applied sciences

MDPI

Article

One-Class Learning for AI-Generated Essay Detection

Roberto Corizzo [1,*] and Sebastian Leal-Arenas [2]

[1] Department of Computer Science, American University, Washington, DC 20016, USA
[2] Department of Linguistics, University of Pittsburgh, Pittsburgh, PA 15260, USA; sal209@pitt.edu
* Correspondence: rcorizzo@american.edu

Abstract: Detection of AI-generated content is a crucially important task considering the increasing attention towards AI tools, such as ChatGPT, and the raised concerns with regard to academic integrity. Existing text classification approaches, including neural-network-based and feature-based methods, are mostly tailored for English data, and they are typically limited to a supervised learning setting. Although one-class learning methods are more suitable for classification tasks, their effectiveness in essay detection is still unknown. In this paper, this gap is explored by adopting linguistic features and one-class learning models for AI-generated essay detection. Detection performance of different models is assessed in different settings, where positively labeled data, i.e., AI-generated essays, are unavailable for model training. Results with two datasets containing essays in L2 English and L2 Spanish show that it is feasible to accurately detect AI-generated essays. The analysis reveals which models and which sets of linguistic features are more powerful than others in the detection task.

Keywords: essay classification; one-class learning; linguistics; L2 writing; ChatGPT

Citation: Corizzo, R.; Leal-Arenas, S. One-Class Learning for AI-Generated Essay Detection. *Appl. Sci.* **2023**, *13*, 7901. https://doi.org/10.3390/app13137901

Academic Editors: Ahmed Rafea and Julian Szymanski

Received: 20 May 2023
Revised: 1 July 2023
Accepted: 3 July 2023
Published: 5 July 2023

1. Introduction

The rapid advancement and availability of Artificial Intelligence (AI) and Natural Language Processing (NLP) technologies to the general public have raised significant concerns regarding the authenticity of written content, especially in academic settings [1,2]. The current academic research on the use of AI tools, such as ChatGPT (Generative Pre-trained Transformer), has predominantly followed two avenues of inquiry: the first explores the applications and potential uses of AI [3], while the second focuses on developing methodologies to discern between human-generated and AI-generated writing. Tools have been developed to classify text, with GPTZero and the TurnItIn extension being the most relevant in academic settings. Nonetheless, one important drawback to most of these detection tools is the fact that English is the only supported language or the language on which they are based. This is problematic in language courses, where written production is part of the curriculum. Moreover, it was observed that GPTZero and Turnitin are often inconsistent in recognizing human written text, making them unreliable to be deployed without a substantial risk of false positives (https://nerdschalk.com/is-gptzero-accurate-detect-chat-gpt-detector-tested/ (accessed on 1 July 2023) ; https://teaching.pitt.edu/newsletter/teaching-center-update-for-june-9-2023/ (accessed on 1 July 2023). Given that information on the algorithmic foundation work of these tools is vague or not publicly disclosed, it is of paramount importance to investigate and design new transparent detection models specially designed for L2 users, a growing population in academic contexts (https://www.statista.com/statistics/233880/international-students-in-the-us-by-country-of-origin/ (accessed on 1 July 2023)), who encounter extra challenges when utilizing the internet [4]. Examples of these challenges include inadequate search techniques, dependence on unsuitable sources, and directly copying and pasting text without proper citation [5–7].

Popular machine learning methods for text classification can be characterized as neural-network-based and feature-based. The former involves neural network model architectures, which extract a hidden representation of the text, or, in some cases, multiple modalities where available, with high discriminating power, and directly conduct the classification task in an end-to-end fashion. The latter adopts NLP tools to extract feature vectors from text, e.g., frequency and fluency, which in turn are used to train external machine-learning classification models, such as Random Forest (RF) and Support Vector Machines (SVMs).

Neural-network-based models include Bi-LSTM, which is particularly popular in sentiment classification [8,9]. Lately, a consistent research effort has been devoted to multi-modal neural network models, which support the combined analysis of multiple modalities, or aspects, of the data. The authors in [10] performed deep multi-task learning, integrating novelty detection, emotion recognition, sentiment prediction, and misinformation detection within the same architecture. The work in [11] leverages multivariate data fusion via Independent Vector Analysis and pre-trained deep learning models to accomplish misinformation detection during high-impact events. Similarly, the authors in [12] proposed the adoption of a Transformer-based model to analyze multiple modalities as intermediate tasks, i.e., rumor score prediction and event classification extracting hidden relationships across various modalities that facilitate the final task. The authors in [13] proposed a multimodal fusion neural network model with dual attention applied to rumor detection. The method extracts text and image features, enhancing its detection capabilities. A Character Text Image Classifier (CTIC) model was proposed in [14], where the model builds upon BERT, Capsule Network, and EfficientNet, involving different types of embeddings (word, character, and sentence level), as well as images. The work in [15] investigated the interplay between news content and users' posting behavior clues in detecting fake news, with convolutional and bidirectional gated recurrent unit neural networks coupled with a self-attention mechanism. The proposed architecture was shown to be effective in learning rich semantic and contextual representations of news for the detection task. The authors in [16] devised a text and audio database with sentences extracted from videos presenting real-world situations and categorized them into three classes: neutral sentences, insulting sentences, and sentences representing unsafe conditions. A deep neural network was proposed to jointly consider text and audio embedding vectors, which resulted in accurate detections of unsafe and insulting situations.

Shifting the focus to feature-based methods, the use of linguistic features for human versus machine-generated text has demonstrated notable advantages. Their transparency, explainability, and discriminative capabilities make them an effective approach to representing the underlying data [17]. The reliance on linguistic features in methods that aim to differentiate human from AI-generated text is based on the assumption that differences exist between these two forms of text across specific dimensions. Feature-based approaches have been employed in the assessment of readability [18,19] and authorship attribution [20]. Text-related feature analysis has encompassed the computation of the distribution of punctuation marks, frequency of punctuation within sentences, and average length of sentences and paragraphs. These measures have been effective when detecting fake news [17,21]. AI-generated texts exhibit distinctive features, such as the frequent repetition of words and expressions [22], as well as the utilization of commonly used words in a given language. These patterns have been detected by automated detection techniques that consider stop-words, unique words, and top lists [23]. Repetitiveness has been represented by the calculation of lexical repetition through the n-gram overlap of words [24]. In contrast to human text, AI-generated text avoids language that is not commonly used. Furthermore, machine-generated texts lack emotional semantics and personal biases found in human language. As a result, sentiment-related words tend to receive higher scores when determining a text's topicality using lexicon-based sentiment analysis tools [25]. Readability refers to the ease with which a piece of written text can be understood by readers. The factors that determine a text's readability are vocabulary complexity, sentence structure and length, and text organization. Diverse readability measurements are available

to determine the suitability of a text for a specific audience [26]. These measures use mathematical formulas to provide an estimation of how accessible and comprehensible a text can be [17]. AI-generated texts have been estimated as easier to understand given the scores obtained by readability indexes [17,18,26]. As noted, human text uses words that are more emotionally charged than machine-generated text. Previous studies have used Part-of-Speech (POS) Tagging [18], an NLP task that involves assigning grammatical labels to each word in a given text based on the part of speech to which it belongs. Thus, words that are associated with emotions, such as adjectives and adverbs, present a higher frequency in human-generated text. In general, linguistic feature extraction is a viable approach to quantitatively represent relevant properties of the language, which supports machine learning models in text classification tasks.

A common limitation of both neural-network-based and feature-based methods stands in their supervised nature, which relies on data availability for both classes to perform model training. Initial studies towards AI-generated essay detection also followed supervised approaches that require both human and AI-generated essays in training data [27]. However, this requirement becomes arduous due to the scarce availability of data for the positive class (phenomenon or anomaly class to detect) and due to the potentially different forms that the phenomenon could take, which are unknown at training time. In such a setting, one-class learning and anomaly detection approaches appear more ideal than fully supervised models.

The adoption of one-class models is particularly important for the AI-generated essay detection task since AI-generated essays are not readily available to train and update the models by the end users of detection tools, i.e., instructors. Therefore, a one-class learning approach would offer the possibility to train models using exclusively human-written essays.

The authors in [28] compare six unsupervised anomaly detection methods of varying complexity to determine a relationship between model complexity and effectiveness in detecting anomaly types. An anomaly detection approach for sequential data is proposed in [29], where tabular data are transformed into an image representation by means of a specialized filter, allowing the model to detect multiple types of outliers simultaneously. The work in [30] proposes a digital-twin-driven GAN-based method for an oil and gas station anomaly detection method. A spatially aware approach is proposed in [31] to deal with the detection of anomalies in geo-distributed smart grid data observed in smart grids. The authors in [32] propose a method to automatically highlight changing trends and anomalous behavior in music streaming data, which also delivers two levels of outlier explanations: the deviation from the model prediction and the explanation of the model prediction. The work in [33] leverages one-class models to perform anomaly detection with active learning in a continual learning setting, focusing on the cybersecurity domain. One-class ensembles with diverse models have shown to be effective in detecting rare biological sequences [34], where available data are characterized by almost only negative examples. Despite the wide range of applications covered by one-class learning studies, approaches addressing textual content are scarce and none so far has been devoted to essay data. On the other hand, supervised neural-network-based and feature-based approaches have been more pervasively applied to text data, dealing with issues pertaining to social media, such as fake news, hate speech, and misinformation. As a result, the effectiveness of different models in the context of education and academia is still unexplored.

The present paper fills this gap, inspired by the ongoing discussion on the impact of ChatGPT in academic contexts. Specifically, this work investigates the impact of linguistic features on the predictive capabilities of one-class learning models on AI-generated essay detection. The contribution of this paper is threefold. First, it explores the adoption of linguistic features and one-class learning models for AI-generated essay detection. Second, it designs two datasets leveraging two real-world essay corpora in two languages, English and Spanish. These essays were written by L2 speakers and were complemented by AI-generated counterparts written by ChatGPT. Third, it proposes an extensive experimental

evaluation that involves multiple one-class learning methods against different sets of linguistic features, measuring and discussing their performance on the AI-generated essay detection task.

2. Materials and Methods

The proposed method consists of two main components: (i) linguistic features, which provide us with a concise numerical description of different text data characteristics, and (ii) one-class models, which are trained using linguistic features extracted from a training set of human essays and are able to extract predictions (human or AI) for a new set of unlabelled essays. A graphical overview of this process is shown in Figure 1.

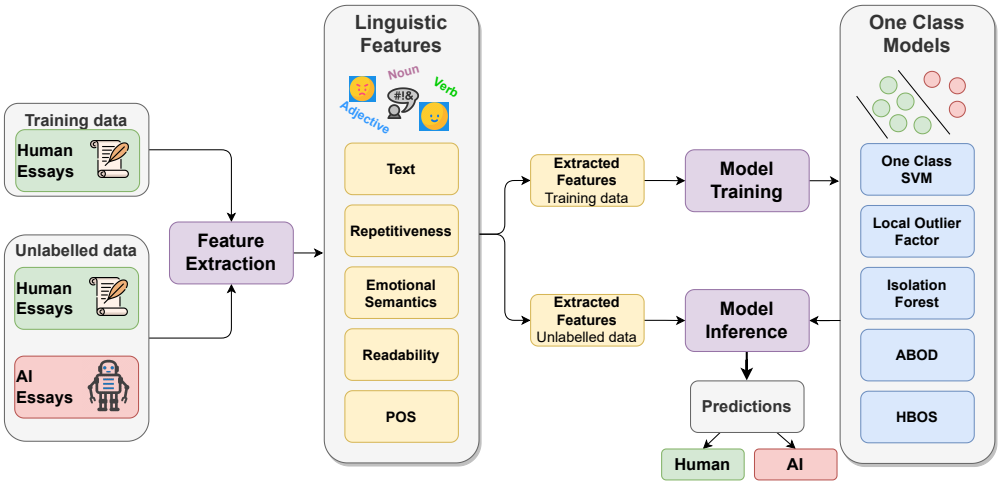

Figure 1. Proposed method for human vs. AI-generated essay classification. Linguistic features are extracted for human essays (negative data) available in a training corpus and are exploited to train one-class models. At prediction time, linguistic features are extracted from unlabelled data, and trained models extract a prediction without any usage of AI-generated essays (positive data) during their training stage.

2.1. Feature Extraction

The method involves the extraction of linguistic features, which aid in the distinction between human and AI-generated texts [19]. The study encompasses the use of five categories: Text, Repetitiveness, Emotional Semantics, Readability, and Part-of-Speech.

