^{m} |\beta_j| \tag{22}$$

For each essay set, a development set was tuned using a 5-fold cross-validation. The linear regression function from a toolbox in SPSS application was used to calculate the essay grade.

4.6. Step 6: Building Neutrosophic Ontology

In order to provide students with feedback, neutrosophic ontology is seen as a necessary prerequisite. The approach first constructs a neutrosophic ontology consisting of commonsense knowledge [19,20]. Neutrosophic ontology is used because it removes ambiguity from the language. A lattice of fuzzy ontologies (consisting of classes, relations, functions, and instances) forms the basis of this framework. The WordNet taxonomy is used to add synonyms (same meaning with different words) and hyponyms (same word with different meanings) to the neutrosophic ontology. The model proceeds by supplementing the neutrosophic ontology with:

4.6.1. Source Text Knowledge

The model builds the base neutrosophic ontology based on source text knowledge using Protégé application. This knowledge is extracted from a set of essays in a shape of concepts and their relations.

4.6.2. Domain Knowledge

The model supplements its understanding of the source text with domain knowledge—specifically, a fuzzy ontology formulation consisting of synonyms and hyponyms—that covers the scope of an essay as a whole. For essays written by students on topics related to genes and biology, for instance, the system automatically incorporates a gene neutrosophic ontology in addition to the standard one.

4.7. Step 7: Information Extraction

The model now extracts the information that will be added to the neutrosophic ontology to give students feedback about their writing mistakes. To do this, these steps are needed [59–61]:

4.7.1. Entity Recognition

The process of identifying grammatical expressions in words written in a native language is called shallow parsing. Chunks (or phrases) such as noun phrases, verb phrases, prepositional phrases, adverbial phrases, and clauses introduced by subordinating conjunctions are all easily identifiable by a parse trees. The use of these pieces is an initial step toward full natural language understanding. True parse tree recognition is a much more difficult issue, but it can yield richer insights into the words being analyzed. The goal was to spot grammatical mistakes before moving on to the Stanford NLP toolkit.

4.7.2. Co-Reference Resolution

Finding when multiple references to the same object exist is essential for natural language processing and providing smart access to written information. Unannotated essay text is processed by the co-reference resolve procedure. Two co-reference resolution models (the Illinois co-reference resolution and the Stanford parser) are used by the model to find and use co-references in an essay, which are then aggregated as extractions into the neutrosophic ontology. Co-references found by both models are combined in the model, which improves the model's precision.

4.7.3. Open Information Extraction

Following these procedures, the model performs information extraction and gives the triples $\{\langle arg_1, rel, arg_2\rangle\}$ where $args$ are noun phrases and rel is a textual piece showing an implied, meaning connection between the two noun phrases. Extractions that are made twice, or that are performed incorrectly, are ignored during the procedure. (e.g., those consisting only of a subject and a relation, while an object is missing). Following the aforementioned procedures, the model begins processing each phrase in turn, eventually adding each extraction to the neutrosophic ontology by means of the logic reasoner. See [61] for more details.

4.8. Step 8: Logical Reasoner

The neutrosophic ontology's consistency is examined by a logical reasoner [62]. When the model has the neutrosophic ontology and the essay's information extractions, it can begin identifying semantic errors. Extracted data is being added to the neutrosophic ontology by the logical reasoner. The model decides that there is a semantic mistake in the essay if the neutrosophic ontology is contradictory or has an unknown concept after adding an extraction. The algorithm remembers the details of the error and uses that information to improve its future feedback. After dropping that relation from neutrosophic ontology, it proceeds to the next extraction.

The model first looks for both items (predicates) in the neutrosophic taxonomy every time an extraction is organized. If a term is not included in the neutrosophic ontology, but it has an equivalent or a co-reference in WordNet [63], the latter can be used to fill in the gap. The method looks for hyponyms of the entity and creates a subclass if the neutrosophic ontology does not already include any equivalents or co-references. (i.e., creating a triplet with a subclass of relation). The only other option is to create a new class or individual in the neutrosophic ontology if all the previously outlined efforts fail. If both entities are already part of the neutrosophic ontology, then the model will determine if the precise connection is missing and add it.

Subsequently, the learner receives feedback. Feedback is the most important aspect of making an automated essay scoring system useful in real life. Feedback gives users an interpretation of the score so that they are convinced about the validity of the score; feedback can also provide an answer on what to do to improve the score. In the current state of research, a system that does not provide feedback can no longer be put to practical use, and those who have used a feedback system no longer want to use a system without feedback.

The Protégé application's Reasoner utility was used for this stage. Protégé 4.3 with RacerPro, Integrated Control Technology Limited, Auckland, New Zealand (https://franz.com/agraph/racer/racer_features.lhtml, accessed on 1 January 2023) was selected as the reasoner that provides inference services for terminological knowledge as well as for representations of knowledge about individuals. RacerPro exploits and combines multiple clues coming from low-level vision processes in a semantically meaningful way. The user interface supports three types of reasoning: (1) consistency checking; (2) classification (incorporating something under a more general category); and (3) instance classification.

