EDUR • Educação em Revista. 20244039:e39773
DOI: http://dx.doi.org/10.1590/0102-469839773t
Preprint: https://doi.org/10.1590/SciELOPreprints.3825

ARTICLE

APPLICATION OF ARTIFICIAL INTELLIGENCE TECHNIQUES FOR CLASSIFICATION OF ESCAPE FROM THE TOPIC IN ESSAYS

CINTIA MARIA DE ARAÚJO PINHO[1]
ORCID: https://orcid.org/0000-0003-0525-5072
<cintia.pinho01@gmail.com>
MARCOS ANTONIO GASPAR[2]
ORCID: https://orcid.org/0000-0002-2422-2455
<marcos.antonio@uni9.pro.br>
RENATO JOSÉ SASSI[2]
ORCID: https://orcid.org/0000-0001-5276-4895
<sassi@uni9.pro.br>

[1] Centro Estadual de Educação Tecnológica Paula Souza. São Paulo (SP), Brazil.
[2] Universidade Nove de Julho. São Paulo (SP), Brazil.

ABSTRACT: The process of manual correction of essays causes some difficulties, among which we point out the time spent for correction and feedback to the student. For institutions such as elementary schools, universities, and the National High School Exam in Brazil (Enem), such activity demands time and cost for the evaluation of the texts produced. Going off-topic is one of the items evaluated in the Enem essay that can nullify the whole essay produced by the candidate. In this context, the automatic analysis of essays with the application of techniques and methods of Natural Language Processing, Text Mining, and other Artificial Intelligence techniques has shown to be promising in the process of automated evaluation of written language. The goal of this research is to compare different AI techniques for the classification of going off-topic in texts and identify the one with the best result to enable a smart correction system for essays. Therefore, computer experiments were carried out to classify these texts to normalize, identify patterns, and classify the essays in 1,320 Brazilian Portuguese essays on 119 different topics. The results indicate that the Convolutional Neural Network classifier obtained greater gain concerning the other classifiers analyzed, both in accuracy and about the results of false positives, the precision of metrics, Recall, and F1-Score. In conclusion, the solution validated in this research contributes to positively impacting the work of teachers and educational institutions, by reducing the time and costs associated with the essay evaluation process.

Keywords: essays, automatic essay evaluation, escape from the topic, artificial intelligence.

APLICAÇÃO DE TÉCNICAS DE INTELIGÊNCIA ARTIFICIAL PARA CLASSIFICAÇÃO DE FUGA AO TEMA EM REDAÇÕES[1][2]

RESUMO: O processo de correção manual de redações acarreta algumas dificuldades, dentre as quais se apontam o tempo dispendido para a correção e a devolutiva de resposta ao aluno. Para instituições como escolas de ensino básico e fundamental, universidades e o Exame Nacional do Ensino Médio (Enem), tal atividade demanda tempo e custo para a avaliação dos textos produzidos. A fuga ao tema é um dos itens avaliados na redação do Enem que pode anular a redação produzida pelo candidato. Neste contexto, a análise automática de redações com a aplicação de técnicas e métodos de Processamento de Linguagem Natural, Mineração de Textos e outras técnicas de Inteligência Artificial tem-se revelado promissora no processo de avaliação automatizada da linguagem escrita. O objetivo desta pesquisa é comparar diferentes técnicas de Inteligência Artificial para classificação de fuga ao tema em textos e identificar aquela com melhor resultado para viabilizar um sistema de correção inteligente de redações. Para tanto, foram executados experimentos computacionais em 1.320 redações elaboradas em língua portuguesa visando a classificação desses textos para normalizar, identificar padrões e categorizar as redações em 119 temas diferentes. Os resultados indicam que o classificador Rede Neural Convolucional obteve maior ganho em relação aos demais classificadores analisados, tanto em acurácia quanto em relação aos resultados de falsos positivos, métricas de precisão, *Recall* e *F1-Score*. Como conclusão, a solução validada nesta pesquisa contribui para impactar positivamente o trabalho de professores e instituições de ensino, por meio da redução de tempo e custos associados ao processo de avaliação de redações.

Palavras-chave: redações, avaliação automática de redações, fuga ao tema, inteligência artificial.

APLICACIÓN DE TÉCNICAS DE INTELIGENCIA ARTIFICIAL PARA CLASIFICACIÓN DE ESCAPE DE LA SUJECIÓN EN ENSAYOS

RESUMEN: El proceso de corrección manual de ensayos presenta dificultades como el tiempo dedicado a la corrección y devolución al alumno. Para las escuelas, las universidades y el Examen Nacional de Enseñanza Secundaria en Brasil (Enem), tal actividad demanda tiempo y costo para la evaluación de los textos producidos. La evasión del tema es uno de los elementos evaluados en la redacción del Enem que puede anular el ensayo. El análisis automático de ensayos con la aplicación de técnicas y métodos de Procesamiento del Lenguaje Natural, Minería de Texto y otras técnicas de Inteligencia Artificial se ha mostrado prometedor en el proceso de evaluación automatizada del lenguaje escrito. El objetivo de esta investigación es comparar diferentes técnicas de Inteligencia Artificial para la clasificación de evasión del tema en textos e identificar aquella con mejor resultado para habilitar un sistema inteligente de corrección de ensayos. Por lo tanto, se llevaron a cabo experimentos computacionales para clasificar estos textos con el fin de normalizar, identificar patrones y clasificar los ensayos en 1.320 ensayos en lengua portuguesa en 119 temas diferentes. Los resultados indican que el clasificador Red Neuronal Convolucional obtuvo mayor gano con relación a los demás clasificadores analizados, tanto en precisión como en relación con los resultados de falsos positivos, métricas de precisión, Recall e F1-Score. La solución validada en esta investigación contribuye a impactar positivamente el trabajo de los docentes y las instituciones educativas, al reducir el tiempo y los costos asociados al proceso de evaluación de ensayos.

Palabras clave: ensayos, evaluación automática de ensayos, escape de tema, inteligencia artificial.