- **Text**: This category possesses features found in textual data, which can be employed as input for machine learning models since they are measurable. The six Text features in this study include distinct type of punctuation marks, number of Oxford commas, number of paragraphs [35], number of full stops, number of commas [35], and average sentence length [36]. Previous studies have demonstrated that human-generated texts typically contain a higher frequency of punctuation mark use [35]. However, the examination of distinctive types of punctuation marks has been largely overlooked. Similarly, the number of paragraphs, full stops, commas, and average sentence length are higher in human texts [35]. The Oxford comma was included (comma before the final element in a list) as a distinguishing feature for its use is prevalent in certain English dialects but absent in Spanish, thus acting as a language transfer feature when analyzing essays written by English-speaking learners of Spanish. Moreover, language, more specifically dialect bias [37], will be evident in AI-generated texts if the Oxford comma is employed.
- **Repetitiveness:** Features in this category were compiled due to the fact that AI-generated texts frequently display the use of similar words and phrases, resulting in monotonous writing and lack of narrative and linguistic diversity [22]. The set of unique n-grams, iterations over the set, and count of the number of times each n-

gram appears in the text are extracted to compute the unigram, bi-gram, and tri-gram overlap in order to quantify repetitiveness properties. The overlap is then calculated as a ratio between the count and the total number of different n-grams [38]. Additionally, the frequency of the words in the data is compared to the top 5K and 10K words in each language [24], thus determining how closely the lexicon in the dataset matches that of everyday speech. What is more, the lexical diversity of human and AI-generated texts can also be studied. Punctuation from essays was removed to extract the list of tokens in order to determine the number of matches with the most-used words.

- **Emotional Semantics**: This set of features represents the texts' emotional tone, expressed through words, phrases, and sentences [39]. Text that has been created by humans has been said to have more subjectivity, complexity, and emotional diversity. AI-generated text, on the other hand, is more likely to be consistent in tone and emotionally neutral [40]. Polarity and Subjectivity, two of the emotional semantic features, are extracted using TextBlob, a lexicon-driven open-source Python text data processing package. Polarity expresses sentiment on a scale from −1.0 to 1.0 [−1.0: negative; 0.0: neutral; 1.0: positive], and Subjectivity indicates the degree by which personal feelings, views, beliefs, opinions, allegations, desires, beliefs, suspicions, and speculations are expressed in the essay, with a range from 0.0 to 1.0 [0.0: very objective; 1.0: very subjective] [41]. Three model-driven techniques were used to capture the semantics of sentiments: Sentiment (ES), Sentiment (Multi-language), and Sentiment Score (Multi-language). For the first one, a Naive Bayes model trained using over 800,000 reviews from El Tenedor, Decathlon, Tripadvisor, Filmaffinity, and eBay was used (https://github.com/sentiment-analysis-spanish/sentiment-spanish (accessed on 1 July 2023)). For Sentiment (Multi-language), the bert-base, A-multilingual-uncased-sentiment model was trained on reviews in different languages: English 150K, Dutch 80K, German 137K, French 140K, Italian 72K, Spanish 50K. From this model, three categories are obtained: 1- and 2-star reviews are considered negative [−1.0], 3-star reviews are neutral [0.0], and 4- and 5-star reviews are positive [1.0]. For the last feature, Sentiment Score (Multi-language), the raw star ratings were normalized in a 0–1 range [0.0: one star, 1.0: five stars], obtaining a continuous score.
- **Readability**: The fourth set of features encompasses the complexity of a text's vocabulary [18,26] through the use of thirteen different indicators, four of them being exclusive to the Spanish language. Each indicator possesses its own unique formula. The Flesch Reading Ease Score indicates how difficult a passage is to understand based on word length and sentence length. Scores range from 0 to 100, with higher scores denoting material that is easier to read and easily understood by elementary school students. Lower scores depict a text that is difficult to read, best understood by people with tertiary education. The Flesch Kincaid Grade Score also determines the readability of a text, but the score that it provides is aligned with a US grade level. Given the formula, there is no upper bound. However, the lowest possible score, in theory, is −3.40. This indicator emphasizes sentence length over word length. The SMOG Index Score estimates the years of education necessary to comprehend a text. The Coleman Liau Index Score measures the understandability of a text, whose output indicates the necessary US grade level to comprehend the text. The score ranges from 1 to 11+. Similarly, The Automated Readability Index Score measures the understandability of a text by providing a representation of the US grade level required to understand the text. Its scores range from 1 to 14, where 1 indicated kindergarten and 14 college student. The Dale–Chall Readability Score uses a list of 3000 words, which could be easily understood by fourth graders. The scores range from 4.9 or lower to indicate that the text is understood by an average fourth grader or lower to 9.9 to denote that the text is understood by college students. The Difficult Words Score provides a value based on the number of syllables a word contains. Words with more than two syllables are considered difficult to understand. The Linsear Write Formula Score provides a score that indicates the grade level of a text considering sentence length and the

number of words with three or more syllables. Values range from 0 to 11+. The Gunning Fog Score estimates the years of formal education required to comprehend a text on the first reading. Scores range from 6, sixth grade, to 17, college graduate. The following four indexes were used in the Spanish dataset. The Fernandez-Huerta Score stems from the Flesch Reading Ease Index, with similar score interpretation, where higher scores indicate that a text is easier to understand than lower scores. The Szigriszt-Pazos Score is also based on the Flesch Reading Ease Index, with similar score interpretation. However, the levels of difficulty of the text can also be associated with a type of text; i.e., a score between 1 and 15 indicates that a text is very hard to understand and that it is a scientific or philosophical publication. Conversely, a score between 86 and 100 indicates that the text is very easy to read and is usually a comicbook or a short story. The Gutiérrez de Polini Score is not an adapatation of any index and was designed for school-level Spanish texts. A low score indicates that the readability of the text is more difficult. The Crawford Score indicates the years of schooling that a person requires in order to understand the text. The utilization of diverse readability indexes provides valuable information about the complexity and accessibility of each text.

- **Part-of-Speech (POS)**: The aforementioned absence of syntactic and lexical variance in machine-generated texts can be quantified by identifying the distribution of Parts of Speech. These word classes are useful to analyze, understand, and construct sentences [42]. POS tagging has been introduced in previous work [18,24] to indicate the relative frequency of word types based on per-sentence count of most classes. Given the complexity of languages, where word order could alter the meaning of a word or where words that use different morphemes possess a similar meaning, SpaCy (https://spacy.io/ (accessed on 1 July 2023)) was used to parse and tag the data according to the linguistic context of each token. In terms of morphologically inflected words, the process by which words are modified to convey grammatical categories (number, tense, person), the lemma (root form; without inflection) becomes a token. A schematic view of all the features considered in this study is shown in Table 1.

Table 1. Linguistic features considered in the study: types, names, and example (where applicable). An asterisk symbol (*) denotes that the feature value is normalized by the essay length (number of words). A plus symbol ($^+$) denotes that the feature is novel, i.e., never encountered by the authors in other text classification works.

Type	Feature	Example
Text	Distinct types of punctuation marks $^+$	. , ; …
	Number of Oxford commas $^+$	“ …Christmas, Easter, and spring break…”
	Number of paragraphs	/
	Number of full stops *	/
	Number of commas *	/
	Average sentence length *	/
Repetitiveness	Unigram Overlap	only …only …
	Bi-gram Overlap	only you …only you …
	Tri-gram Overlap	only you can …only you can …
	Matches in the 5K most common words *	from, your, have …
	Matches in the 10K most common words *	scared, infringement, bent …
Emotional Semantics	Polarity	$(-1.0, 1.0)$
	Subjectivity	$(0.0, 1.0)$
	Sentiment (ES) $^+$	$(-1.0, 1.0)$
	Sentiment (Multi-language) $^+$	$(-1.0, 1.0)$
	Sentiment Score (Multi-language) $^+$	$(0.0, 1.0)$

Table 1. *Cont.*

Type	Feature	Example
Readability	Flesch Reading Ease Score	(0, 100)
	Flesch Kincaid Grade Score	(−3.40, no limit)
	Smog Index Score	(1, 240)
	Coleman Liau Index Score	(1, 11+)
	Automated Readability Index	(1, 14)
	Dale–Chall Readability Score	(0, 10)
	Difficult Words Score	Varies
	Linsear Write Formula Score	(0, 11+)
	Gunning Fog score	(6–17)
	Fernández-Huerta Score (SPAN)	(0, 100)
	Szigriszt-Pazos Score (SPAN)	(0, 100)
	Gutiérrez de Polini Score (SPAN)	(0, 100)
	Crawford Score (SPAN)	(6–17)
Part-of-Speech (POS)	ADJ (Adjective) *	current, long
	ADP (Adverbial Phrase) *	during the week, right there
	ADV (Adverb) *	hopefully, differently
	AUX (Auxiliary) *	"…as you **have** seen…"
	CCONJ (Coordinating Conjunction) *	and, but, or
	DET (Determiner) *	the, a, an
	NOUN *	teachers, adoption
	NUM *	two, hundreds
	PRON (Pronoun) *	I, you, him, hers
	PROPN (Proper Noun) *	Sweden, United States
	PUNCT (Punctuation) *	. ,
	SCONJ (Subordinating Conjunction) *	although, because, whereas
	SYM (Symbol) *	symbol that is not punctuation, e.g., $
	VERB *	mirror, use, have
	SPACE (Number of spaces) *	count of number of spaces

2.2. One-Class Models

Once linguistic features are extracted, they are used to train and extract inferences using one-class learning models. A set of popular and diverse models is adopted in the study. Specifically:

- **One-Class Support Machines (OCSVMs)** [43]: They conceptually operate in a similar way to Support Vector Machines, which identify a hyperplane to separate data instances from two classes. However, the one-class learning counterpart uses a hyperplane to encompass all of the background data instances (human essays). Solving the OCSVM optimization problem corresponds to solving the dual quadratic programming problem:

$$\min_{\alpha} \frac{1}{2} \sum_{ij} \alpha_i \alpha_j K(x_i, x_j)$$

subject to the constraints $0 \leq \alpha_i \leq \frac{1}{\nu l}$ and $\sum_i \alpha_i = 1$, where α_i is the weight for the instance i, vectors with non-zero weights are defined as support vectors and determine the optimal hyperplane, ν is a parameter that represents a trade-off between the distance of the hyperplane from the origin and the number of instances covered by the hyperplane, l is the number of instances in training data, and $K(x_i, x_j)$ is the kernel function. Leveraging the kernel function to project input vectors into a feature space allows for nonlinear decision boundaries. Specifically, a feature map can be defined as:

$$\phi : X \to \mathbb{R}^N,$$

where ϕ maps training vectors from the input space X to a dimensional feature space, and the kernel function is defined as:

$$K(x, y) = \langle \phi(x), \phi(y) \rangle$$

The adoption of kernel values $K(x, y)$ allows to avoid the explicit computation of feature vectors, with great improvement in the computational efficiency. Common kernels include Linear, Radial Basis Function (RBF), Polynomial, and Sigmoid. Following the training phase, the OCSVM-learned hyperplane can categorize a new data instance (essay) as regular/normal (human) or different/anomaly (AI-generated) with regard to the training data distribution based on its geometric location within the decision boundary.

- **Local Outlier Factor (LOF)** [44]: The method measures the deviation in local density of data instances (essays) with respect to their neighbors. The anomaly score returned by LOF is based on the ratio between the local density of the data instance and the average local density of the nearest neighbors. Considering the k-distance (A) as the distance of instance A from its k-th nearest neighbors, it is possible to define the notion of reachability distance:

$$RD_k(A, B) = max\{k - d(B), distance(A, B)\}.$$

Instances that belong to the k nearest neighbors of B are considered to be equally distant. The local reachability density of an instance A defined as $lr_k(B)$ is the inverse of the average reachability distance of the object A from its neighbors. Local reachability densities are compared with those of the neighbors as:

$$LOF_k(A) := \frac{\Sigma_{B \in N_k(A)} \frac{lr_k(B)}{rr_k(A)}}{|N_k(A)|} = \frac{\Sigma_{B \in N_k(A)} lr_k(B)}{|N_k(A)| \cdot lr_k(A)},$$

which represents the average local reachability density of neighbors divided by the local reachability density of the instance. A value lower than 1 is indicative of a higher density than neighbors (inlier), whereas a value greater than 1 is indicative of a lower density than neighbors (outlier).

- **Isolation Forest** [45]: It uses a group of tree-based models and calculates an isolation score for every data instance (essay). The average distance that lies from the tree's root to the leaf associated with the data instance, corresponding to the number of partitions required to reach the instance, is used to compute the anomaly score. Considering that more noticeable variations in values relatively correspond to shorter paths in the tree, the method uses this information to distinguish one abnormal observation (AI-generated essays) from the rest (human essays). The anomaly score is defined as

$$s(x, n) = 2^{-\frac{E(h(x))}{c(n)}},$$

where $h(x)$ is the path length of the instance x, $c(n)$ is the average path length of unsuccessful search in the Binary Search Tree, and n is the number of external nodes. An anomaly score close to 1 indicates that an instance has a high chance to be an anomaly (AI-generated essay), whereas scores smaller than 0.5 are indicative of a normal instance (human essay).