5. Experimental Results

Experiments were run on benchmark datasets given as part of the Automated Essay Scoring challenge hosted on the Kaggle platform, Kaggle Inc., Google LLC, San Francisco, United States (www.kaggle.com/datasets, accessed on 1 January 2023) [64]. Student essays in response to six distinct topics can be found in the databases (essay discussion questions). The American students who participated were in grades 7, 8, and 10 (aged 12, 13, and 15, respectively). Persuasive writing (T1) and source-based essays (T2) were used. A persuasive essay is one in which the writer attempts to get the reader to agree with your point of view. Source based essays are ones in which your own opinion should not be the predominant aspect of your answer. At least two human expert evaluators pre-score each training session. The databases were pre-organized into training and testing groups by their creators. Scoring models were constructed using the same training and test sets, and predicted precision was determined by comparing the two. The characteristics of the used datasets are shown in Table 3. The suggested model was build using three software:

Table 3. Dataset Description.

Characteristic	DS_1	DS_2	DS_3	DS_4	DS_5	DS_6	DS_7	DS_8
Type of essay	T1	T1	T2	T2	T2	T2	T2	T2
No. of essays	1783	1800	1726	1771	1805	1800	600	568
Mean number of words	366.40	381.19	108.69	94.39	122.29	153.64	387.4	113.24
Standard deviation of number of words	120.40	156.44	53.3	51.68	57.37	55.92	156.82	65.0
Range of grades	2–12	1–6	0–3	0–3	0–4	0–4	1–4	0–3
Mean grade	8.35	3.42	1.85	1.43	2.41	2.72	3.42	1.9

- Stanford NLTK Kite, Stanford University, Stanford, CA, USA (www.nlp.stanford.edu/software, accessed on 1 January 2023) [65]: Linguistic attributes were extracted from the text by using the Natural Language Toolkit (NLTK) for natural language processing from Stanford University. It consists of several modules, but only three were used: the Stanford parser, used to separate phrase structure; the Stanford Named Entity Recognition (NER), used to label sequences of words in a text that are the names of things, such as person and company names or gene and protein names; and The Stanford co-reference resolution module, used to find mentions of the same entity in a text, such as when "Theresa May" and "she" refer to the same person.
- Protégé Stanford University, Stanford, CA, USA (https://protege.stanford.edu/, accessed on 1 January 2023) [66]: Ontology was built using the Protégé application for both Fuzzy and neutrosophic logics. It is a free, open-source ontology editor and a knowledge management system. Protégé provides a graphic user interface to define ontologies. It also includes deductive classifiers to validate that models are consistent and to infer new information based on the analysis of an ontology.
- Matlab R2022b [67]. The suggested model has been implemented in Matlab. It is used to calculate all the mathematical values which lead to the essay grade. In general, Matlab has several advantages: (1) it allows the testing algorithms immediately without recompilation. It permits typing something at the command line or executing a section in the editor and immediately shows the results, and greatly facilitating algorithm development; (2) MATLAB's built-in graphing tools and Graphical User

Interface (GUI) builder ensures that customizing the data and models helps the user interpret their data more easily for quicker decision making; (3) MATLAB's functionality can be expanded by adding toolboxes. These are sets of specific functions that could provide more specialized functionality.

The suggested model has been built using a laptop computer (HP pavilion g6 Series) with the following specifications:

- Processor: Intel (R), Core (TM) i5 CPU, 5200U @ 2.20 GHz (4CPUs) 2.20 GHz.
- RAM: 16 GB.
- System type: 64-bit operating system, and Microsoft Windows 10 professional as running operating system.

The following metrics were used to assess the accuracy of the prediction models [68,69]:

- Exact agreement measure (Z). The Z-score is the percentage of essays that were given the same score by both a human evaluator and the AEE model. In this case, we count the number of scores assigned by the AES system that agree with the human raters relative to the total number of scores (in percent).

- The Kappa statistic, also known as the quadratic weighted Kappa (*Kp*), is a measure of inter-rater reliability (how closely two raters' opinions coincide). The quadratic weighted kappa will find agreement between human evaluation score and system evaluation score and produces value ranging from 0 to 1. This scale usually falls somewhere between 0 (completely arbitrary agreement between evaluators) and 1 (complete agreement between graders). If there are *S* possible ratings (1, 2,... *S*), and two evaluators have scored each article, then we can say that essay responses has *S* possible ratings. (e.g., human and computer). We can determine this measure by performing the following:

$$kp = 1 - \frac{\sum_{i,j} W_{i,j}\, O_{i,j}}{\sum_{i,j} W_{i,j}\, e_{i,j}} \tag{23}$$

where *W* represents weights, *O* represents the actual rating matrix, and *e* represents the expected rating matrix. The weights assigned by each evaluator are stored in a S- by- S matrix called $W_{i,j}$, where *i* is the *i*-th weight and *j* is the *j*-th weight.

$$W_{i,j} = \frac{(i-j)^2}{(S-1)^2} \tag{24}$$

Matrix of evaluations that have been collected by constructing a S -by- S histogram (agreement) matrix *O* over the essay ratings, where $O_{i,j}$ represents the number of essays assigned a grade of *i* by grader A and a grade of *j* by grader B, and where *RE* represents an S-by-S histogram of predicted grades:

$$RE_{sc,j} = \frac{H_{Ai}\, .H_{Bj}}{No} \tag{25}$$

Number of essays graded by grader A with score *sc* is denoted by H_{Ai} , i = 1, ..., *S*, *No* = total number of essays graded. The sum of *RE* and *O* is equal after normalization with *No*.