---

[1] The translation of this article into English was funded by Conselho Nacional de Desenvolvimento Científico e Tecnológico - CNPq/Brasil.

[2] The Editors participating in the open peer review process: Suzana dos Santos Gomes e Eucidio Pimenta Arruda

INTRODUCTION

Writing is a practice of great importance, whether in the academic, corporate world, or even in social life. According to Gomes (2020), writing is a skill to be developed, essential for the development of human cognition. In addition to being part of the individual's social activities and professional context, writing is part of their personal growth and teaching-learning process. The ability to communicate through writing remains paramount.

For students who want to enter higher education, writing a good essay can facilitate this process (Squarisi; Salvador, 2020). However, the development of writing in students is still a challenge. In 2019, Decree number 9,765 was instituted, relating to the National Literacy Policy (PNA- *Política Nacional de Alfabetização*), which highlights that "[...] education is a central concern of nations in the 21st century [...] The results obtained by Brazil in international assessments and the national indicators reveal a serious problem in teaching and learning reading and writing" (Brasil, 2019, p. 5). Therefore, this decree emphasizes the need to implement better conditions for teaching and learning reading and writing skills throughout the country.

In this sense, the teacher plays an important role in developing skills related to reading and writing. Lesme (2021) interviewed an Enem evaluator which was applied in 2021, who corrected around one hundred to 150 essays daily, taking three to five minutes to correct each essay. Starlles (2022), when interviewing another evaluator, found that she took an average of one and a half minutes to correct each essay and that she only reached that time after years of experience and carrying out repetitive work.

In the classroom, these times can be higher, as the analysis carried out by Universia, with the participation of the Training and Qualification Center for the National Secondary Education Examination (Enem- *Exame Nacional do Ensino Médio*), highlighted that a teacher can take from forty seconds to ten minutes to correct an essay, according to the quality of the writing presented in the text (Universia, 2015).

To give a practical example of the volume of data presented in the previous paragraphs, if a Portuguese language teacher evaluates an essay for each student out of a total of 500 students, it may take him eight to eighty-three hours to correct the essays which time will vary according to his/her experience. This estimate considers that the professional spends, respectively, between one and ten minutes to evaluate each essay.

The above scenario shows that teachers experience difficulties when individually evaluating texts from different students. A study carried out by Riolfi and Igreja (2010) points out that "[...] teachers dedicate only 6% of their time in the classroom to teaching writing and that, in some cases, after correcting the texts, the teachers made comments orally on the essays, ignoring other individual textual problems in their collective presentation to the class" (Riolfi; Igreja, 2010, p. 321).

In research conducted by Pinho et al. (2020), more than 60% of the teachers analyzed used less than 25% of their time for teaching and returning essays. Given this reality, it is important to develop solutions that make it possible to optimize the teacher's work by reducing the time and costs associated with the text evaluation process. Such contributions would save resources, which could be used for other teacher activities, such as preparing classes, planning content, updating skills, and being available to answer students' questions, among other possibilities.

In the case of the student, it should be noted that the knowledge absorbed during primary and secondary education is assessed in the Enem, as this exam measures students' performance at the end of secondary education. Such assessment is one of the main ways of entering higher education in Brazil. In 2019, "[...] ENEM was responsible for 32% of selections in selection processes for entry into higher education courses" (Tokarnia, 2019, n.p.).

The National Institute of Educational Studies and Research Anísio Teixeira (Inep- *O Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira*) disclosed how the essays are evaluated. The texts are corrected by more than 5 thousand evaluators, who correct 150 texts every three days. For every 50 essays, the evaluator receives two of them already evaluated by another specialist (Inep, 2022, n.p.). Thus, each essay is corrected by two teachers who are unaware of the grade given by the other, in addition to not knowing who the evaluated candidate is (Inep, 2020). When there is a discrepancy in the scores given by the two evaluators, a third evaluator is called. This process occurs when there is a difference of more

than one hundred points in the total grade awarded or if there is a difference of 80 points per competency (Inep, 2022).

This manual process takes a long time to disseminate results to candidates, in addition to consuming resources to maintain this methodology. To give an idea of the scale of this process, in 2023 there were 3,476,226 registrations for the Enem, which demonstrates the scale and complexity of the essay evaluation system implemented in each edition of the exam.

This essay is the only discursive item in the Enem test, whose text produced by the student is evaluated according to five skills. Each competency evaluated counts to 200 points for the candidate who wrote the essay, so that, if he/she masters the five competencies, he/she can score a maximum of one thousand points as a grade awarded for the essay (Inep, 2022).

However, there are unsatisfactory results in student performance in this type of exam. When evaluating the Enem 2022 microdata made available by Inep at the beginning of 2023 (Inep, 2023), 2,355,395 candidates were present every day. It was identified that approximately 129 thousand students scored zero in the Enem essay and only 32 individuals achieved the maximum grade in their textual production. According to the database analyzed, the main reasons that led to the award of a zero score were blank writing (43.9%), avoiding the topic (24.6%), and copying the motivating text (20.8%). Thus, once the candidate prepares the essay, it appears that the biggest reason for awarding a minimum grade is related to avoiding the proposed topic. This was corroborated by a study carried out by Diana (2021), which highlights this problem in the preparation of essays, which results in the student receiving a very low grade or even canceling the Enem test.

Evading the topic is related to competency 2 established by Enem and occurs when the candidate writes a text that does not contain any reference to the thematic phrase indicated in the proposition established for the essay. Therefore, if the candidate deviates from the established proposal, the evaluator does not need to continue with the correction of the writing.

In this context, this work deals exclusively with writing deviations that touch on competency 2 established by Enem, that is, avoiding the topic. The set of circumstances reported previously highlights the importance of studying a way of helping teachers speed up the process of evaluating discursive texts. To this end, an alternative is to provide support for the creation of an intelligent essay correction system. The idea of such a solution is not to completely replace the teacher's work in evaluating texts but to make the evaluation process more agile by providing indications and notes of possible flaws in the student's writing due to escaping from the proposed topic.