- **Angle-Base Outlier Detection (ABOD)** [46,47]: is a widely adopted method that calculates the variance of weighted cosine scores between each data instance and its neighbors and uses that variance as the anomaly score. It has the ability to reduce the frequency of false positive detections by effectively identifying relationships in high-dimensional spaces between each instance and its neighbors rather than interactions among neighbors.

- **Histogram-based Outlier Score (HBOS)** [48]: It is a straightforward statistical approach for anomaly detection that assumes feature independence. The main concept is the generation a histogram for each feature of the data. Subsequently, for each instance, the method multiplies the inverse height of the bins associated to it, providing an assessment of the density of all features. This behavior is conceptually similar to the Naive Bayes algorithm for classification, which is known for multiplying all independent feature probabilities. Histograms represent a quick and effective way to identify anomalies (AI-generated essays). Even though feature relationships are ignored (i.e., the method assumes feature independence), this simplification allows the method to converge quickly. HBOS builds histograms in two different modalities: (i) static bin sizes and a preset bin width, and (ii) dynamic bins with a close-to-equal number of bins.
- **AutoEncoder:** A neural network model used to learn efficient representation of unlabeled data in an unsupervised manner. It learns two functions: an encoding function that transforms the input data, and a decoding function that recreates the input data from the encoded representation. The learned representation can be used to address detection tasks by performing the reconstruction of new data instances and comparing their reconstruction error with that of the learned distribution.

2.3. Research Questions

This experimental study is devised to address the following research questions:

- **RQ1:** Is it possible to accurately detect human vs. AI-generated essays using one-class learning models, i.e. without exploitation of AI-generated essays for model training?
- **RQ2**: Do linguistic features allow one-class models to accurately classify human vs. AI-generated essays, and which ones are the most relevant for essay classification?
- **RQ3:** Are there substantial differences in detection accuracy in essays written in L2 English and L2 Spanish?

In order to answer the RQs, specific datasets, setup, and metrics are used.

2.4. Datasets

The dataset (the datasets are available at the following public repository: https://github.com/rcorizzo/one-class-essay-detection (accessed on 1 July 2023)) for this study encompasses essays written by L2 English learners, L2 Spanish learners, and AI-generated counterparts. For the human-generated essays, two corpora were used:

1. **L2 English:** the Uppsala Student English Corpus (USE) (https://ota.bodleian.ox.ac.uk/repository/xmlui/handle/20.500.12024/2457 (accessed on 1 July 2023)), which consists of 1489 essays written by Swedish university students learning English. The majority of the essays belong to students in their first term of full-time studies. Essays found in the 'a2' folder in the repository were used for two main reasons: the substantial amount of essays in that folder and the nature of the task. Learners wrote argumentative essays on diverse topics. While argumentative essays are linguistically more complex than introductory or personal experience texts, they are easier to generate using AI. The language proficiency of the students (A2) reflects the performance of a basic user of the language. The total number of human-generated essays in English was 335.
2. **L2 Spanish:** To obtain L2 Spanish data, the UC Davis Corpus of Written Spanish, L2 and Heritage Speakers (COWSL2HS (https://github.com/ucdaviscl/cowsl2h (accessed on 1 July 2023))) was used. The corpus includes essays written by students in university-level Spanish courses. Essays written only by L2 learners were considered. Similarly to L2 English essays, the language proficiency of the learners was that of a basic user and topics were varied; 350 essays were compiled for analysis.

The two datasets were complemented with the same amount of AI-generated essays. ChatGPT wrote essays in both languages. For English, the following propmpt was used: "Write an 800-word essay on [topic] as a second language speaker". The same instruction

was used to collect Spanish texts: "Escribe un ensayo de 800 palabras sobre [tema] como un hablante de segunda lengua". [topic] and [tema] were matched for each student essay in the dataset. Data collection occurred in December 2022, a few days after the release of ChatGPT. Thus, drawbacks occurred when generating essays using this tool: word limits or paragraph lengths were not considered in the output provided; requesting the telling of a personal story yielded no results since ChatGPT has no experiences; after the first inquiries on a topic, the output exhibited a high degree of repetition; and there exists a limit on the number of inquiries within an hour. The English dataset is comprised of 670 essays, 335 found in the corpus and 355 generated by AI. The Spanish dataset is made up of 700 essays, equally distributed between human (350) and machine-generated (350) essays. Table 2 shows examples of the dataset.

Table 2. Short essay samples in our datasets. Second paragraph in an essay for the English dataset (top), and first sentences of an essay for the Spanish dataset (bottom).

Topic: Capital punishment	
Human	**Generated**
. . . The right to life is the most fundamental part of the Human Rights Agenda. To claim the right to life for innocent people is not very difficult. It goes without saying. It is considerably harder to defend people with horrible crimes on their consciences; mass-murderers, war criminals, paedophiles etc. But that is how universal rights works. They applies to everyone, everywhere and during all circumstances. . .	. . . One of the main reasons why capital punishment is not good is that it is not an effective deterrent to crime. Studies have shown that the death penalty does not have a greater deterrent effect on crime than other forms of punishment, such as life imprisonment. Additionally, the death penalty can also have a negative impact on the crime rate, as it can lead to a lack of public trust in the justice system, which can in turn lead to a rise in crime. . .
Topic: Una persona famosa	
Human	**Generated**
Hay un artista y se llama Kanye West. El tiene muchos dinero. El tiene cuarenta y cuartro anos de edad. El es de estados unicos y es americano. El tiene dos ojos grandes. El lleva una camiseta blanca y pantalones negro o azul. El lleva zapatos blancos. Tiene una esposa y llamada Kim Kardashian West. El tiene una hija y un hijo. Sus nombres son North West y Saint West. El escribe musica El trabaja con artistas de musica . . .	Una persona famosa que me viene a la mente es Albert Einstein. Einstein es conocido en todo el mundo como uno de los científicos más importantes y brillantes de la historia. Nacido en Alemania en 1879, Einstein comenzó a mostrar una gran aptitud para las matemáticas y la física desde temprana edad. A los 16 años, se matriculó en la Universidad Politécnica de Zurich, donde obtuvo su título en 1900 . . .

2.5. Setup and Metrics

One-class learning models are trained following an evaluation scheme that resembles a stratified 5-fold cross-validation, with the exception that training folds contain exclusively human essays and testing folds contain both human and AI-generated essays. As a result, models are trained and evaluated 5 times and results are aggregated. For OneClassSVM, experiments were performed with different kernel types (RBF, Polynomial, Linear) and results were reported separately. For LocalOutlierFactor, experiments with different numbers of neighbors (3, 5, 10) were performed. For AutoEncoder, a model architecture with three hidden layers of 10, 5, and 10 neurons was used. The model was trained for 200 epochs with a batch size of 32. For all other methods and parameters, considering that the experimental conditions are standard with respect to the common one-class learning setting, the recommended configurations proposed in the respective software libraries (SkLearn (https://scikit-learn.org/ (accessed on 1 July 2023)) and PyOD (https://pyod.readthedocs.io/en/latest/index.html (accessed on 1 July 2023))) were used. Commonly used metrics for machine-learning-based classification tasks, such as Precision, Recall, and F1-Score, were adopted.

3. Results

RQ1: The results in Table 3 for English indicate that OneClassSVM (RBF kernel) achieves the best performance in terms of F1-Score in three of the five settings, i.e., Text

(0.6694), Emotional Semantics (0.6436), and POS (0.7223). AutoEncoder presents exceptional F1-Score performance in the Repetitiveness setting (0.9553), outperforming all methods in this setting. Isolation Forest presents a very high F1-Score performance in Readability (0.8942), surpassing all other methods. The results for OneClassSVM with other kernels show relatively low F1-Score performance, except for the Polynomial kernel with Readability features, which possesses the highest performance level (0.7292). LocalOutlierFactor yields satisfactory results in the Readability setting, where its performance increases exponentially: 3 (0.7582), 5 (0.7995), and 10 (0.8236). The method is unsatisfactory in all other settings. HBOS exhibits excellent F1-Score performance in the Repetitive setting (0.9027). However, its performance is sub-optimal in Readability (0.7085) and falls below random performance in the remaining settings. ABOD presents subpar F1-Scores in all settings.

Results for the Spanish data in Table 4 show that OneClassSVM (RBF kernel) achieves the best performance in four of the five settings, i.e., Text (0.5917), Repetitiveness (0.5199), POS (0.4968), and Emotional Semantics (0.4881). The results for OneClassSVM with other kernels present a subpar F1-Score performance, except for the Linear kernel with Readability features, which represents the highest performance across all settings (0.7238) and methods. LocalOutlierFactor presents a poor F1-Score across all neighbor configurations (3, 5, 10) in the Repetitiveness (0.4426, 0.4369, 0.4223) and Emotional Semantics (0.4663, 0.4501, 0.4684) settings. Its performance is notably poorer in the remaining three settings. IsolationForest showcases similar F1-Score performance to LocalOutlierFactor in Readability (0.5788) and Emotional Semantics (0.4499) settings. Its performance is subpar in Repetitiveness (0.3424), POS settings (0.3247), and Text (0.3132). ABOD, HBOS, and AutoEncoder present unsatisfactory performance (below random) in all settings.

RQ2: Table 3 shows that linguistic features aid in the accurate prediction of human vs. AI-generated essays in English. Based on the average F1-Score of each setting across all methods, Readability (0.7166) is the best-performing feature, preceded by Repetitiveness, POS (0.5345), and Emotional Semantics (0.5120). Text had the lowest performance (0.3994). The maximum F1-Score in each setting across all methods reveals that Repetitiveness (0.9553) scored the highest, preceded by Readability (0.8942), POS (0.7223), Text (0.6694), and Emotional Semantics (0.6436). Readability and Repetitiveness are the best-performing features, as shown by both average and maximum F1-Score across all methods. For Spanish (Table 4), average F1-Scores revealed that Readability had the highest score (0.5056), preceded by Emotional Semantics (0.4165), Repetitiveness (0.4189), and POS (0.3699). Text features present the poorest performance (0.3509). On the other hand, maximum F1-Scores reveal a different order of features, from highest to lowest: Readability (0.7238), Text (0.5917), Repetitiveness (0.5199), POS (0.4968), and Emotional Semantics (0.4881). Readability emerges as the top-performing setting, displaying the highest average and maximum performance despite the generally sub-optimal results.

The results obtained with the extraction of confusion matrices show that the secondary diagonal of confusion matrices offers a different perspective than performance metrics. For models with multiple configurations (OneClassSVM and LocalOutlierFactor), the confusion matrix obtained with the best configuration (kernel type and number of neighbors, respectively) is provided.

For the English dataset, the Text setting (see Figure 2) shows 344 (AutoEncoder), 340 (ABOD), 331 (HBOS), 262 (IsolationForest), 332 (LocalOutlierFactor), and 207 (OneClassSVM) incorrectly classified essays out of 671. The difference between the best-performing method (OneClassSVM) and the second-best (IsolationForest) method is substantial and the results are not optimal. The Repetitiveness setting (see Figure 3) highlights 294 (ABOD), 265 (LocalOutlierFactor), 262 (OneClassSVM), 65 (HBOS), 39 (IsolationForest), and 30 (AutoEncoder) incorrectly classified essays, making AutoEncoder and IsolationForest the top-performing methods.

Table 3. English dataset: detection accuracy via 5-fold stratified cross-validation. Highest F1-Scores in bold.