To validate the suggested model, a set of experiments was conducted to analyze the potential benefits of the spatial measures and their significance to predictive accuracy. Then, experiments continued to compare the predictive accuracy of the suggested model by comparing its results to those of other state-of-the-art systems. The last experiment was to check the error rate of the suggested detection system. All experiments were conducted after transforming the essay into a semantic space by moving a window over it by 10 words. To determine if the inclusion of semantic attributes improves model performance, experiment 1 (semantic attribute significance) compares the proposed model to two related statistical-based AES systems: AGE (Automated Grader for Essays with only linguistic and content attributes) and AGE+ (the AGE system with additional coherence attributes) [51,70,71].

The proposed system for assessing writing was more effective. Quadratic weighted kappa and exact agreement are displayed in Tables 4 and 5, respectively, for AEG, AGE+, and the suggested model. The results reveal that the model's accuracy significantly increases by approximately 6% to 7% over AGE and AGE+. In the exact agreement measure, the proposed model increases by 6% in the case of AGE and by 5% in the case of AGE+ for all datasets. The proposed model's superiority is predicated on the fact that the AGE system simply validates syntactic characteristics. Only syntax and coherence are validated by the AGE+ system. The suggested approach, however, takes into account all three characteristics (syntax, coherence, and consistency) simultaneously, leading to better results. Syntax means grammar and spelling mistakes, coherence means that ideas in a paragraph flow smoothly from one sentence to the next, and consistency means that the paragraphs of the essay are linked well and evolve around the main idea of the essay.

Table 4. Comparison of AEG, AGE+ and the Suggested Model for different data set in terms of *Kp*.

Models	DS_1	DS_2	DS_3	DS_4	DS_5	DS_6	DS_7	DS_8
AGE System	0.91	0.75	0.82	0.84	0.89	0.73	0.88	0.79
AGE+ system	0.93	0.80	0.84	0.85	0.88	0.77	0.89	0.80
Neutrosophic ontology model	0.95	0.83	0.88	0.86	0.91	0.84	0.91	0.82

Table 5. Comparison of AEG, AGE+ and the Suggested Model for different data set in terms of *Z*.

Models	DS_1	DS_2	DS_3	DS_4	DS_5	DS_6	DS_7	DS_8
AGE System	0.88	0.72	0.78	0.75	0.77	0.74	0.86	0.77
AGE+ system	0.90	0.79	0.82	0.81	0.79	0.76	0.87	0.79
Neutrosophic ontology model	0.94	0.82	0.85	0.83	0.90	0.78	0.90	0.81

In experiment two (a comparative study with commercial systems), the aim is to compare the suggested model to the state-of-the-art systems that were mentioned in [71]. Project Essay Grade (PEG) is one such system; it employs a database of training essays graded by humans to create a model that evaluates the quality of unscored essays using sophisticated statistical methods. E-Rater assigns grades to essays after identifying and extracting a collection of features that stand for key components of writing quality. Next, a statistical model is used to aggregate these features and assign weights to them based on a statistical procedure, ultimately yielding an estimated score. The Lexile Writing Analyzer makes use of a web-based application called the Lexile Analyzer to calculate the Lexile level of polished, edited, and complete writing. There is now an automated scoring market in the United States, and these commercial systems have cornered more than 97% of it [71,72]. There is no feedback given by these systems; they simply provide grades.

The results shown in Table 6 confirm that the proposed model is better than the systems compared with it in terms of quadratic weighted kappa and average accuracy. As not all the systems are offered for public testing, their results were acquired from the work presented in [73] and from the Kaggle website. The evaluated systems are organized in descending order of the average Kappa value, and commercial systems are ranked according to average accuracy. The proposed model was better than PEG by 3.86%, E-rater by 5.55%, and Lexical by 2%. These systems use statistical methods to extract the lemmas (tokens after stemming) into a vector, and then they begin dealing with these lemmas as numbers only, which means they do not care about the word order or the coherence and consistency of the essay. Meanwhile, in the suggested approach, extra features, such as spatial analysis and spatial autocorrelation, are used to assess the essay, improving accuracy.

Table 6. Comparison of the proposed model with other state-of-the-art commercial systems in terms of quadratic weighted Kappa, attained on different datasets.

AEE Systems	DS_1	DS_2	DS_3	DS_4	DS_5	DS_6	DS_7	DS_8	Average	Rank
PEG	0.84	0.72	0.84	0.86	0.83	0.82	0.85	0.75	0.81	2
e-rater	0.83	0.70	0.83	0.83	0.80	0.79	0.81	0.71	0.78	3
Lexile	0.69	0.63	0.68	0.61	0.65	0.66	0.59	0.64	0.64	4
Proposed Model	0.94	0.82	0.89	0.87	0.90	0.83	0.89	0.78	0.83	1

In experiment 4 (Comparative study with machine learning models), the suggested model is compared to several machine learning models in the field of essay scoring. These models are the recurrent neural network [74], which predicts a score for a given essay, defined as a sequence of words, by following multi-layered neural networks; similarity measures [75,76], in which student answers are compared with model answers and then scores are allocated; and word order graph [77], which uses word-order based graph representation for text. The word order graph is able to capture the order of the tokens and their lexicon-semantic. Then, it utilizes the matching technique for identifying the degree of relatedness across tokens and phrases. The suggested model was superior for evaluating essays. Quadratic weighted kappa was used to compare the results in the three experiments. Table 7 indicates that the proposed model results were significantly better than the neural network-based model (CO-Attention) by 10%. Table 8 shows that the proposed model exceeded the model that utilizes the Jaccard coefficient by 11%, the model that utilizes the dice coefficient by 15%, and the model that utilizes the cosine coefficient by 17%. The results in Table 9 show that the proposed model is better than the algorithm that uses word graphs by 5.5%.