Despite the possible positive contributions of applying Artificial Intelligence solutions in teaching activities, it is also necessary to indicate aspects that establish points of attention regarding this phenomenon. In this sense, Pinto (2005) warns about the non-neutrality of technology in its applications and, consequently, its impact on individuals, professionals, organizations, and society. A study carried out by Silveira and Barros (2021) regarding the skills of higher education teachers indicated that 34.09% of the skills analyzed could in the future be absorbed by Artificial Intelligence at some point in the development of the technologies associated with it. Therefore, this scenario shows future consequences for the teacher's activities that should not be ignored in the context of the application of Artificial Intelligence in the teaching profession.

However, in the specific context of this study, the use of digital and intelligent tools can bring significant improvements to the process of correcting dissertation texts. The implementation of solutions based on Artificial Intelligence can contribute to this demand such as Natural Language Processing and Text Mining, in addition to different text classification techniques.

Therefore, this research aims to compare different Artificial Intelligence techniques for classifying missing topics in texts and identifying those that bring better results to enable an intelligent essay correction system to support teachers. To achieve this objective, a comparative study was carried out on different Artificial Intelligence techniques aimed at supporting the teaching-learning process, seeking to carry out experiments to indicate which techniques achieve better results.

In addition to the comparison between the classifiers, this research also presents the results of experiments already applied on a web platform called CRIA (Artificial Intelligence Writing Corrector-*Corretor de Redações por Inteligência Artificial*). The purpose of this platform is to provide instant analysis of essays simulating Enem guidelines and grades, indicating deviations made in the text. Educational

institutions can make the CRIA platform available to their students to encourage teacher-student interaction.

When the students submit their essays to the platform, they obtain the grade instantly and then receive detailed corrections with specific indications, including the indication of escape from the topic, which shows the probability of escape occurring and marking in the text of words that adhere to the essay topic proposal. By correcting the points indicated by the CRIA platform, the student will learn from their mistakes. At the end of this process, the student sends the essay to the teacher responsible for the Portuguese language curricular component. This teacher will analyze the corrections suggested by the Artificial Intelligence solution and the changes made by the student, being able to change the deviations highlighted and the grades awarded, or even add the indication of more errors in the writing. The proposed platform aims to reduce work overload and optimize the management of teacher classes.

RESEARCH ON SMART SOLUTIONS FOR WRITING ANALYSIS

The application of information technology to education has sparked research to assist teachers in the process of correcting and identifying problems in students' learning of textual production. In the automatic analysis of textual cohesion in essays, a study carried out by Nobre and Pellegrino (2010) automatically identified cohesion problems in 90% of the argumentative and dissertation texts analyzed in the experiment conducted by the authors. The results of the automated solution applied in the experiment were compatible with the grades given in corrections made by human evaluators. The authors also state that the correction carried out by a computer program does not suffer from certain external interferences, such as fatigue and mood changes, always allowing for equal assessment and analysis. However, an important counterpoint to be highlighted is that Artificial Intelligence solutions can eventually carry biases and prejudices in their structure. Even so, the automated process reduces the workload of the human evaluator and proves to be a tool to support the correction process carried out by human evaluators.

Santos Júnior (2017) developed research to improve the quality of automatic evaluation of dissertation texts by applying Natural Language Processing and Artificial Neural Networks. In his experiment, the author sought to generically address deviations in essays, without specifically evaluating each assessment skill. The applied neural network should correctly determine the score from zero to one thousand to be assigned to the essay. To this end, 18 essay themes were evaluated, indicating the results of each theme having been generated separately. The best result achieved in this experiment gave grades to the essays with an error rate of plus or minus one hundred points.

Cândido and Webber (2018) describe the possibilities of assertively treating the coherence and cohesion of essays using natural language processing tools. The study they carried out used linguistic elements and computational techniques to evaluate essays. The experiments carried out compared the analysis carried out by software and the evaluations made by two human experts. Convergent results were found in 70% of the cases analyzed. It is considered that such initial results are promising for the development of solutions for the automatic evaluation of essays, opening new research possibilities.

Passero (2018) developed a project to detect topic leaks in newsrooms by applying natural language processing and machine learning techniques. The author implemented leak detection models considering textual analysis techniques, using textual semantic similarity and some techniques, such as linear regression and support vector machines. In their experiments, 2,151 cases of essays without straying from the topic were used, in addition to 12 examples of essays that were out of topic. The best results of this experiment were obtained using the KFF-A algorithm, with average accuracy between 81.13% and 96.76%. The author reports that his solution still has a false positive rate of 4.24%, which detects that the writing had an escape from the topic, when in fact it did not. In this case, the presence of the human evaluator would still be essential.

Ramisch (2020) specifically investigated the recurrence of syntactic deviations in essays and possible correlations with certain linguistic attributes of the sentences prepared. However, in his research, essays that were canceled or strayed from the topic were eliminated. The best results obtained by the test corpus were with the Logistic Regression algorithm, which achieved 75.62% accuracy.

Riedo (2020) developed an instrument based on Artificial Intelligence techniques and procedures for the qualitative evaluation of written productions in distance education. His solution considered the relationship between the concepts elaborated, and, due to variations in the human way of expressing in writing, those not initially present in the conceptual basis of assessment could be "learned" to expand the base. As the main contribution of this study, the solution developed was capable of qualitatively discriminating written productions.

Bittencourt Júnior (2020) proposed the automatic evaluation of essays using deep neural networks. For the learning process, a set of essays with 18 different themes was used. The study sought to evaluate the five competencies stipulated by Enem. As result, the proposal of a new multi-theme architecture is proposed, based on the hypothesis that the characteristics learned by the network applied to correct a certain theme could help improve the performance of other themes, improving the results obtained in each automated assessment.