Setting = Text	**Precision**	**Recall**	**F1-Score**
OneClassSVM (RBF kernel)	0.7607	0.6915	**0.6694**
OneClassSVM (Polynomial kernel)	0.3092	0.3294	0.3110
OneClassSVM (Linear kernel)	0.3092	0.3294	0.3110
LocalOutlierFactor (3 neighbors)	0.5169	0.5052	0.4084
LocalOutlierFactor (5 neighbors)	0.4817	0.4948	0.3783
LocalOutlierFactor (10 neighbors)	0.4395	0.4903	0.3527
IsolationForest	0.6465	0.6095	0.5836
ABOD	0.3312	0.4933	0.3324
HBOS	0.5978	0.5067	0.3587
AutoEncoder	0.4297	0.4873	0.3534
Setting = Repetitiveness	Precision	Recall	F1-Score
OneClassSVM (RBF kernel)	0.6141	0.6095	0.6054
OneClassSVM (Polynomial kernel)	0.5082	0.5082	0.5082
OneClassSVM (Linear kernel)	0.5067	0.5067	0.5067
LocalOutlierFactor (3 neighbors)	0.6764	0.6051	0.5611
LocalOutlierFactor (5 neighbors)	0.7163	0.6036	0.5446
LocalOutlierFactor (10 neighbors)	0.7113	0.6021	0.5435
IsolationForest	0.9473	0.9419	0.9417
ABOD	0.7268	0.5618	0.4651
HBOS	0.9105	0.9031	0.9027
AutoEncoder	0.9563	0.9553	**0.9553**
Setting = Emotional Semantics	Precision	Recall	F1-Score
OneClassSVM (RBF kernel)	0.6699	0.6528	**0.6436**
OneClassSVM (Polynomial kernel)	0.4918	0.4918	0.4918
OneClassSVM (Linear kernel)	0.5291	0.5291	0.5286
LocalOutlierFactor (3 neighbors)	0.6085	0.5648	0.5168
LocalOutlierFactor (5 neighbors)	0.6065	0.5574	0.5005
LocalOutlierFactor (10 neighbors)	0.6214	0.5708	0.5215
IsolationForest	0.6516	0.6110	0.5835
ABOD	0.6968	0.5350	0.4155
HBOS	0.5674	0.5067	0.3657
AutoEncoder	0.6606	0.5768	0.5138
Setting = Readability	Precision	Recall	F1-Score
OneClassSVM (RBF kernel)	0.7526	0.5112	0.3569
OneClassSVM (Polynomial kernel)	0.8318	0.7466	0.7292
OneClassSVM (Linear kernel)	0.8324	0.7481	0.7310
LocalOutlierFactor (3 neighbors)	0.8166	0.7675	0.7582
LocalOutlierFactor (5 neighbors)	0.8408	0.8048	0.7995
LocalOutlierFactor (10 neighbors)	0.8563	0.8271	0.8236
IsolationForest	0.8943	0.8942	**0.8942**
ABOD	0.7581	0.6006	0.5291
HBOS	0.8072	0.7273	0.7085
AutoEncoder	0.8525	0.8376	0.8358
Setting = POS	Precision	Recall	F1-Score
OneClassSVM (RBF kernel)	0.8154	0.7392	**0.7223**
OneClassSVM (Polynomial kernel)	0.6065	0.5618	0.5098
OneClassSVM (Linear kernel)	0.5835	0.5812	0.5781
LocalOutlierFactor (3 neighbors)	0.7457	0.6080	0.5446
LocalOutlierFactor (5 neighbors)	0.7513	0.5991	0.5280
LocalOutlierFactor (10 neighbors)	0.7501	0.5708	0.4776
IsolationForest	0.6678	0.5887	0.5342
ABOD	0.7326	0.5693	0.4782
HBOS	0.5932	0.5127	0.3800
AutoEncoder	0.7014	0.6289	0.5925

Table 4. Spanish dataset: detection accuracy via 5-fold stratified cross-validation. Highest F1-Score in bold.

Setting = Text	**Precision**	**Recall**	**F1-Score**
OneClassSVM (RBF kernel)	0.6190	0.6043	**0.5917**
OneClassSVM (Polynomial kernel)	0.3102	0.3329	0.3123
OneClassSVM (Linear kernel)	0.3102	0.3329	0.3123
LocalOutlierFactor (3 neighbors)	0.3641	0.4600	0.3443
LocalOutlierFactor (5 neighbors)	0.3409	0.4657	0.3354
LocalOutlierFactor (10 neighbors)	0.2582	0.4629	0.3187
IsolationForest	0.2511	0.4514	0.3132
ABOD	0.2489	0.4957	0.3314
HBOS	0.2475	0.4900	0.3289
AutoEncoder	0.2426	0.4714	0.3204
Setting = Repetitiveness	Precision	Recall	F1-Score
OneClassSVM (RBF kernel)	0.5200	0.5200	**0.5199**
OneClassSVM (Polynomial kernel)	0.3988	0.4029	0.3968
OneClassSVM (Linear kernel)	0.3972	0.4014	0.3952
LocalOutlierFactor (3 neighbors)	0.5369	0.5171	0.4426
LocalOutlierFactor (5 neighbors)	0.5697	0.5257	0.4369
LocalOutlierFactor (10 neighbors)	0.5752	0.5229	0.4223
IsolationForest	0.3971	0.4886	0.3424
ABOD	0.2482	0.4929	0.3301
HBOS	0.2486	0.4943	0.3308
AutoEncoder	0.4578	0.4929	0.3599
Setting = Emotional Semantics	Precision	Recall	F1-Score
OneClassSVM (RBF kernel)	0.4898	0.4900	**0.4881**
OneClassSVM (Polynomial kernel)	0.3568	0.3614	0.3562
OneClassSVM (Linear kernel)	0.3828	0.3843	0.3824
LocalOutlierFactor (3 neighbors)	0.5305	0.5186	0.4663
LocalOutlierFactor (5 neighbors)	0.5177	0.5100	0.4501
LocalOutlierFactor (10 neighbors)	0.5387	0.5229	0.4684
IsolationForest	0.4859	0.4900	0.4499
ABOD	0.5947	0.5100	0.3689
HBOS	0.4830	0.4986	0.3496
AutoEncoder	0.5561	0.5157	0.4094
Setting = Readability	Precision	Recall	F1-Score
OneClassSVM (RBF kernel)	0.7514	0.5057	0.3459
OneClassSVM (Polynomial kernel)	0.8218	0.7357	0.7168
OneClassSVM (Linear kernel)	0.8244	0.7414	**0.7238**
LocalOutlierFactor (3 neighbors)	0.7134	0.6329	0.5946
LocalOutlierFactor (5 neighbors)	0.7055	0.6129	0.5637
LocalOutlierFactor (10 neighbors)	0.6382	0.5671	0.5033
IsolationForest	0.6904	0.6186	0.5788
ABOD	0.2486	0.4943	0.3308
HBOS	0.2482	0.4929	0.3301
AutoEncoder	0.4845	0.4971	0.3684
Setting = POS	Precision	Recall	F1-Score
OneClassSVM (RBF kernel)	0.4971	0.4971	**0.4968**
OneClassSVM (Polynomial kernel)	0.4100	0.4100	0.4100
OneClassSVM (Linear kernel)	0.3787	0.3843	0.3771
LocalOutlierFactor (3 neighbors)	0.5200	0.5029	0.3673
LocalOutlierFactor (5 neighbors)	0.5216	0.5029	0.3652
LocalOutlierFactor (10 neighbors)	0.5311	0.5043	0.3681
IsolationForest	0.2701	0.4757	0.3247
ABOD	0.2478	0.4914	0.3295
HBOS	0.2482	0.4929	0.3301
AutoEncoder	0.3171	0.4671	0.3297

The Emotional Semantics setting (see Figure 4) showcases 331 (HBOS), 312 (ABOD), 288 (LocalOutlierFactor), 284 (AutoEncoder), 261 (IsolationForest), and 233 (OneClassSVM) incorrectly classified essays. These results show that this setting does not allow for accurate essay detection for all models, as evident by the large amount of miscategorized data. The Readability setting (see Figure 5) presents 268 (ABOD), 183 (HBOS), 170 (OneClassSVM), 116 (LocalOutlierFactor), 109 (AutoEncoder), and 71 (IsolationForest) incorrectly classified essays. This setting is, overall, effective for it yields better results than the Text and Emotional Semantics settings; however, it does not outperform Repetitiveness. The POS setting (see Figure 6) shows 327 (HBOS), 289 (ABOD), 276 (IsolationForest), 263 (LocalOutlierFactor), 249 (AutoEncoder), and 175 (OneClassSVM) incorrectly classified essays. Its performance falls between Text and Readability.

For the Spanish dataset, the Text setting (see Figure 7) shows 384 (IsolationForest), 378 (LocalOutlierFactor), 370 (AutoEncoder), 357 (HBOS), 353 (ABOD), and 277 (OneClass SVM) incorrectly classified essays out of 700. These results are close to the random performance with this setting, with OneClassSVM being the best-performing method. The Repetitiveness setting (see Figure 8) highlights 358 (IsolationForest), 355 (ABOD), 355 (AutoEncoder), 354 (HBOS), 338 (LocalOutlierFactor), and 336 (OneClassSVM) incorrectly classified essays. Across all the methods in this setting, the performance is consistently poor, as indicated by the minimal deviation in the number of misclassified essays. The Emotional Semantics setting (see Figure 9) showcases 357 (IsolationForest), 357 (OneClassSVM), 351 (HBOS), 343 (ABOD), 339 (AutoEncoder), and 334 (LocalOutlierFactor) incorrectly classified essays. These results confirm that Emotional Semantics do not offer highly predictive information for the detection task, as shown in the English dataset. The Readability setting (see Figure 10) presents 355 (HBOS), 354 (ABOD), 352 (AutoEncoder), 267 (IsolationForest), 257 (LocalOutlierFactor), and 181 (OneClassSVM) incorrectly classified essays. This setting is more effective than others, as observed by performance metrics. In fact, OneClassSVM, LocalOutlierFactor, and IsolationForest are able to reach results that are above random performance. The POS setting (see Figure 11) shows 373 (AutoEncoder), 367 (IsolationForest), 356 (ABOD), 355 (HBOS), 352 (OneClassSVM), and 347 (LocalOutlierFactor) incorrectly classified essays, falling between Text and Readability. Overall, these results appear to be close to random performanc and in line with the results obtained in the Emotional Semantics, Repetitiveness, and Text settings for the English dataset.

RQ3: The results obtained from the English (Table 3) and Spanish datasets (Table 4) reveal that there is a substantial difference in model performance. Considering the best outcomes obtained within each setting across all methods for English vs. Spanish essays, the following side-by-side F1-Score comparisons are obtained: Text (0.6694 vs. 0.5917), Repetitiveness (0.9553 vs. 0.5199), Emotional Semantics (0.6436 vs. 0.4881), Readability (0.8942 vs. 0.7238), and POS (0.7223 vs. 0.4968).

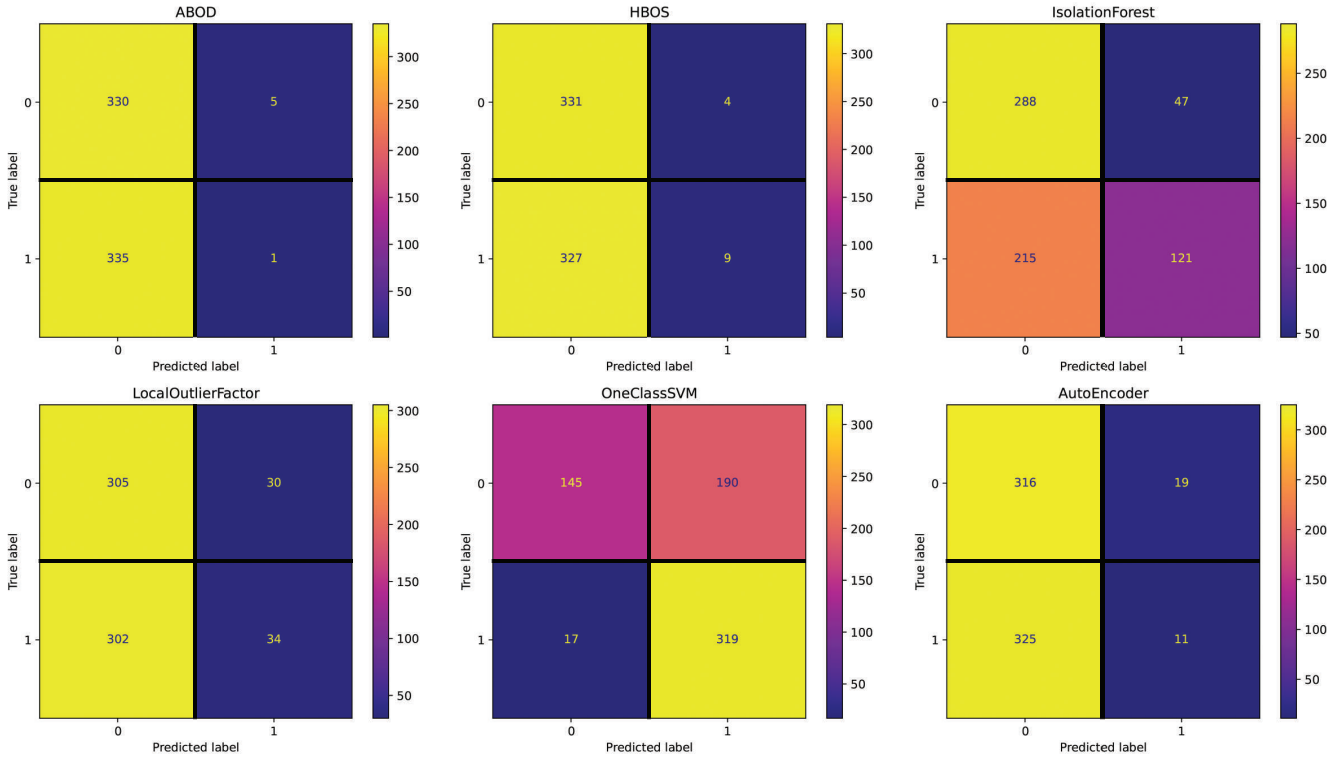

Figure 2. Confusion matrices reporting the number of correctly (main diagonal) and incorrectly (secondary diagonal) predicted essays for all methods (English dataset—setting: Text).