Table 7. Comparison between the proposed model and neural network-based AEE model on selected dataset in terms of *Kp*.

AEE Systems	DS_3	DS_4	DS_5	DS_6
Neural Network based model (CO-ATTN)	0.70	0.81	0.80	0.82
Proposed Model	0.87	0.86	0.91	0.84

Table 8. Comparison between the proposed Model Similarity Measures-based AEE models in terms of *Kp*.

	Jaccard Coefficient	Dice Coefficient	Cosine Coefficient	Proposed Model
DS_1	0.71	0.77	0.79	0.93
DS_2	0.32	0.35	0.43	0.83
DS_3	0.25	0.28	0.43	0.87
DS_4	0.43	0.51	0.63	0.86

Table 9. Comparison between Word Graph Model (AutoSAS) and the Proposed Model in terms of *Kp*.

	AutoSAS	Proposed Model
DS_1	0.83	0.92
DS_2	0.82	0.82
DS_3	0.85	0.87
DS_4	0.84	0.86

The superiority of the proposed model relies on the fact that recurrent neural network is not considered the best kind of neural network for all natural language processing tasks, which perform poorly on sentences with lots of words [75]. The reason is, if the sequence is a sentence of five words, the network will contain one layer for each word, which would require heavy processing and take a lot of time. Moreover, the proposed

model utilizes more than one semantic attribute to check similarity. In addition to that, the proposed model exceeded the models that use similarity measures because it utilized LSA to overcome the drawbacks of similarity measures. Similarity measures cannot deal with synonyms (words with the same meaning) or polysemy (the same word with different meaning). Using LSA transformed the essay into semantic space, so every word was treated as a point in the space, and as a result, it was judged semantically well. The superiority of the proposed model over the algorithm that uses lexicon-semantic matching comes from transferring the essay into a semantic space and utilizing spatial measures to check the coherence and consistency of the points in this semantic space (the essay's parts). The AutoSAS system used word2vector and doc2vector techniques to extract features. These two techniques have a great drawback, which is that they do not define the sub-linear relationships explicitly.

The purpose of experiment four (Essay Error Analysis) is to evaluate the suggested model's error detection capabilities. Sixty tourism-related phrases were constructed artificially. Forty-one statements were marked as correct and nineteen as erroneous to reflect the base truth. It was designed to test how well the automatic error detection model can identify incorrect statements using only a common sense neutrosophic ontology as input. Both the sensitivity (the percentage of incorrect sentences accurately identified as such) and the specificity (the percentage of correct sentences correctly identified as such) of the suggested model were evaluated. Putting the suggested model to the test, it was found to be both highly specific (100%) and highly sensitive (80.2%). Since the model assumes that every phrase is accurate unless it finds a mistake in the sentence, a precision of 100% was to be anticipated. By successfully navigating the fuzziness of concepts, the incorporation of neutrosophic logic within precise ontology yielded an 80.2% sensitivity. The accuracy loss is eliminated, and language concepts are conveyed effectively.

An additional set of experiments (comparative study with a fuzzy ontology model) was carried out to prove that neutrosophic logic coupled with an ontology builder is an effective tool for modeling semantic characteristics in the realm of computer-assisted essay grading. In this context, we compared our suggested method to one that employs fuzzy ontology to represent semantic characteristics for automatic essay scoring, as discussed in the author's previous work [53]. The results in Table 10 confirm the research hypothesis that using the neutrosophic ontology for modeling an essay's semantic features will enhance the system's correctness in terms of *kp*. The suggested combination achieved a 4% increase in *kp* when compared to the nearest combination that shields between fuzzy logic and ontology. For each case in the overlapping region, the neutrosophic ontology specifies the level of truth, indeterminacy, and falsehood. Therefore, it yields superior results when compared to fuzzy ontology. Only the instance's fuzzy membership value is provided by a fuzzy ontology. In neutrosophic sets, all three measures (Truth, Falsehood, and indeterminacy) are independent, and how one affects another in decision-making depends on their sum. If sum < 1, we have incomplete information (we do not know all information), while if sum = 1, we have complete information, as in an intuitionistic fuzzy set.

Table 10. Comparison of the proposed model with other state-of-the-art fuzzy ontology-based AEE systems in terms of quadratic weighted Kappa, attained on different datasets.

Models	DS_1	DS_2	DS_3	DS_4	DS_5	DS_6	DS_7	DS_8
Fuzzy Ontology model	0.90	0.78	0.86	0.82	0.87	0.80	0.87	0.79
Neutrosophic Ontology model	0.95	0.83	0.88	0.86	0.91	0.84	0.91	0.82

6. Conclusions

With the rise of online education, students' online writing is becoming more and more popular. However, scoring these writings manually is time-consuming and tiring. That is why an efficient automated scoring system is urgently needed. Automated essay scoring systems are developing into systems with richer functions than the previous simple scoring

systems. Its purpose is not only to score essays but also as a learning tool to improve the writing skills of users. This imposes challenges on the automated evaluation system: (1) finding better features to score the essay semantically; (2) checking the coherence and consistency of the essay; and (3) providing feedback to students in order to help them recognize their mistakes and avoid them in the future. In general, the contributions of this paper are presented as follows: (1) Utilizing latent semantic analysis (LSA) to convert the essay into a compact vector that contains the main features of the essay; (2) Applying both spatial analysis and spatial autocorrelation to handle the essay's semantic coherence and consistency. (3) A neutrosophic ontology was applied to detect the semantic errors in the essay; the output is provided to students in the form of feedback.