The research indicated so far has brought advances in the application of Artificial Intelligence techniques for automated text correction. This research, in addition to applying different classification techniques to detect topic leaks in newsrooms, also seeks to compare different classification algorithms, in addition to establishing a relationship with an Artificial Neural Network. Such comparisons were important for understanding that there is often no need to use a neural network to solve all the prediction problems inherent to text classification. In this sense, what expresses the effective need to use a neural network is when the amount of data no longer converges in a way equivalent to the simplest Machine Learning algorithms. Thus, in this research, it could be seen that these algorithms converged in a very similar way, that is, demonstrating that it is not necessary to spend the computational expense required by a neural network, both concerning training and application of the generated models.

ARTIFICIAL INTELLIGENCE APPLIED TO EDUCATION

Artificial Intelligence is focused on the field of knowledge associated with language, intelligence, reasoning, learning, and problem-solving. Preuss, Barone, and Henriques (2020) argue that Artificial Intelligence techniques applied to different areas of study bring numerous benefits. For Russo (2020) and Ludermir (2021), such techniques can solve increasingly complex problems, bringing efficiency, meaning and agility.

According to Müller (2018), the application of Artificial Intelligence in education has been widely discussed, although it serves a limited number of learning scenarios, as intelligent machines operate within the limits of their system. Therefore, intelligent systems applied to education must provide support for teachers and improve their work. In addition, Müller (2018) argues that Artificial Intelligence research in education is promising, while machines are adjusting to the individual needs of each professional.

Artificial Intelligence uses different techniques to provide information based on large volumes of data. Text Mining, Natural Language Processing, and Machine Learning stand out. The latter encompasses intelligent classification techniques, with emphasis on artificial neural networks (Eggers; Schatsky; Viechnicki, 2017; Hariri; Fredericks; Bowers, 2019). These techniques will be discussed in the next topics to compose a brief overview of the set of methods mentioned.

Text mining (TM)

Morais and Ambrósio (2007) define TM as a knowledge discovery process that uses data analysis and extraction techniques from texts, sentences, or just words. In the view of Gonçalves (2012) and Souza (2019), TM involves the application of computational algorithms that process texts and identify useful and implicit information, which normally could not be retrieved using traditional query methods, as the information contained in these texts cannot be obtained directly.

Natural Language Processing

Natural Language Processing is a subarea of Artificial Intelligence that studies human communication using computational methods. Thus, the aim is to convert human natural language into

a formal representation, so that it becomes more easily manipulated by machines. Many natural language processing applications are based on language models that define a probability distribution over sequences of words, characters, or bytes in a natural language (Coneglian, 2018; Goodfellow; Yoshua, 2016).

Natural language processing searches for patterns and indicators that help understand the text under analysis. Thus, studies of natural language processing and machine learning increasingly converge due to the large amount of data that is generated daily since through this data, the computer learns (Coneglian, 2018; Goodfellow; Yoshua, 2016).

Natural language processing has focused on the treatment and analysis of masses of unstructured data, especially in text format, which has led to the emergence of different areas of activity such as answering systems for user questions, translations made by machines, voice and dialogue recognition, document classification, text recognition in images and sentiment analysis in texts (PNL [...], 2019; Prates, 2019).

Artificial intelligence techniques

Currently, the main intelligent techniques are inserted in the context of Machine Learning. This area is dedicated to the study of prediction and inference algorithms, which seek to simulate the brain as a learning machine on computers. Machine Learning includes statistical and Artificial Intelligence techniques to enable machines to make the most of their tasks based on data extracted from experience. Thus, algorithms can learn from this data, identify patterns, and make decisions with little human intervention (Bianchi, 2020; Muylaert, 2020; Russo, 2020).

In the context of machine learning, there are the following types of learning, as indicated by Carvalho (1994) and Waltrick (2020):

- Supervised Learning: the model must be taught on what should be done. In this sense, a set of labeled data must be provided for the model to learn, with this data being partitioned between sets for training and testing. This type of learning is generally applied when the objective is to classify or predict future occurrences;
- Unsupervised Learning: occurs when no external agent is indicating the desired response to the input patterns. Unlike previous learning, an unlabeled data set is given and the model is not taught what the end goal is. This type of learning is generally applied when the objective is to generate clusters;
- Reinforcement Learning: occurs when an external critic evaluates the answer provided. It is used in cases where the problem is not related to a dataset, but there is an environment to deal with, such as a game scenario or a city where autonomous cars drive. It uses the "trial and error" method, in which success is equivalent to a reward, while an error is equivalent to punishment.

The classifiers used in Machine Learning seek to organize objects between different categories, and, to this end, the classifier model analyzes the set of data provided. In this set, each data already contains a label indicating which category it belongs to, to "learn" how to classify other new data. In classification, the algorithms that implement this process are called classifiers (Han; Kamber; Pei, 2011; Ramos et al., 2018).

For Affonso et al. (2010), text classification is a technique used to automatically assign one or more predefined categories to a corpus under analysis. The most common applications are text indexing, text mining, message categorization, news, summaries, and periodical publication archives. In computer systems, the classification process involves techniques for extracting the most relevant information from each category, in addition to using this information to teach the system to correctly classify documents.

Deep Learning is a branch of Machine Learning based on a set of algorithms that attempt to model high-level abstractions of data. Some of its representations are inspired by the interpretation of information processing and communication patterns in the nervous system (Baberjee, 2020; Bianchi,

2020; Premlatha, 2019). According to the authors, Deep Learning is one of the Machine Learning architectures that has been applied to tagging parts of speech, translating, and classifying texts.

Classifiers

To compare different Artificial Intelligence techniques for classifying topic avoidance in essays, we used supervised learning classifiers, as explained previously. The classifier algorithms highlighted below are from the Scikit-Learn library (2021a), which uses Machine Learning techniques. All classifiers indicated below were applied in the experiments carried out in this work.

Multilayer Perceptron (MLP): The MLPClassifier is a neural network that has more than one layer of neurons. In cases where there is no possibility of a single straight line separating the classes, the MLP generates a classification plan (Affonso et al., 2010; Leite, 2018). The algorithm used for MLP training is called backpropagation and consists of four steps: initialization, activation, weight training, and iteration. The idea of the backpropagation algorithm is, based on the calculation of the error that occurred in the output layer of the neural network, to recalculate the value of the vector weights of the last layer of neurons (Leite, 2018; Moreira, 2018).