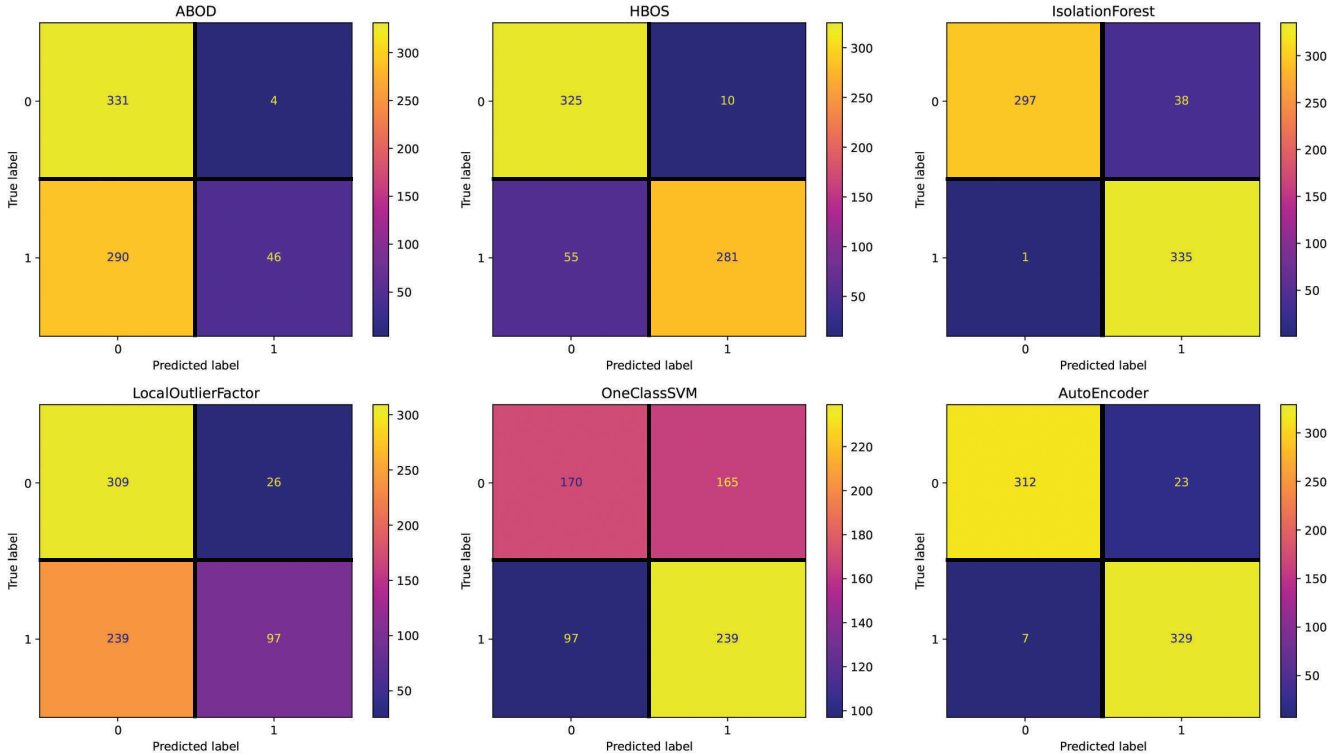

Figure 3. Confusion matrices reporting the number of correctly (main diagonal) and incorrectly (secondary diagonal) predicted essays for all methods (English dataset—setting: Repetitiveness).

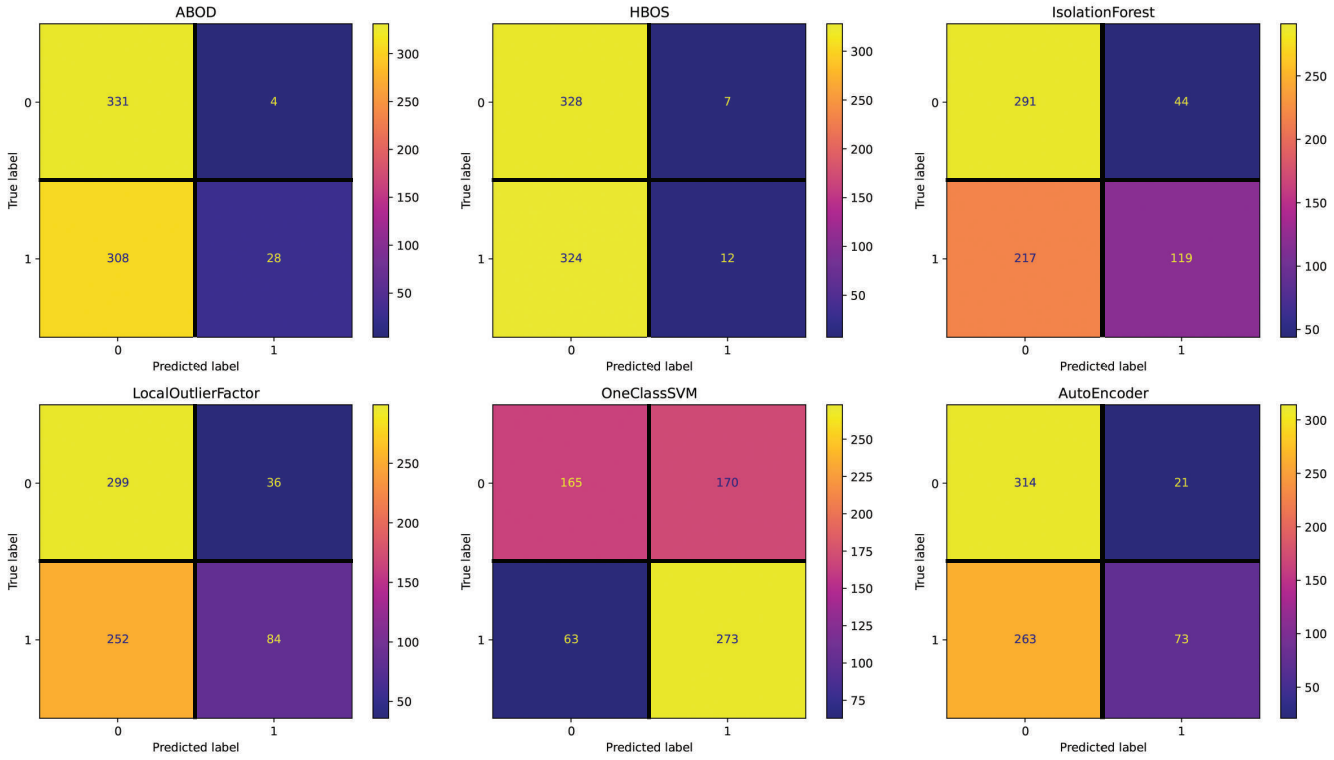

Figure 4. Confusion matrices reporting the number of correctly (main diagonal) and incorrectly (secondary diagonal) predicted essays for all methods (English dataset—setting: Emotional Semantics).

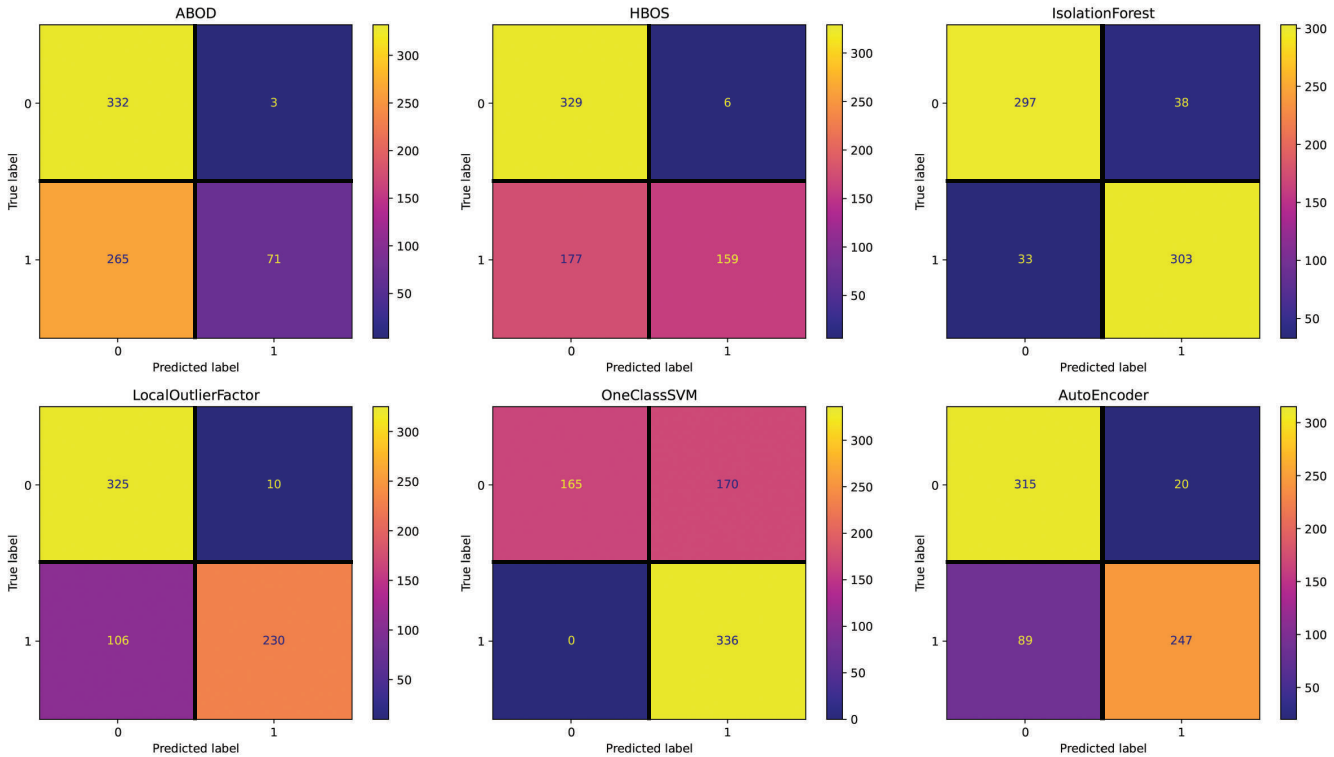

Figure 5. Confusion matrices reporting the number of correctly (main diagonal) and incorrectly (secondary diagonal) predicted essays for all methods (English dataset—setting: Readability).

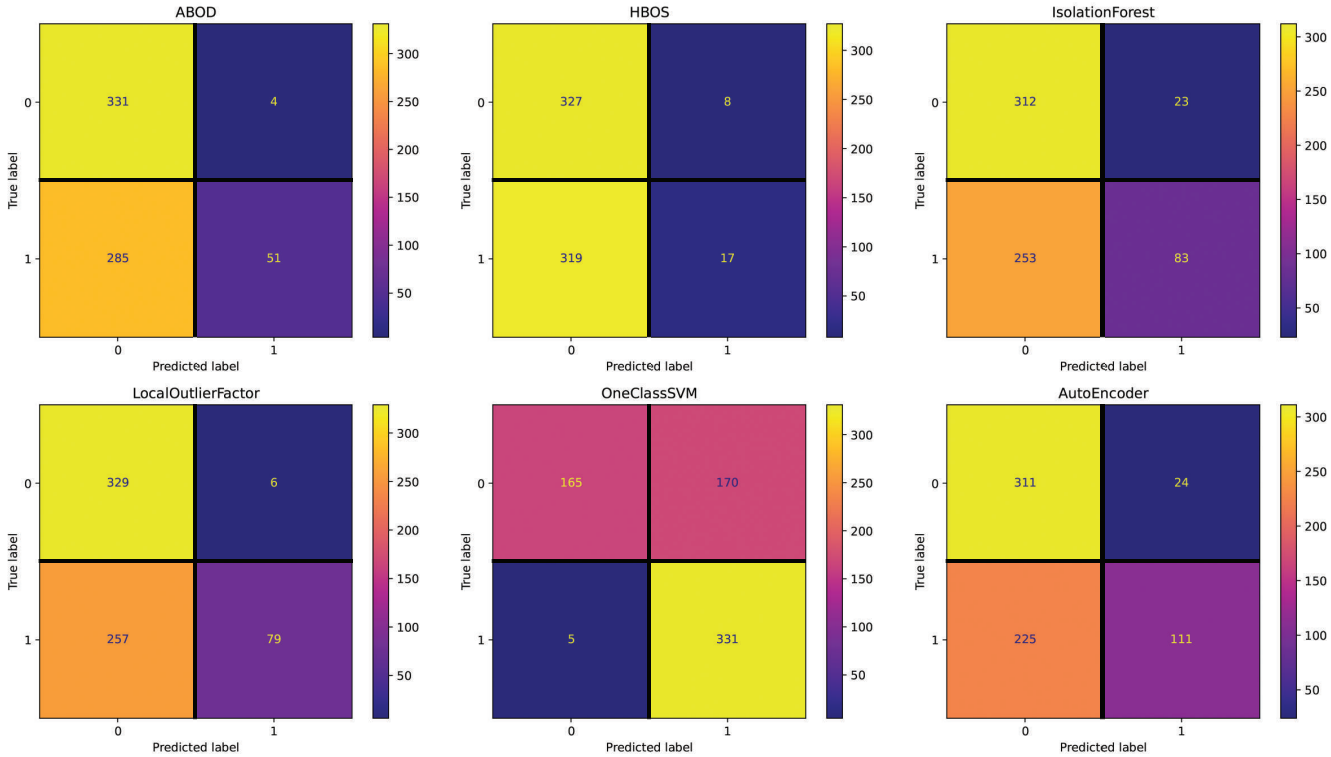

Figure 6. Confusion matrices reporting the number of correctly (main diagonal) and incorrectly (secondary diagonal) predicted essays for all methods (English dataset—setting: POS).