The proposed model can be used in schools to evaluate students' essays and paragraphs in English exams. This model can be beneficial in evaluating essays in TOEFEL and ILTS exams. The proposed model was proven through experiments to be reliable and valid. It outperformed the commercial systems that score essays by approximately 4%. The proposed model faces some limitations, which are: (1) it cannot evaluate handwritten essays; (2) there was no way to detect plagiarism; (3) if the student illustrates his writing with poetry, "The waves beside them danced", the model will not be able to identify its semantic.

The open challenges for future work are scattered over different approaches that are used through the development of the proposed AEE model. First and foremost, different semantic attributes will be further developed. Alternative approaches to LSA and TF-IDF for transforming text into semantic space will be tested to discover how the alternatives impact the results. In addition to that, the problem of co-reference will be investigated. Furthermore, the automatic error detection model shall be upgraded by trying to use other external sources to determine if statements in an essay are true and consistent. Incorporating inference rules into the type-2 neutrosophic ontology is one of the future goals since it will aid in the detection of implicit errors and unstated facts and relations in an essay. Moreover, new approaches for unsupervised learning and a taxonomy for taxonomy-based learning shall also be developed, since the current approach uses only WordNet as the underlying taxonomy. The last goal is to have it evaluate the handwritten essays.

Author Contributions: Conceptualization, S.M.D. and R.A.A.; methodology, S.M.D. and A.A.E.; software, R.A.A.; validation, S.M.D. and A.A.E.; formal analysis, S.M.D. and A.A.E.; investigation, S.M.D., R.A.A. and A.A.E.; resources, R.A.A.; data curation, S.M.D. and R.A.A.; writing—original draft preparation, S.M.D. and R.A.A.; writing—review and editing, A.A.E. and R.A.A.; visualization, R.A.A.; supervision, S.M.D.; project administration, S.M.D.; funding acquisition, R.A.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Datasets for this research are available in www.kaggle.com/datasets, accessed on 1 January 2023.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Goura, V.; Moulesh, M.; Madhusudanarao, N.; Gao, X. An efficient and enhancement of recent approaches to build an automated essay scoring system. *Procedia Comput. Sci.* **2022**, *215*, 442–451. [CrossRef]
2. Beseiso, M.; Alzubi, O.; Rashaideh, H. A novel automated essay scoring approach for reliable higher educational assessments. *J. Comput. High. Educ.* **2021**, *33*, 727–746. [CrossRef]
3. Lee, A.; Luco, A.; Tan, S. A Human-centric automated essay scoring and feedback system for the development of ethical reasoning. *Educ. Technol. Soc.* **2023**, *26*, 147–159.
4. Vijaya, S.; Guruvyas, K.; Patil, P.; Acharya, J. Essay scoring systems using AI and feature extraction: A review. In *Proceedings of Third International Conference on Communication, Computing and Electronics Systems*; Springer: Singapore, 2022; pp. 45–57.