Decision Trees: The Decision Tree Classifier is a supervised learning algorithm used in classification and regression tasks. The decision tree is a method for approximating discrete-valued target functions, where the learned function is represented by a decision tree. Decisions are made based on a set of "if-then" rules (Mitchell, 1997). Decision trees represent one of the most simplified forms of a decision support system. From a set of data, the algorithm creates a representation of the knowledge embedded there, in tree format (Pessanha, 2019).

Random Forests: RandomForestClassifier is a supervised learning algorithm that creates a forest at random; The "forest" created is a combination (ensemble) of decision trees, in most cases trained with the bagging method. The main idea of the bagging method is that the combination of learning models increases the overall result (Costa da Silva, 2018).

Gradient Boosting: The GradientBoostingClassifier algorithm is a generalization of boosting for arbitrary differentiable loss functions. This algorithm is an accurate and effective procedure that can be used for regression and classification problems in a variety of areas, including web search classification or ecology, for example (Scikit-learn, 2021b).

The Gradient Boosting algorithm is included in the Ensemble group of classifiers. This classifier uses a combination of weak learner results to produce a better predictive model. A weak learner is defined as a classifier that is slightly correlated with the true classification. This means that in a weak learner, performance on any given training set is slightly better than chance prediction. In the Boosting technique, each weak learner is trained with a set of data, sequentially and in an adaptive manner, whereby a base model depends on the previous ones and, in the end, they are combined in a deterministic way (Silva, 2020).

AdaBoost: The basic principle of AdaBoostClassifier is to tune a sequence of weak learners. It is an ensemble method that trains and deploys decision trees serially, that is, on repeatedly modified versions of the data (Scikit-Learn, 2021b). AdaBoost can be used to boost the performance of any machine learning algorithm. These are models that achieve just above chance accuracy in a classification problem (Brownlee, 2016).

Stochastic Gradient Descent (SGD): SGDClassifier is a simple yet efficient algorithm employed to fit linear classifiers and regressors under convex loss functions such as Support Vector Machines and Logistic Regression (Scikit-learn, 2021c). SGD has been successfully applied to large-scale, sparse machine-learning problems often encountered in text classification and Natural Language Processing. The advantages of applying this algorithm are efficiency and ease of implementation, as it offers many opportunities for code adjustment. The SGDClassifier class implements a simple stochastic gradient descent learning routine that supports different loss functions and penalties for classification (Scikit-learn, 2021c).

Support Vector Machines (SVM): it is a set of supervised learning algorithms used for classification, regression, and outlier detection (an outlier or result that deviates from the average). SVMs

are effective in high-dimensional spaces and when the number of dimensions is greater than the number of samples (Scikit-learn, 2021d).

In addition to the Machine Learning classifiers highlighted above, there are also classifiers based on Deep Learning, such as the Convolutional Neural Network (Scikit-learn, 2021e).

Convolutional Neural Network (RNC): Deep Learning algorithm that can take an input image, assign importance (weights and biases that can be learned) to various aspects and objects in the image, and be able to differentiate one from the other. Convolutional Neural Networks are responsible for advances in image classification, forming the core of most current computer vision systems, from automatic Facebook photo tagging to self-driving cars. More recently, Convolutional Neural Networks have been applied to Natural Language Processing problems, for which promising results were obtained. Britz (2015), Carneiro (2020), and Rodrigues (2018) highlight that Convolutional Neural Networks are very effective for textual classification tasks.

Classifier evaluation metrics

Two main steps are performed for a technique to classify the data in a database. In the first step, the model is generated that learns through data training, normally using 70% to 80% of the base. This partitioning is called Cross-Validation (CV), which is a technique widely used to evaluate the performance of machine learning models. In the second stage, the separate data are tested, between 30% and 20% of the base, to estimate the performance of the technique, measuring the correctness of the model (Han; Kamber; Pei, 2011; Ramos et al., 2018). After classification, it is necessary to evaluate the performance of the classifier using some metrics for this purpose. To distinguish between the actual class and the predicted class, the labels "Positive (P)" and "Negative (N)" are used, which are used for the class predictions produced by a model.

According to Ramos et al. (2018), given a classifier and an instance to classify, a Confusion Matrix (CM) is generated with four possible results (Chart 1):

a) True Positive (TP), when the evaluated label is true and the model resulted in a positive value, indicating the correctness of the model;
b) False Negative (FN), when there was an error in the model that predicted the negative class when the real value was positive, indicating a model error;
c) True Negative (TN), when the evaluated label is negative and the model resulted in a negative value, indicating the correctness of the model;
d) False Positive (FP), when the evaluated label is positive and the model returns a negative value, indicating a model error.

Chart 1 – Confusion matrix

		Preview	
		Yes	**No**
Real	**Yes**	True Positive (TP)	False Negative (FN)
	No	False Positive (FP)	True Negative (TN)

Source: Rodrigues (2019).

The Confusion Matrix seen in Chart 1 demonstrates how the previously mentioned nomenclatures are arranged. From these, evaluation metrics such as accuracy, precision, Recall, and F1-score (Rodrigues, 2019) are taken, as shown in Chart 2.

Chart 2 – Assessment metrics

Metrics	Description	Formula
Accuracy	It indicates the overall performance of the model. Among all classifications, how many did the model classify correctly.	$\frac{TP + TN}{TP + FV + FP + FN}$
Precision	Among all the positive class classifications that the model made, how many are correct.	$\frac{TP}{TP + FP}$
Recall /Revocação/Sensibilidade:	Among all positive class classifications as expected value, how many are correct.	$\frac{TP}{TP + FN}$
F1-Score	The Harmonic mean between precision and recall.	$\frac{2 * Precision * Recall}{Precision + Recall}$

Source: Adapted from Rodrigues (2019).