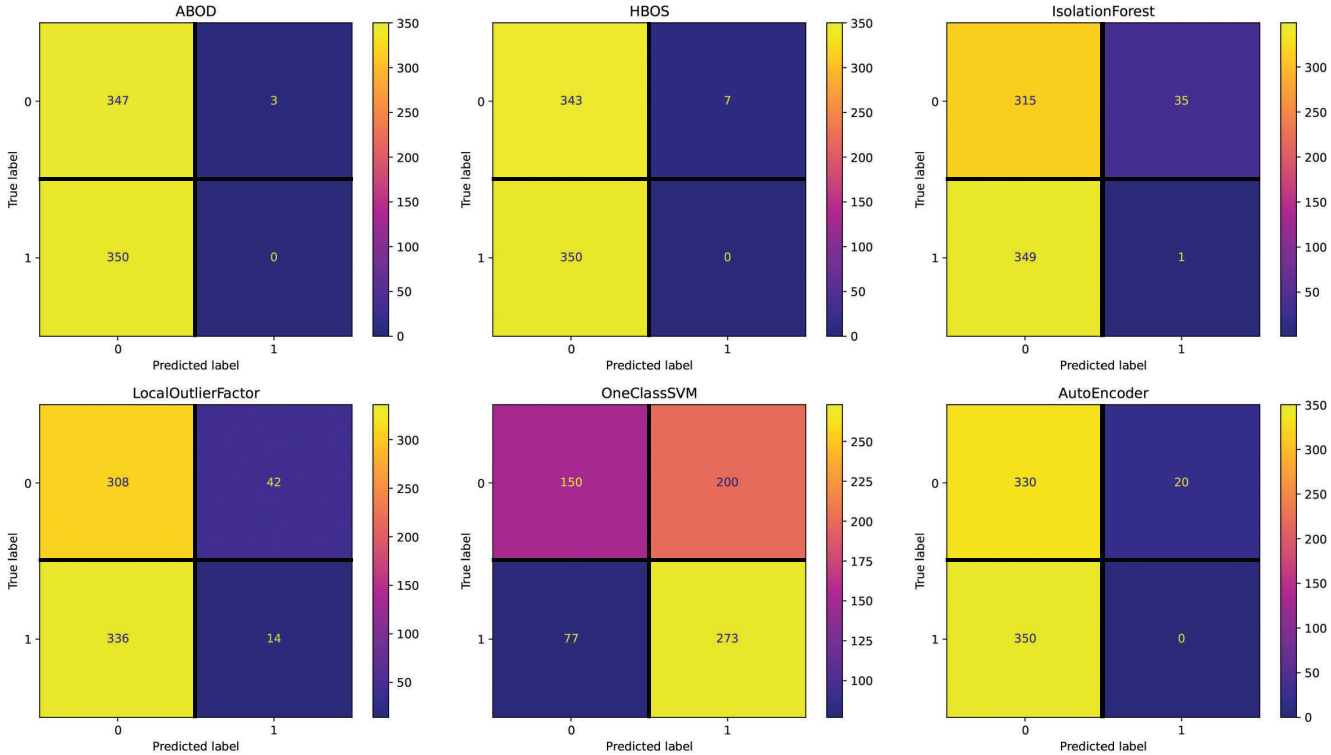

Figure 7. Confusion matrices reporting the number of correctly (main diagonal) and incorrectly (secondary diagonal) predicted essays for all methods (Spanish dataset—setting: Text).

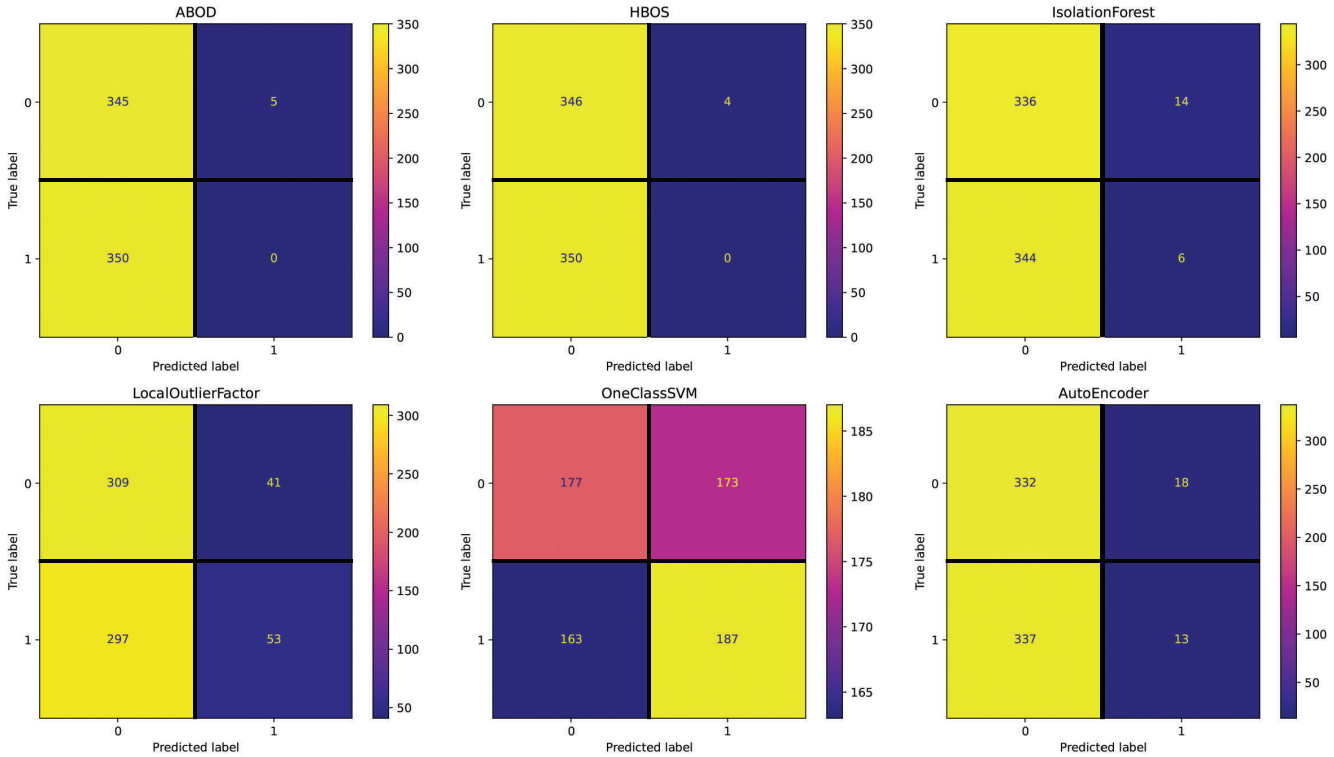

Figure 8. Confusion matrices reporting the number of correctly (main diagonal) and incorrectly (secondary diagonal) predicted essays for all methods (Spanish dataset—setting: Repetitiveness).

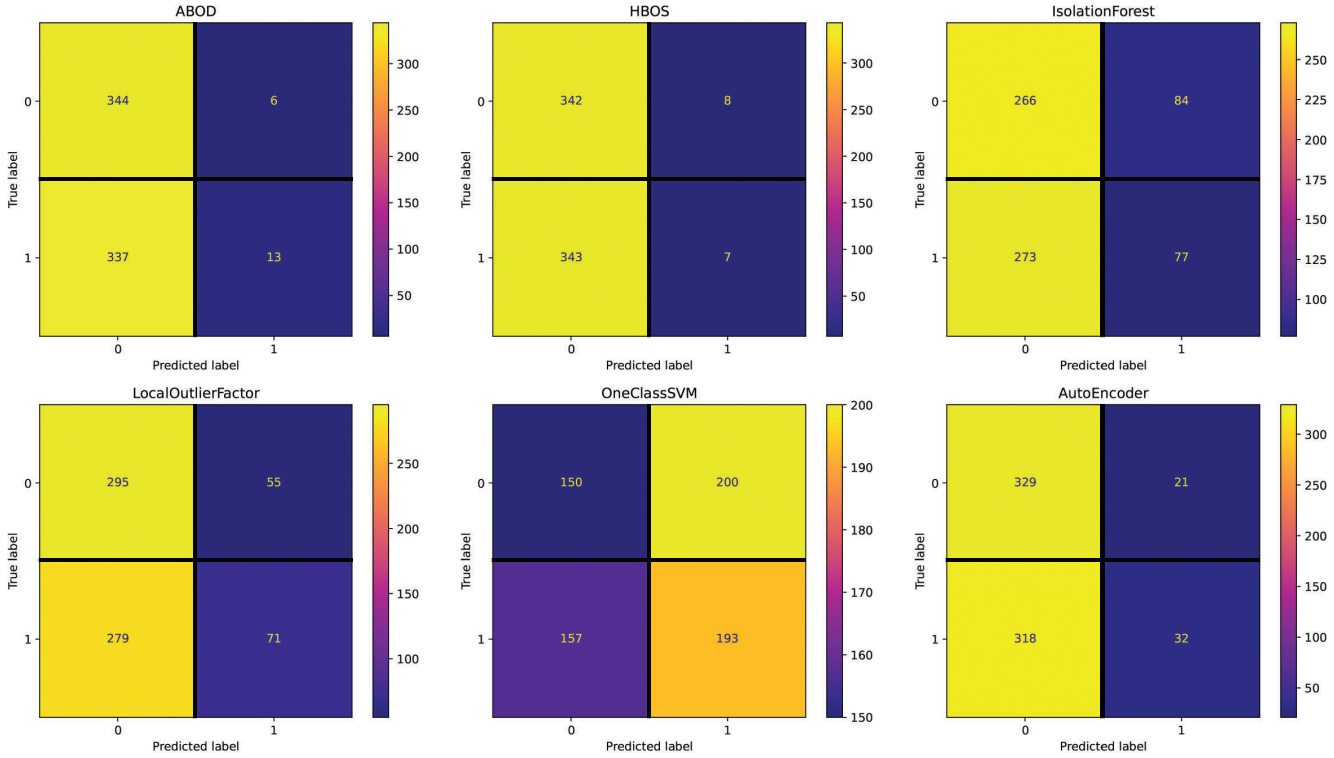

Figure 9. Confusion matrices reporting the number of correctly (main diagonal) and incorrectly (secondary diagonal) predicted essays for all methods (Spanish dataset—setting: Emotional Semantics).