5. Rokade, A.; Patil, B.; Rajani, S.; Revandkar, S.; Shedge, R. Automated grading system using natural language processing. In Proceedings of the Second International Conference on Inventive Communication and Computational Technologies, Coimbatore, India, 20–21 April 2018; pp. 1123–1127.
6. Smith, G.; Haworth, R.; Žitnik, S.; Smith, G.G.; Haworth, R.; Žitnik, S. Computer Science Meets Education: Natural Language Processing for Automatic Grading of Open-Ended Questions in eBooks. *J. Educ. Comput. Res.* **2020**, *58*, 1227–1255. [CrossRef]
7. Somers, R.; Cunningham-Nelson, S.; Boles, W. Applying natural language processing to automatically assess student conceptual understanding from textual responses. *Australas. J. Educ. Technol.* **2021**, *37*, 98–115. [CrossRef]
8. Nandini, V.; Maheswari, P.U. Automatic assessment of descriptive answers in online examination system using semantic relational features. *J. Supercomput.* **2020**, *76*, 4430–4448. [CrossRef]
9. Hendre, M.; Mukherjee, P.; Preet, R.; Godse, M. Efficacy of Deep Neural Embeddings based Semantic Similarity in Automatic Essay Evaluation. *Int. J. Comput. Digit. Syst.* **2020**, *10*, 1379–1389. [CrossRef]
10. Hoblos, J. Experimenting with latent semantic analysis and latent Dirichlet allocation on automated essay grading. In Proceedings of the Seventh International Conference on Social Networks Analysis, Management and Security, Paris, France, 14–16 December 2020; pp. 1–7.
11. Bhatt, R.; Patel, M.; Srivastava, G.; Mago, V. A graph based approach to automate essay evaluation. In Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics, Toronto, ON, Canda, 11–14 October 2020; pp. 4379–4385.
12. Rao, M. Automated evaluation of Telugu text essays using latent semantic analysis. *Turk. J. Comput. Math. Educ.* **2021**, *12*, 5299–5302.
13. Jain, S.; Seeja, K.; Jindal, R. Computing semantic relatedness using latent semantic analysis and fuzzy formal concept analysis. *Int. J. Reason. -Based Intell. Syst.* **2021**, *13*, 92–100. [CrossRef]
14. Adel, E.; El-Sappagh, S.; Barakat, S.; Hu, J.; Elmogy, M. An Extended Semantic Interoperability Model for Distributed Electronic Health Record Based on Fuzzy Ontology Semantics. *Electronics* **2021**, *10*, 1733. [CrossRef]
15. Dhiman, S.; Thukral, A.; Bedi, P. OHF: An ontology based framework for healthcare. In *Proceedings of Third International Conference in Artificial Intelligence and Speech Technology*; Springer International Publishing: Cham, Switzerland, 2022; pp. 318–328.
16. Urbieta, I.; Nieto, M.; García, M.; Otaegui, O. Design and Implementation of an Ontology for Semantic Labeling and Testing: Automotive Global Ontology (AGO). *Appl. Sci.* **2021**, *11*, 7782. [CrossRef]
17. Fathian, A. Semantic publishing: A semantic representation of scholarly publications based on the SPAR ontologies. *Librariansh. Inf. Organ. Stud.* **2021**, *32*, 23–55.
18. Indra, M.; Govindan, N.; Satya, R. Thanasingh. Fuzzy rule based ontology reasoning. *J. Ambient. Intell. Humaniz. Comput.* **2021**, *12*, 6029–6035. [CrossRef]
19. Hernández, N.; Chacón, L.; Cruzaty, L.; Zumba, G.; Chávez, W.; Quispe, J. Model based on neutrosophic ontologies for the study of entrepreneurship competence. *Neutrosophic Sets Syst.* **2022**, *51*, 57.
20. Bhutani, K.; Aggarwal, S. Experimenting with neutrosophic ontologies for medical data classification. In Proceedings of the IEEE Workshop on Computational Intelligence: Theories, Applications and Future Directions, Kanpur, India, 14–17 December 2015; pp. 1–6.
21. Abbasi-Moud, Z.; Hosseinabadi, S.; Kelarestaghi, M.; Eshghi, F. CAFOB: Context-aware fuzzy-ontology-based tourism recommendation system. *Expert Syst. Appl.* **2022**, *199*, 116877. [CrossRef]
22. Morente-Molinera, J.; Trillo, F.C.; Pérez, I.; Herrera-Viedma, E. Managing Group Decision Making criteria values using Fuzzy Ontologies. *Procedia Comput. Sci.* **2022**, *199*, 166–173. [CrossRef]
23. Rahayu, N.; Ferdiana, R.; Kusumawardani, S. A systematic review of ontology use in E-Learning recommender system. *Comput. Educ. Artif. Intell.* **2022**, *3*, 100047. [CrossRef]
24. Cross, V.; Chen, S. Fuzzy ontologies: State of the art revisited. In *Proceedings of the International Conference of the North American Fuzzy Information Processing Society*; Springer International Publishing: Fortaleza, Brazil, 2018; pp. 230–242.
25. Ricardo, J.; Fernández, A.; Vázquez, M. Compensatory fuzzy logic with single valued neutrosophic numbers in the analysis of university strategic management. *Int. J. Neutrosophic Sci.* **2022**, *18*, 151–159. [CrossRef]
26. Broumi, S.; Dhar, M.; Bakhouyi, A.; Bakali, A.; Talea, M. Medical diagnosis problems based on neutrosophic sets and their hybrid structures: A 1]. *Neutrosophic Sets Syst.* **2022**, *49*, 1–8.
27. Ramesh, D.; Sanampudi, S.; Ramesh, D.; Sanampudi, S.K. An automated essay scoring systems: A systematic literature review. *Artif. Intell. Rev.* **2022**, *55*, 2495–2527. [CrossRef] [PubMed]
28. Ifenthaler, D. Automated essay scoring systems. In *Handbook of Open, Distance and Digital Education*; Springer Nature: Singapore, 2022; pp. 1–15.
29. Li, J. English Writing Feedback Based on Online Automatic Evaluation in the Era of Big Data. *Mob. Inf. Syst.* **2022**, *2022*, 9884273. [CrossRef]
30. Li, X.; Long, X.; Chen, S.L. A Review of Research on Automatic Scoring of English Reading. In *Proceedings of the ICNC-FSKD 2022: Advances in Natural Computation, Fuzzy Systems and Knowledge Discovery*; Springer: Berlin/Heidelberg, Germany, 2023; pp. 773–780.
31. Sriwanna, K. Text classification for subjective scoring using K-nearest neighbors. In Proceedings of the International Conference on Digital Arts, Media and Technology, Chiang Mai, Thailand, 1–4 March 2018; pp. 139–142.