The metrics presented can be used in different types of classifiers, such as those demonstrated in previous topics: Multilayer Perceptron (MLP); Decision Tree; Random Forest; Gradient Boosting; Adaboosting; Support Vector Machine; and Convolutional Neural Networks.

RESEARCH METHOD AND MATERIALS

This is applied and experimental research carried out through the application of algorithms and measurement of the results obtained in developed experiments.

Database and experiment platform

To conduct the experiments, a sample of 1,320 essays distributed across 119 different themes was used. This corpus was extracted from the database available in the open-source UOL Portal (2019) repository and from the platform developed by Pinho et al. (2020), to set up a repository of essays corrected by different teachers and student levels. In this case, the essays were corrected by a teacher, in learning environments focused on the parameters applied in Enem. All the essays that make up the corpus used in the experiments had already been corrected by teachers, therefore having the respective grades and comments assigned by teachers at the end of the correction process.

Regarding its structure, the database has twelve columns, as exemplified in Chart 3. The columns considered for conducting the first experiments were: essay, theme, and escape, with the latter (escape) being responsible for the classification process, consisting of the target attribute of this study.

Chart 3 – Database structure used in the initial phase of the experiments

Essay	Theme	Motivating Text	Comp score1	Comp score2	Comp score3	Comp score4	Comp score5	Total	Evaluator comment	Escape
Text with the essay	Essay topic	Text provided in the proposal talking about the subject of the topic	Score between: 0, 40, 80, 120, 160 e 200.	Score between: 0, 40, 80, 120, 160 e 200.	Score between: 0, 40, 80, 120, 160 e 200.	Score between: 0, 40, 80, 120, 160 e 200.	Score between: 0, 40, 80, 120, 160 e 200.	0 to 1000.	Evaluator's comment	Yes or No

Source: Created by the authors.

The distribution of essay grades given by teachers to compose the database can be seen in Table 1, considering that the initial objective of this research is to analyze whether the essay deviated from the proposed theme. There were 230 essays classified as "out of topic", representing 17% of the total database considered for the experiments in this research.

Table 1 – Distribution of essay grades

Evaluated Essays		
Quantity	**Score**	**Percentage**
266	0 points. (Obs.: 230 essays were classified as avoiding the topic).	20.1%
33	20 to 100 points	2.5%
172	120 to 300 points	13%%
293	320 to 500 points	22.2%
252	520 to 700 points	19.2%
192	720 to 900 points	14.5%
112	920 to 1.000 points	8.5%
Total: 1,320		**Total: 100.0%**

Source: Crated by the authors.

Data distribution

To separate the database, cross-validation was used, to prevent only a portion of training and test data from being very similar. This is because, in this case, when there were tests with new data that were very different from the trained model, the results would be unsatisfactory. Working with different distributions makes it possible to reduce the risk of bias when working with just one sample. In this way, the data were separated into three different sets, following the idea demonstrated in Chart 4, which shows the three different groups generated to carry out the experiments.

These groups were mixed to generate the three samples. Thus, three experiments were carried out with each classifier. The proposal was to use the same data sets, both in the Sklearn classifiers and in the Convolutional Neural Network. Thus, it was possible to compare the results between the techniques, as they used the same distributions. The process visualized in Chart 4 makes it possible to avoid data variance, in addition to making it possible to understand whether the experiments carried out brought the same average results with different combinations.

Chart 4 – Demonstration of cross-validation for the training and testing process

1st sample 1,320 essays	**2nd sample 1,320 essays**	**3rd sample 1,320 essays**
Drill	Test 2	Test 1
Test 1	Drill	Test 2
Test 2	Test 1	Drill

Source: Created by the authors.

To use cross-validation in training, the script was inserted into the training process, using the Sklearn cross_val_predict library. Therefore, the results demonstrated in the results presentation and discussion section already include such data distribution. Chart 5 shows the process performed using the Sklearn cross_val_predict library.

Chart 5 – Example of applying the Sklearn cross_val_predict library

```
>>> from sklearn.model_selection import cross_val_predict
# I need to have the database imported
# then separate the input data into (x) and target column (y)
# Configuration of the Classification Model, with all hyperparameters
>>>y_pred = cross_val_predict(model, x, y, cv=3)
```

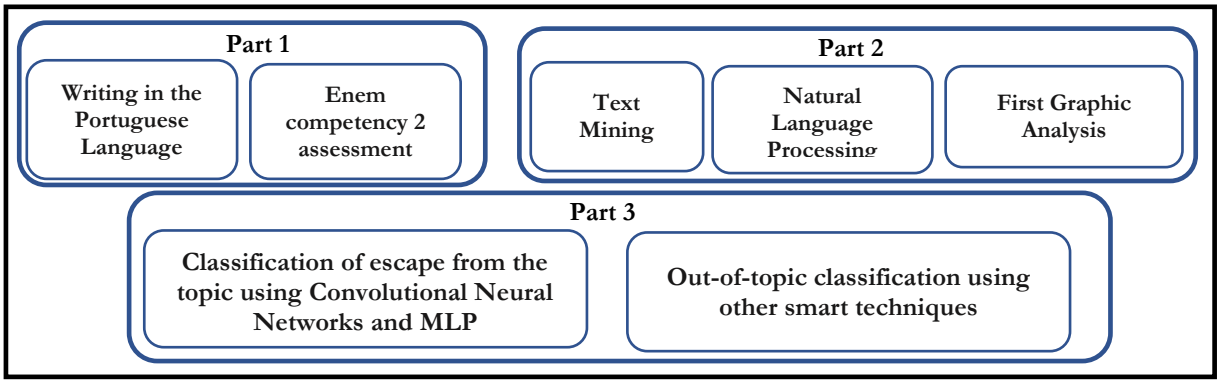
Source: Created by the authors.

The application of the library shown in Chart 5 occurs when the model is trained when cross-validation is carried out to determine which data set obtained the best result.