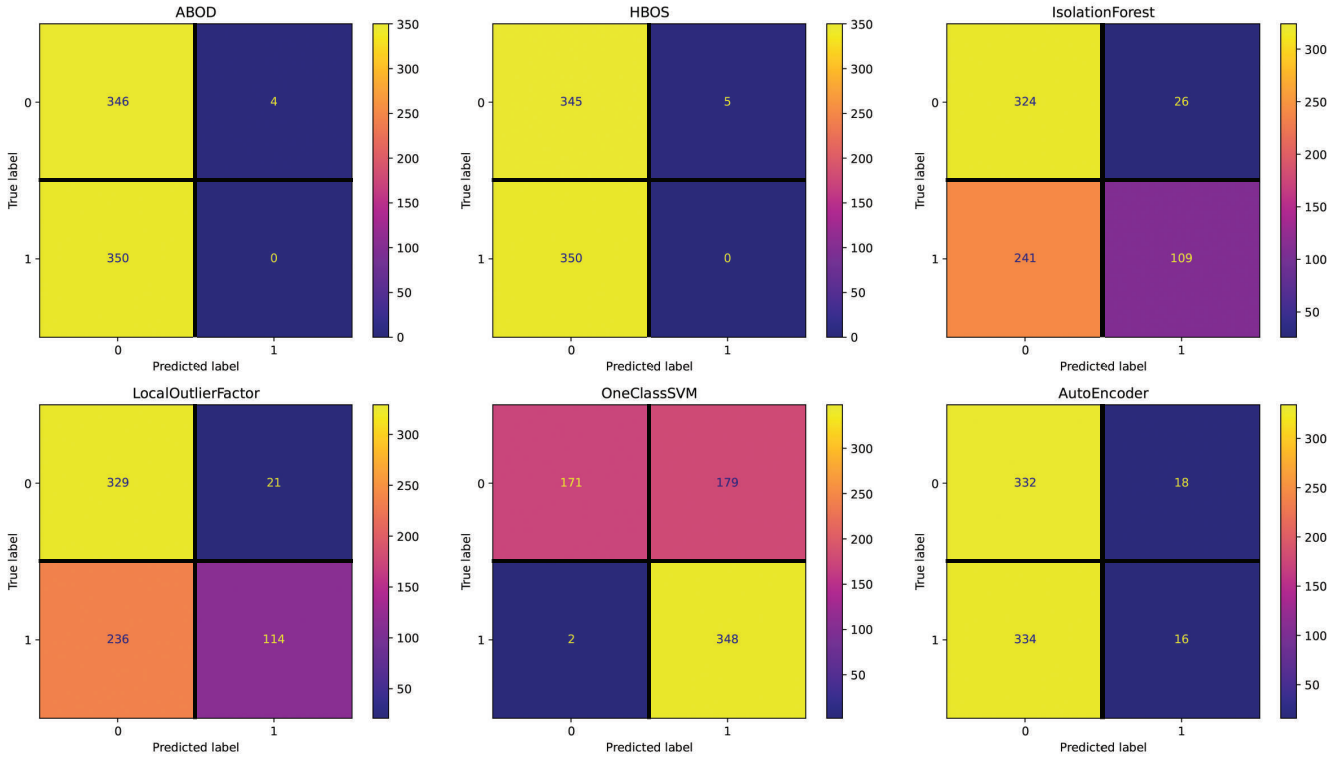

Figure 10. Confusion matrices reporting the number of correctly (main diagonal) and incorrectly (secondary diagonal) predicted essays for all methods (Spanish dataset—setting: Readability).

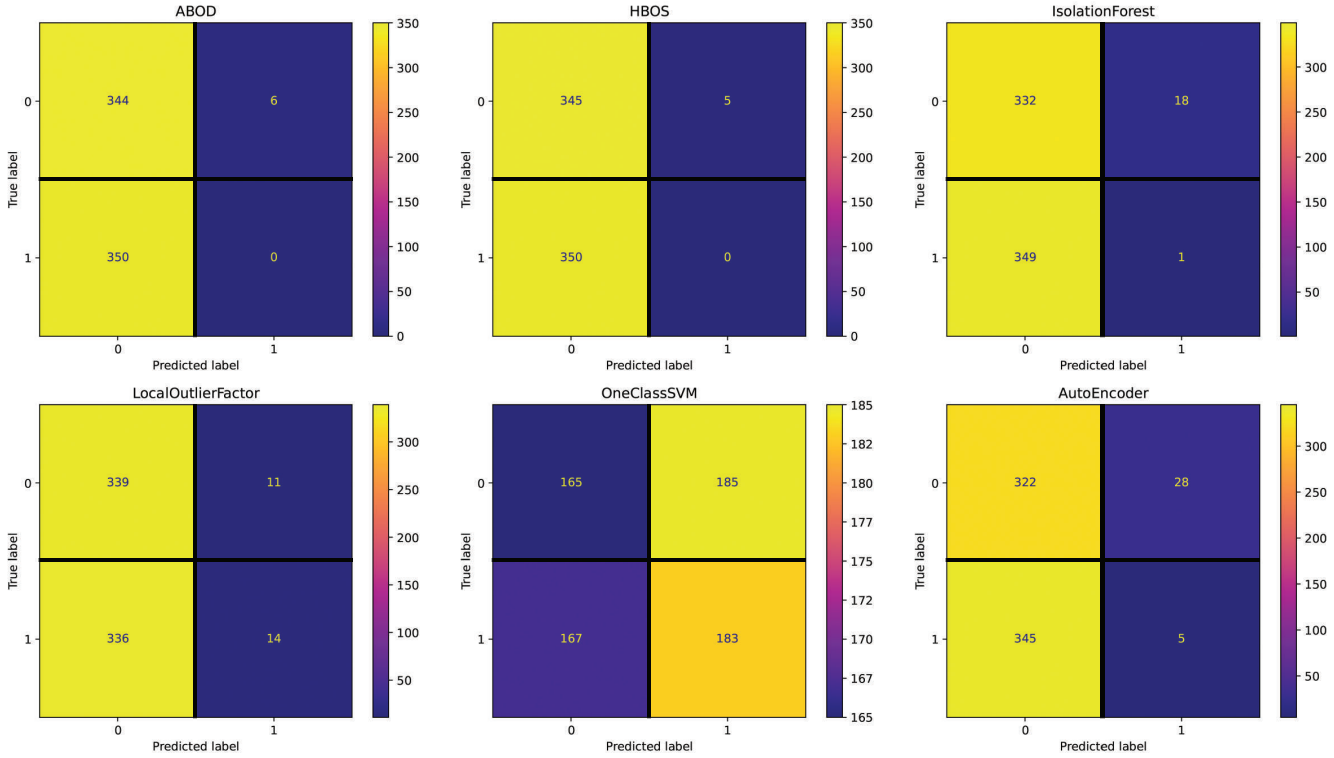

Figure 11. Confusion matrices reporting the number of correctly (main diagonal) and incorrectly (secondary diagonal) predicted essays for all methods (Spanish dataset—setting: POS).

4. Discussion

In this section, the experimental results obtained for the AI-generated detection task are discussed.

RQ1: This question pertains to the accurate detection of human vs. AI-generated essays using one-class learning models. The results showed that AutoEncoder achieved the best results on the English dataset, followed by IsolationForest. Analyzing this result from the perspective of recent research works involving one-class learning methods, a similar pattern was observed in [33], where AutoEncoder significantly outperformed other approaches in terms of both model accuracy and robustness to contamination in training data. The same behavior was observed in [31], where AutoEncoder outperformed both IsolationForest and OneClassSVM in the context of geo-distributed sensor network data, and in [29,30,32], where deep learning approaches based on neural networks such as AutoEncoder yielded remarkable detection performances. This result shows that such approaches are particularly effective when the ability to catch the non-linearities across features is beneficial for the detection task.

However, it is noteworthy that this phenomenon was not systematically observed in the literature. For instance, the study in [28] compared deep learning and classical machine learning methods and identified that classical methods outperformed deep learning approaches on the considered time series domains. The work in [34] also showcases the effectiveness of one-class ensembles of classical machine learning approaches in complex domains. This result, which seems to contradict what is observed in the present paper for the English dataset, is actually in line with the results for the Spanish dataset. On that dataset, OneClassSVM, which is regarded as a classical machine learning approach for one-class learning, yielded the best performance, outperforming other approaches including AutoEncoder. Therefore, different one-class learning methods present varying robustness and performance in the presence of different data characteristics.

Overall, the general results of the experiments are significant. Their practical implication is that it is possible to train models for the detection task that yield accurate predictions without exploiting positively labelled training data (AI-generated essays), which are usually unavailable to end users, i.e., instructors. Thus, the classification of human vs. AI-generated essays using one-class learning methods without leveraging positively labeled data during the model training process is an achievable undertaking.

RQ2: Regarding the importance of linguistic features for accurate essay detection, the results showed that Readability was on average the best-performing setting for the English and Spanish datasets. This finding contradicts the results in other previously published studies on fake news [21,49], which deemed Readability as the features with the lowest detection capabilities. Such an anomaly can be explained by the number and complexity of the features used. The authors in [21,49] used seven and four features, respectively. The current study employs thirteen features, including some that are specific to academic texts and the Spanish language. Furthermore, Readability possessed the highest F1-Score performance across all methods in Spanish and the second-highest in English. Readability refers to the complexity of the vocabulary of a text; given that essays display elaborate and specific lexicon, as opposed to that of fake news, it is not surprising for these linguistic features to rank high.

In the English data, Repetitiveness obtained the highest F1-Score. This setting was the third-highest in Spanish. As noted in [50], AI essays are repetitive, not only in the use of vocabulary but also in the expression of ideas. Essays generated by ChatGPT exhibited great repetition of ideas and phrases in the current study, particularly those generated in English.

Based on the assumption that machine-generated text presents shorter and less complex sentences, less use of adjectives and adverbs, and overall shorter sentence length [18,24], POS tagging was expected to have higher classifying power. However, the results indicated that it was the fourth-best in English and the least competitive in Spanish. While sentence length is a differentiating feature between AI- and human-generated essays, it is not

competitive enough when compared to semantic features. Overall, the results showed that linguistic features enabled one-class models to effectively classify human vs. AI-generated essays, an unexplored venue in detection studies.

RQ3: This research question addresses the differences in detection accuracy in English and Spanish essays. The results highlighted that Readability is the setting that could be considered optimal for both languages. This may be the case because of the features used in this setting. Overall, lower model performance in the Spanish dataset is observed. This phenomenon could be due to two main reasons. First, human essays were produced by L2 learners of English and Spanish, which could exhibit language transfer, a process by which learners use their knowledge of other languages and apply it to another [51]. Thus, linguistic features that highly depend on meaning—Emotional Semantics, Readability, and POS—may have been influenced by learners' inaccuracy and/or misspelling of words. Second, Spanish is a morphologically synthetic language, relying on inflection to express syntactic meaning, whereas English is analytic in nature, using words and word order to convey syntactic meaning [52]. More inaccurate use of morphological inflections was present in Spanish human-generated essays compared to the English dataset. Additionally, the use of the Oxford comma is found in certain dialects of English. Evidence of its use in Spanish has not been reported in the literature. Oxford commas in L2 Spanish essays were found since the essays were written by L1 (First Language) American English students. However, this single feature alone could not help the model discern between both text modalities. What is more, Spanish essays generated by ChatGPT exhibited the use of the Oxford comma, depicting a strong influence of the English language in ChatGPT outcomes. Thus, considering that GPT has been mostly trained on English texts (https://seo.ai/blog/how-many-languages-does-chatgpt-support (accessed on 1 July 2023)), it is likely that the class overlap between essays written by L2 learners of Spanish with English as their L1 and AI-generated essays is smaller than that observed for the English dataset, resulting in a more difficult detection task for one-class learning models with the Spanish dataset. Average sentence length is longer in Spanish than in English [53]. Considering the aforementioned influences, the complexity of the detection task using simple features in the one-class learning setting was increased. Identifying effective features and models for the Spanish language is a crucial aspect to be further investigated in future research.

5. Conclusions and Future Work

This paper attended to the gaps in the literature for essay data analysis, which often involve supervised neural-network-based and feature-based methods applied to English data. Additionally, the effectiveness of one-class learning methods, considered more suitable for detection tasks, had never been explored in the context of essay data. Previous research has predominantly concentrated on L1 texts or has overlooked the implications of not knowing if the text comes from a native speaker of the language or a speaker who uses the language as a second means of communication. The noteworthy contribution of this paper resides in the composition of the analyzed dataset, which consists of essays written in English and Spanish by L2 learners, and in addressing the AI-generated essay detection task using one-class learning models.

This paper assessed the detection performance of different models in varied settings, where positively labeled data, i.e., AI-generated essays, are unavailable for model training. Experiments with two datasets containing essays in L2 English and L2 Spanish showed that it is feasible to accurately detect AI-generated essays, especially those that are written in English. The results revealed that the AutoEncoder model for English, OneClassSVM for Spanish, and Readability as a linguistic feature are more powerful than others in the detection task.

In summary, this paper addressed the task of AI-generated essay detection, emphasizing its importance given the adoption of ChatGPT by the general public, which raised concerns with regard to academic integrity. Overall, the extracted results should guide and

stimulate researchers and practitioners in the design of new and effective methods, as well as the adoption of new linguistic features, for this detection task. To this end, the results presented and published datasets that can be used as a benchmark for future works. Future work should also focus on analyzing and improving the detection accuracy for Spanish essays. Moreover, custom one-class-learning methods as well as more sophisticated deep-learning model architectures for AI-generated essay detection that do not exclusively rely on linguistic features are to be explored.

Author Contributions: Conceptualization: R.C. and S.L.-A., methodology: R.C. and S.L.-A., software: R.C., validation: R.C. and S.L.-A., data curation: S.L.-A. and R.C., investigation: R.C. and S.L.-A., writing—original draft preparation: R.C. and S.L.-A., writing—review: R.C. and S.L.-A., editing: R.C. and S.L.-A., visualization: R.C., and supervision: R.C. 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: The code to reproduce the experiments and the datasets adopted in this paper are available at the following URL: https://github.com/rcorizzo/one-class-essay-detection (accessed on 19 May 2023).

Acknowledgments: The authors acknowledge the support of American University and of NVIDIA through the donation of a Titan V GPU.