32. Zupanc, K.; Bosnić, Z. Increasing accuracy of automated essay grading by grouping similar graders. In Proceedings of the 8th International Conference on Web Intelligence, Mining and Semantics, Novi Sad, Serbia, 25–27 June 2018; pp. 1–6.
33. Contreras, J.; Hilles, S.; Abubakar, Z. Automated essay scoring with ontology based on text mining and NLTK tools. In Proceedings of the International Conference on Smart Computing and Electronic Enterprise, Shah Alam, Malaysia, 11–12 July 2018; pp. 1–6.
34. Contreras, J.; Hilles, S.; Abubakar, Z. Automated Essay Scoring using Ontology Generator and Natural Language Processing with Question Generator based on Blooms Taxonomy's Cognitive Level. *Int. J. Eng. Adv. Technol.* **2019**, *9*, 2448–2457. [CrossRef]
35. Ajetunmobi, S.; Daramola, O. Ontology-based information extraction for subject-focused automatic essay evaluation. In Proceedings of the International Conference on Computing Networking and Informatics, Lagos, Nigeria, 29–31 October 2017; pp. 1–6.
36. Li, X.; Chen, M.; Nie, J. SEDNN: Shared and enhanced deep neural network model for cross-prompt automated essay scoring. *Knowledge-Based Syst.* **2020**, *210*, 106491. [CrossRef]
37. Huda, A.; Putri, M.; Awalluddin, A.; Sholeha, R. Text summarization of hadiths in Indonesian language using the combination of fuzzy logic scoring and latent semantic analysis (LSA). In Proceedings of the International Conference on Wireless and Telematics, Yogyakarta, Indonesia, 21–22 July 2022; pp. 1–4.
38. Mittal, H.; Devi, M. Computerized evaluation of subjective answers using hybrid technique. In *Proceedings of the International Conference on Innovations in Computer Science and Engineering*; Springer: Singapore, 2016; pp. 295–303.
39. Fauzi, M.; Utomo, D.; Setiawan, B.; Pramukantoro, E. Automatic essay scoring system using N-gram and cosine similarity for gamification based E-learning. In Proceedings of the International Conference on Advances in Image Processing, Beijing, China, 17–20 September 2017; pp. 151–155.
40. Zhang, B.; Liu, Y.; Zhang, B.; Liu, Y. Construction of English Translation Model Based on Neural Network Fuzzy Semantic Optimal Control. *Comput. Intell. Neurosci.* **2022**, *2022*, 9308236. [CrossRef] [PubMed]
41. Zhang, F.; Yu, L.; Shen, J. Automatic Scoring of English Essays Based on Machine Learning Technology in a Wireless Network Environment. *Secur. Commun. Netw.* **2022**, *2022*, 9336298. [CrossRef]
42. Sadanand, V.; Guruvyas, K.; Patil, P.; Acharya, J.; Suryakanth, S. An automated essay evaluation system using natural language processing and sentiment analysis. *Int. J. Electr. Comput. Eng.* **2022**, *12*, 6585–6593. [CrossRef]
43. Guo, X. An automatic scoring method for Chinese-English spoken translation based on attention LSTM. *EAI Endorsed Trans. Scalable Inf. Syst.* **2022**, *9*, e13. [CrossRef]
44. Wang, Q.; Wan, Y.; Feng, F. Human–machine collaborative scoring of subjective assignments based on sequential three-way decisions. *Expert Syst. Appl.* **2023**, *216*, 119466. [CrossRef]
45. Tashu, T.; Maurya, C.; Horvath, T. Deep Learning Architecture for Automatic Essay Scoring. *arXiv* **2022**, arXiv:2206.08232.
46. Birla, N.; Jain, M.K.; Panwar, A. Automated assessment of subjective assignments: A hybrid approach. *Expert Syst. Appl.* **2022**, *203*, 117315. [CrossRef]
47. Sahu, A.; Bhowmick, P.K. Feature Engineering and Ensemble-Based Approach for Improving Automatic Short-Answer Grading Performance. *IEEE Trans. Learn. Technol.* **2019**, *13*, 77–90. [CrossRef]
48. Mishra, P.; Parikh, R.; Sharma, P.; Parikh, R.; Joshi, D. Automatic Content Analyzer. *Int. J. Sci. Res. Comput. Sci. Eng. Inf. Technol.* **2019**, *5*, 818–822.
49. Farag, Y.; Yannakoudakis, H.; Briscoe, T. Neural automated essay scoring and coherence modeling for adversarial crafted input. *arXiv* **2018**, arXiv:1804.06898.
50. Li, X.; Chen, M.; Nie, J.; Liu, Z.; Feng, Z.; Cai, Y. Coherence-based automated essay scoring using self-attention. In *Proceedings of the China National Conference on Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data*; Springer International Publishing: Berlin/Heidelberg, Germany, 2018; pp. 386–397.
51. Zupanc, K.; Bosnić, Z. Automated essay evaluation with semantic analysis. *Knowledge-Based Syst.* **2017**, *120*, 118–132. [CrossRef]
52. Srivastava, K.; Dhanda, N.; Shrivastava, A. Optimization of Window Size for Calculating Semantic Coherence Within an Essay. *Adv. Distrib. Comput. Artif. Intell. J.* **2022**, *11*, 147–158. [CrossRef]
53. Darwish, S.; Mohamed, S. Automated essay evaluation based on fusion of fuzzy ontology and latent semantic analysis. In *Proceedings of the International Conference on Advanced Machine Learning Technologies and Applications*; Springer International Publishing: Berlin/Heidelberg, Germany, 2020; pp. 566–575.
54. Diniz-Filho, J.A.F.; Barbosa, A.C.O.F.; Collevatti, R.G.; Chaves, L.J.; Terribile, L.C.; Lima-Ribeiro, M.S.; Telles, M.P.C. Spatial autocorrelation analysis and ecological niche modelling allows inference of range dynamics driving the population genetic structure of a Neotropical savanna tree. *J. Biogeogr.* **2016**, *43*, 167–177. [CrossRef]
55. Susanti, M.N.I.; Ramadhan, A.; Warnars, H.L.H.S. Automatic essay exam scoring system: A systematic literature review. *Procedia Comput. Sci.* **2023**, *216*, 531–538. [CrossRef] [PubMed]
56. Chan, K.K.Y.; Bond, T.; Yan, Z. Application of an Automated Essay Scoring engine to English writing assessment using Many-Facet Rasch Measurement. *Lang. Test.* **2023**, *40*, 61–85. [CrossRef]
57. Ke, Z.; Ng, V. Automated Essay Scoring: A Survey of the State of the Art. In Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI-19), Macao, China, 10 August 2019; Volume 19, pp. 6300–6308.
58. Sharma, A.; Kabra, A.; Kapoor, R. Feature enhanced capsule networks for robust automatic essay scoring. In *Proceedings of the European Conference of Machine Learning and Knowledge Discovery in Databases*; Springer International Publishing: Berlin/Heidelberg, Germany, 2021; pp. 365–380.