Activity flow for executing experiments

In addition to the steps previously defined for the research, a flow of activities was also generated for the elaboration of the experiments, as shown in Figure 1. The sequence developed initially consists of essays in Portuguese to evaluate deviations established by Enem competency 2 (away from the topic), the focus of this work.

Figure 1 – Activity flow



Source: Created by the authors.

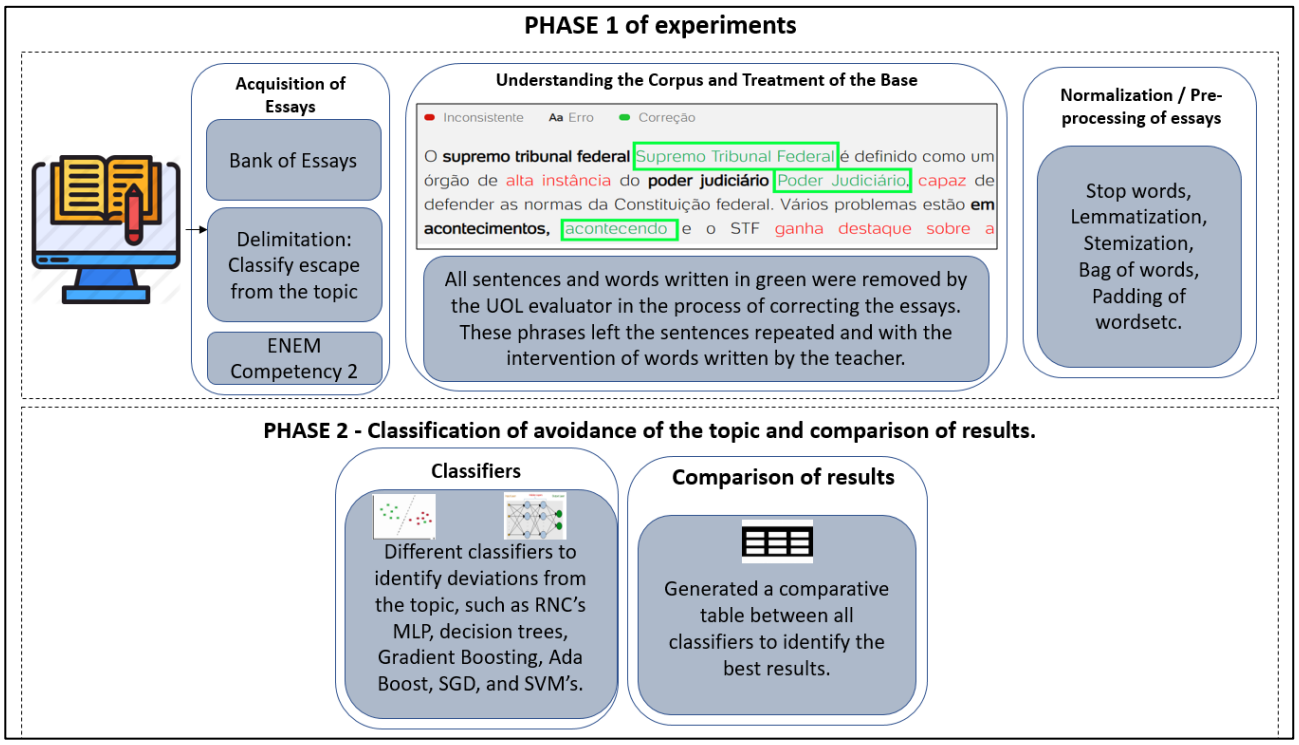
As shown in Figure 1, the steps relevant to Text Mining, Natural Language Processing and, finally, base training were carried out, using Convolutional Neural Networks and Multilayer Perceptron. In addition to the neural networks, training was also carried out with other classification techniques such as Decision Tree Classifier, Random Forest Classifier, Gradiente Boosting, Ada Boost, Stochastic Gradiente Descent, and Support Vector Machine. The expected results after applying the model and selected techniques sought to classify the escape from the topic as positive or negative, considering the essays under analysis to generate the future solution to be validated.

Details of the experiments

After presenting the flow of experiments, we sought to detail their sequence, as shown in Figure 2. Analyzing phase 1 of the sequence shown in Figure 2, initially a database with 1,320 essays was used to identify writing deviations in competency 2 (avoidance of the topic). These essays were then analyzed by the researchers, and at this stage, it was identified that some had repeated words and phrases. They were entered in the correction process by the evaluator, with all sentences and words highlighted in green. Thus, such repetitions were removed, and the texts returned to their original form.

Continuing the analysis of Figure 2, after removing the redundant excerpts, generated in the teachers' evaluation, the next step was to apply the first Natural Language Processing and Text Mining techniques to normalize the documents and thus prepare them for the first graphical analysis, clustering (grouping) and preparation of training models.

Figure 2 – Sequence of experiments



Source: Created by the authors.

In the data normalization phase, a function was created in Python, which performed all the necessary processing. At this stage, the Spacy library was used, and a function was created to make all characters lowercase, carry out the process of tokenization, lemmatization, removal of stop words, and removal of special characters from the essays in the corpus under analysis. In Figure 3, the Python function script developed to perform data normalization can be seen.

Figure 3 – Data normalization script developed in Python

```
def preprocessamento(text):
    result = []
    pos_tag = ['ADJ', 'NOUN', 'VERB', 'PROPN']
    texto = re.sub(u'[^a-zA-Z0-9áéíóúÁÉÍÓÚâêîôÂÊÎÔãõÃÕçÇ: ]', '', text )
    doc = pln(texto.lower())
    for token in doc:

        if(token.text in stop_words or token.text in pontuacoes):
            continue

        if(token.pos_ in pos_tag):
            result.append(token.text)

        texto = ' '.join([str(elemento) for elemento in result if not elemento.isdigit()])


    return texto
```

Source: Created by the authors.

In Figure 4 below, an example of the result of a standardized essay theme is shown. The entire theme was transformed into lowercase letters, without special characters and without stop words, which would be the words "ao" and "ou". The lemmatization process can also be visualized, in which there is the action of deflecting a word to determine its lemma, as can be seen in the word "*fumo*", which became "*fumar*", as well as in the word "*combate*", which became "*combater*".