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

References

1. Sallam, M. ChatGPT utility in healthcare education, research, and practice: Systematic review on the promising perspectives and valid concerns. *Healthcare* **2023**, *11*, 887. [CrossRef] [PubMed]
2. Lund, B.D.; Wang, T. Chatting about ChatGPT: How may AI and GPT impact academia and libraries? *Libr. Hi Tech News* **2023**, *40*, 26–29. [CrossRef]
3. King, M.R.; ChatGPT. A conversation on artificial intelligence, chatbots, and plagiarism in higher education. *Cell. Mol. Bioeng.* **2023**, *16*, 1–2. [CrossRef]
4. Slaouti, D. The World Wide Web for academic purposes: Old study skills for new? *Engl. Specif. Purp.* **2002**, *21*, 105–124. [CrossRef]
5. Stapleton, P. Writing in an electronic age: A case study of L2 composing processes. *J. Engl. Acad. Purp.* **2010**, *9*, 295–307. [CrossRef]
6. Crothers, E.; Japkowicz, N.; Viktor, H. Machine Generated Text: A Comprehensive Survey of Threat Models and Detection Methods. *arXiv* **2022**, arXiv:2210.07321.
7. Bostrom, N.; Yudkowsky, E. The ethics of artificial intelligence. In *Artificial Intelligence Safety and Security*; Chapman and Hall/CRC: Boca Raton, FL, USA, 2018; pp. 57–69.
8. Arbane, M.; Benlamri, R.; Brik, Y.; Alahmar, A.D. Social media-based COVID-19 sentiment classification model using Bi-LSTM. *Expert Syst. Appl.* **2023**, *212*, 118710. [CrossRef]
9. Li, W.; Qi, F.; Tang, M.; Yu, Z. Bidirectional LSTM with self-attention mechanism and multi-channel features for sentiment classification. *Neurocomputing* **2020**, *387*, 63–77. [CrossRef]
10. Kumari, R.; Ashok, N.; Ghosal, T.; Ekbal, A. A multitask learning approach for fake news detection: Novelty, emotion, and sentiment lend a helping hand. In Proceedings of the 2021 International Joint Conference on Neural Networks (IJCNN), Shenzhen, China, 18–22 July 2021; pp. 1–8.
11. Damasceno, L.P.; Shafer, A.; Japkowicz, N.; Cavalcante, C.C.; Boukouvalas, Z. Efficient Multivariate Data Fusion for Misinformation Detection During High Impact Events. In Proceedings of the Discovery Science: 25th International Conference, DS 2022, Montpellier, France, 10–12 October 2022; pp. 253–268.
12. Jing, Q.; Yao, D.; Fan, X.; Wang, B.; Tan, H.; Bu, X.; Bi, J. TRANSFAKE: Multi-task Transformer for Multimodal Enhanced Fake News Detection. In Proceedings of the 2021 International Joint Conference on Neural Networks (IJCNN), Shenzhen, China, 18–22 July 2021; pp. 1–8.
13. Han, H.; Ke, Z.; Nie, X.; Dai, L.; Slamu, W. Multimodal Fusion with Dual-Attention Based on Textual Double-Embedding Networks for Rumor Detection. *Appl. Sci.* **2023**, *13*, 4886. [CrossRef]
14. Prasad, N.; Saha, S.; Bhattacharyya, P. A Multimodal Classification of Noisy Hate Speech using Character Level Embedding and Attention. In Proceedings of the 2021 International Joint Conference on Neural Networks (IJCNN), Shenzhen, China, 18–22 July 2021; pp. 1–8.
15. Alghamdi, J.; Lin, Y.; Luo, S. Does Context Matter? Effective Deep Learning Approaches to Curb Fake News Dissemination on Social Media. *Appl. Sci.* **2023**, *13*, 3345. [CrossRef]

16. Allouch, M.; Mansbach, N.; Azaria, A.; Azoulay, R. Utilizing Machine Learning for Detecting Harmful Situations by Audio and Text. *Appl. Sci.* **2023**, *13*, 3927. [CrossRef]
17. Rubin, V.L.; Conroy, N.; Chen, Y.; Cornwell, S. Fake news or truth? Using satirical cues to detect potentially misleading news. In Proceedings of the Second Workshop on Computational Approaches to Deception Detection, San Diego, CA, USA, 17 June 2016; pp. 7–17.
18. Feng, L.; Jansche, M.; Huenerfauth, M.; Elhadad, N. A comparison of features for automatic readability assessment. In Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010), Beijing, China, 23–27 August 2010.
19. Argamon-Engelson, S.; Koppel, M.; Avneri, G. Style-based text categorization: What newspaper am I reading. In Proceedings of the AAAI Workshop on Text Categorization, Madison, WI, USA, 26–27 July 1998; pp. 1–4.
20. Koppel, M.; Argamon, S.; Shimoni, A.R. Automatically categorizing written texts by author gender. *Lit. Linguist. Comput.* **2002**, *17*, 401–412. [CrossRef]
21. Pérez-Rosas, V.; Kleinberg, B.; Lefevre, A.; Mihalcea, R. Automatic detection of fake news. *arXiv* **2017**, arXiv:1708.07104.
22. Holtzman, A.; Buys, J.; Du, L.; Forbes, M.; Choi, Y. The Curious Case of Neural Text Degeneration. In Proceedings of the International Conference on Learning Representations, New Orleans, LA, USA, 6–9 May 2019.
23. Ippolito, D.; Duckworth, D.; Callison-Burch, C.; Eck, D. Automatic Detection of Generated Text is Easiest when Humans are Fooled. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, 5–10 July 2020; pp. 1808–1822.
24. Fröhling, L.; Zubiaga, A. Feature-based detection of automated language models: Tackling GPT-2, GPT-3 and Grover. *Peerj Comput. Sci.* **2021**, *7*, e443. [CrossRef] [PubMed]
25. Gehrmann, S.; Harvard, S.; Strobelt, H.; Rush, A.M. GLTR: Statistical Detection and Visualization of Generated Text. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, ACL 2019, Florence, Italy, 28 July–2 August 2019; p. 111.
26. Crossley, S.A.; Allen, D.B.; McNamara, D.S. Text readability and intuitive simplification: A comparison of readability formulas. *Read. Foreign Lang.* **2011**, *23*, 84–101.
27. Corizzo, R.; Leal-Arenas, S. A Deep Fusion Model for Human vs. Machine-Generated Essay Classification. In Proceedings of the 2023 International Joint Conference on Neural Networks (IJCNN), Broadbeach, Australia, 18–23 June 2023; pp. 1–8.
28. Rewicki, F.; Denzler, J.; Niebling, J. Is It Worth It? Comparing Six Deep and Classical Methods for Unsupervised Anomaly Detection in Time Series. *Appl. Sci.* **2023**, *13*, 1778. [CrossRef]
29. Ryan, S.; Corizzo, R.; Kiringa, I.; Japkowicz, N. Pattern and anomaly localization in complex and dynamic data. In Proceedings of the 2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA), Boca Raton, FL, USA, 16–19 December 2019; pp. 1756–1763.
30. Lian, Y.; Geng, Y.; Tian, T. Anomaly Detection Method for Multivariate Time Series Data of Oil and Gas Stations Based on Digital Twin and MTAD-GAN. *Appl. Sci.* **2023**, *13*, 1891. [CrossRef]
31. Corizzo, R.; Ceci, M.; Pio, G.; Mignone, P.; Japkowicz, N. Spatially-aware autoencoders for detecting contextual anomalies in geo-distributed data. In Proceedings of the Discovery Science: 24th International Conference, DS 2021, Halifax, NS, Canada, 11–13 October 2021; Springer: Berlin, Germany, 2021; Volume 24, pp. 461–471.
32. Herskind Sejr, J.; Christiansen, T.; Dvinge, N.; Hougesen, D.; Schneider-Kamp, P.; Zimek, A. Outlier detection with explanations on music streaming data: A case study with danmark music group ltd. *Appl. Sci.* **2021**, *11*, 2270. [CrossRef]
33. Faber, K.; Corizzo, R.; Sniezynski, B.; Japkowicz, N. Active Lifelong Anomaly Detection with Experience Replay. In Proceedings of the 2022 IEEE 9th International Conference on Data Science and Advanced Analytics (DSAA), Shenzhen, China, 13–16 October 2022; pp. 1–10.
34. Kaufmann, J.; Asalone, K.; Corizzo, R.; Saldanha, C.; Bracht, J.; Japkowicz, N. One-class ensembles for rare genomic sequences identification. In Proceedings of the Discovery Science: 23rd International Conference, DS 2020, Thessaloniki, Greece, 19–21 October 2020; Springer, 2020; Volume 23, pp. 340–354.
35. Baly, R.; Karadzhov, G.; Alexandrov, D.; Glass, J.; Nakov, P. Predicting factuality of reporting and bias of news media sources. *arXiv* **2018**, arXiv:1810.01765.
36. Horne, B.D.; Adali, S. This just in: Fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news. In Proceedings of the Eleventh International AAAI Conference on Web and Social Media, Montreal, QC, Canada, 15–18 May 2017.
37. Hube, C.; Fetahu, B. Detecting biased statements in wikipedia. In Proceedings of the Companion Proceedings of the Web Conference, Lyon, France, 23–27 April 2018; pp. 1779–1786.
38. Moroney, C.; Crothers, E.; Mittal, S.; Joshi, A.; Adalı, T.; Mallinson, C.; Japkowicz, N.; Boukouvalas, Z. The case for latent variable vs deep learning methods in misinformation detection: An application to covid-19. In Proceedings of the Discovery Science: 24th International Conference, DS 2021, Halifax, NS, Canada, 11–13 October 2021; Springer: Berlin, Germany, 2021; Volume 24, pp. 422–432.
39. Wang, W.; Yu, Y.; Sheng, J. Image retrieval by emotional semantics: A study of emotional space and feature extraction. In Proceedings of the 2006 IEEE International Conference on Systems, Man and Cybernetics, Taipei, Taiwan, 8–11 October 2006; Volume 4, pp. 3534–3539.
40. Vosoughi, S.; Roy, D.; Aral, S. The spread of true and false news online. *Science* **2018**, *359*, 1146–1151. [CrossRef]

41. Bonta, V.; Janardhan, N.K.N. A comprehensive study on lexicon based approaches for sentiment analysis. *Asian J. Comput. Sci. Technol.* **2019**, *8*, 1–6. [CrossRef]
42. Voutilainen, A. Part-of-speech tagging. In *The Oxford Handbook of Computational Linguistics*; Oxford University Press: Oxford, UK, 2003; pp. 219–232.
43. Schölkopf, B.; Williamson, R.C.; Smola, A.J.; Shawe-Taylor, J.; Platt, J.C. Support vector method for novelty detection. In *Advances in Neural Information Processing Systems*; MIT Press: Cambridge, MA, USA, 2000; pp. 582–588.
44. Breunig, M.M.; Kriegel, H.P.; Ng, R.T.; Sander, J. LOF: Identifying density-based local outliers. In Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data, Dallas, TX, USA, 15–18 May 2000; pp. 93–104.
45. Liu, F.T.; Ting, K.M.; Zhou, Z.H. Isolation forest. In Proceedings of the 2008 Eighth IEEE International Conference on Data Mining, Pisa, Italy, 15–19 December 2008; pp. 413–422.
46. Kriegel, H.; Schubert, M.; Zimek, A. Angle-based outlier detection in high-dimensional data. In Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Las Vegas, NV, USA, 24–27 August 2008; pp. 444–452.
47. Pham, N.; Pagh, R. A near-linear time approximation algorithm for angle-based outlier detection in high-dimensional data. In Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Beijing, China, 12–16 August 2012; pp. 877–885.
48. Goldstein, M.; Score, A.D.H.b.O. A fast Unsupervised Anomaly Detection Algorithm. In Proceedings of the KI-2012: Poster and Demo Track, 35th German Conference on Artificial Intelligence, Saarbrücken, Germany, 24–27 September 2012; pp. 59–63.
49. Choudhary, A.; Arora, A. Linguistic feature based learning model for fake news detection and classification. *Expert Systems with Applications* **2021**, *169*, 114171. [CrossRef]
50. Zhu, T. From Textual Experiments to Experimental Texts: Expressive Repetition in "Artificial Intelligence Literature". *arXiv* **2022**, arXiv:2201.02303.
51. Selinker, L. Language transfer. *Gen. Linguist.* **1969**, *9*, 67.
52. Haspelmath, M.; Michaelis, S.M. Analytic and synthetic. In Proceedings of the Language Variation-European Perspectives VI: Selected Papers from the Eighth International Conference on Language Variation in Europe (ICLaVE 8), Leipzig, Germany, 27–29 May 2017; pp. 3–22.
53. Filippova, K. Multi-sentence compression: Finding shortest paths in word graphs. In Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010), Beijing, China, 23–27 August 2010; pp. 322–330.