59. Gomaa, W.; Fahmy, A. Ans2vec: A scoring system for short answers. In *Proceedings of the International Conference on Advanced Machine Learning Technologies and Applications*; Springer International Publishing: Berlin/Heidelberg, Germany, 2020; pp. 586–595.
60. Azman, A.; Alksher, M.; Doraisamy, S.; Yaakob, R.; Alshari, E. A framework for automatic analysis of essays based on idea mining. In *Proceedings of the Computational Science and Technology*; Springer: Singapore, 2020; pp. 639–648.
61. Alobed, M.; Altrad, A.; Bakar, Z.; Zamin, N. Automated Arabic essay scoring based on hybrid stemming with WordNet. *Malays. J. Comput. Sci.* **2021**, *31*, 55–67. [CrossRef]
62. Li, T.; Wang, S.; Lillis, D.; Yang, Z. Combining Machine Learning and Logical Reasoning to Improve Requirements Traceability Recovery. *Appl. Sci.* **2020**, *10*, 7253. [CrossRef]
63. Lee, Y.; Ke, H.; Yen, T.; Huang, H.; Chen, H. Combining and learning word embedding with WordNet for semantic relatedness and similarity measurement. *J. Assoc. Inf. Sci. Technol.* **2020**, *71*, 657–670. [CrossRef]
64. Polak, J.; Cook, D. A study on student performance, engagement, and experience with Kaggle in class data challenges. *J. Stat. Data Sci. Educ.* **2021**, *29*, 63–70. [CrossRef]
65. Ofoghi, B.; Mahdiloo, M.; Yearwood, J. Data Envelopment Analysis of linguistic features and passage relevance for open-domain Question Answering. *Knowl.-Based Syst.* **2022**, *244*, 108574. [CrossRef]
66. Sivakumar, R.; Arivoli, P. Ontology visualization PROTÉGÉ tools—A review. *Int. J. Adv. Inf. Technol.* **2011**, *1*, 1–11.
67. Farooq, M.; Saqlain, M. The selection of LASER as surgical instrument in medical using neutrosophic soft set with generalized fuzzy TOPSIS, WSM and WPM along with MATLAB coding. *Neutrosophic Sets Syst.* **2021**, *40*, 29–44.
68. Dettori, J.R.; Norvell, D.C. Kappa and Beyond: Is There Agreement? *Glob. Spine J.* **2020**, *10*, 499–501. [CrossRef]
69. Warrens, M.J. Kappa coefficients for dichotomous-nominal classifications. *Adv. Data Anal. Classif.* **2021**, *15*, 193–208. [CrossRef]
70. Park, Y.; Choi, Y.; Park, C.; Lee, K. EssayGAN: Essay data augmentation based on generative adversarial networks for automated essay scoring. *Appl. Sci.* **2022**, *12*, 5803. [CrossRef]
71. Zupanc, K.; Bosnic, Z. Automated essay evaluation augmented with semantic coherence measures. In Proceedings of the IEEE International Conference on Data Mining, Virtual Conference, 7–10 December 2014; pp. 1133–1138.
72. Wilson, J.; Roscoe, R.D. Automated Writing Evaluation and Feedback: Multiple Metrics of Efficacy. *J. Educ. Comput. Res.* **2020**, *58*, 87–125. [CrossRef]
73. Kakkonen, T.; Myller, N.; Sutinen, E.; Timonen, J. Comparison of dimension reduction methods for automated essay grading. *J. Educ. Technol. Soc.* **2008**, *11*, 275–288.
74. Liang, G.; On, B.; Jeong, D.; Kim, H.; Choi, G. Automated essay scoring: A Siamese bidirectional LSTM neural network architecture. *Symmetry* **2018**, *10*, 682. [CrossRef]
75. Dasgupta, T.; Naskar, A.; Dey, L.; Saha, R. Augmenting textual qualitative features in deep convolution recurrent neural network for automatic essay scoring. In Proceedings of the 5th Workshop on Natural Language Processing Techniques for Educational Applications, Melbourne, Australia, 19 July 2018; pp. 93–102.
76. Young, T.; Hazarika, D.; Poria, S.; Cambria, E. Recent trends in deep learning based natural language processing. *IEEE Comput. Intell. Mag.* **2018**, *13*, 55–75. [CrossRef]
77. Jiang, Z.; Liu, M.; Yin, Y.; Yu, H.; Cheng, Z.; Gu, Q. Learning from graph propagation via ordinal distillation for one-shot automated essay scoring. In Proceedings of the Web Conference, Ljubljana, Slovenia, 19 April 2021; pp. 2347–2356.