Figure 4 – Visualization of texts after normalization

Writing topic not normalized: Combating smoking: authoritarianism or government duty

Normalized essay theme: combat smoking authoritarianism govern duty

Source: Created by the authors.

In the last phase of the experiments, the techniques that established the classification of the essays under analysis were applied. Then, a comparison was generated between all the techniques, to highlight the particularities of the application of each of them in the experiments, as shown in Table 6.

Chart 6 – Details of the experiments

Etapas/bibliotecas/experimentos realizados	
1)	Import of the libraries necessary to start the experiments: Spacy, Pandas, Numpy, MatplotLib, and Scikit-learn.
2)	Extraction of the database with 1,320 essays.
3)	Data pre-processing, cleaning the database with the normalization process, removing stop words, removing special characters, stemming, stemming, and padding of words.
4)	Before the model training process, we sought to collect information about the texts, generating some histograms to understand the database.
5)	Still, in the phase of understanding the database and looking for ways to help teachers find out which topics students have the most difficulty writing, the process of similarity between the essays and their themes was then carried out. This was carried out using words related to the essay written by the student and the motivating text provided by the institution.
6)	For the training process, essays with grades higher than 499 points and essays that strayed from the topic were used. We used 628 essays for the training process, and, for the testing process, 209 essays were used (test essays correspond to 33% of the trained essays). In the training process, cross-validation was used. In this step, we check how the model's accuracy behaves in different categories and classes.
7)	After dividing the database into training and testing, the first classifier tested was the convolutional neural network. To input the neural network into the training process, the first step was to transform the target variable "escape" into numerical values "0" and "1". Avoidance of the theme is now represented by "1", and non-escape from the theme by "0".
8)	Next, all texts needed to be vectorized, going through the padding process, and creating a vector of numbers that represent each word. At this stage, each essay and theme proposal was vectorized to be in the same dimension. For the Convolutional Neural Network, the "tfds"' library from "tensorflow_datasets" was used.
9)	Once all data preparation was carried out, the next step was to configure the neural network model, determining the number of layers of the neural network, the number of neurons for each layer, the number of filters, the number of documents evaluated to update the weights and the number of epochs to execute the training.
10)	After the training process, the next step was to use the generated model applied to the 209 separate essays for the testing phase.
11)	After applying the first test, the confusion matrix was generated with the results and accuracy of the experiment.
12)	Another test was carried out, this time with essays that had not yet been trained, that is, those with grades lower than 500.
13)	For the next classifiers, the process of normalizing the essays was also carried out. As with the Convolutional Neural Network, it was necessary to create a new representation of (discrete) textual values, since the classifiers used only understand numbers. Therefore, it was necessary to convert the raw data, which is in text format, to numeric format. This must happen before passing the essays to the classifier. The vectorization process for Scikit-learn classifiers was carried out using TfidfVectorizer.
14)	The essays that had already been divided into training and testing were then configured in each Scikit-learn classifier, using the hyperparameters defined in the library documentation.
15)	The results of the experiments will be demonstrated in the next section.

Source: Created by the authors.

The results of the experiments mentioned in Chart 6 are shown in the next section, to facilitate the visualization of how such experiments are presented to end users (teachers and students). It

is worth mentioning that these results are available on the CRIA platform developed from these experiments.

PRESENTATION AND ANALYSIS OF RESULTS

The first experiment carried out to detect patterns in the essays that indicated a departure from the topic was the application of the grammatical marking technique, also known as post-tagging, from the Spacy library. In this experiment, verbs, nouns, and adjectives were extracted from both the essay proposal (motivating text and theme) and the essay, considering the terms that provide the most meaning to the writing.

After obtaining the terms with the highest semantic value from each text, these were then compared to find the words that the student wrote that adhered to the proposed theme for preparing the essay. In Figure 5, an example of the analysis of the correlation between the written essay and the essay proposal for cases that deviate from the topic is shown.

Figure 5 – Example of correlation between essay and essay proposal – Essays that stray from the topic

Violence and drugs: the role of the user {‘ drugs: 2’}
Violence and drugs: the role of the user {‘violence’: 3, ‘security’: 1, ‘brazil’: 2, ‘man’: 1, ‘thought’: 1, ‘country’: 1}

Source: Created by the authors.

Figure 5 shows two essays classified as escaping the topic and the relationship between the essay prepared by the student and the proposed topic entitled “Violence and Drugs: The Role of the User”. When analyzing the content of Figure 5, it is possible to identify in the first essay that only the word “drugs” appeared twice in the entire essay, demonstrating that the student did not use any other word related to the proposal that was indicated by the institution applying the essay. The second essay shows that the student used six words that may be related to the previously indicated theme, words that were found in the essay proposal. However, it was possible to verify that the student did not use the words “drugs” and “user”, as well as other words that were related to the proposal, but that could have been used in different contexts, not just on this theme.

Figure 6 illustrates the correlation between the written essay and the essay proposal for two essays classified without straying from the topic, with the first receiving a score of 700, and the second receiving a score of 900.

Figure 6 – Example of correlation between essay and essay proposal – Essays without straying from the topic

Violence and drugs: the role of the user {'violence': 3, 'drugs': 9, 'commerce': 1, 'illegals': 1, 'law': 1, 'supply' 1:, 'demand': 1, 'politics': 4, ' prohibitionist': 2, 'special': 1, 'country': 3, 'decriminalization': 4, 'debate': 2}
Violence and drugs: the role of the user {'violence': 3, 'drugs': 3, 'trafficking': 3, 'commerce': 3, 'health': 1, 'law': 2, 'offer': 1, 'guilt', 'prohibition' : 3, 'illegality': 1, 'parcel': 1, 'marijuana': 1, 'brazil': 1, 'population': 1, 'finance': 1, 'consumption': 3, 'regulation': 1 , 'problem': 1}

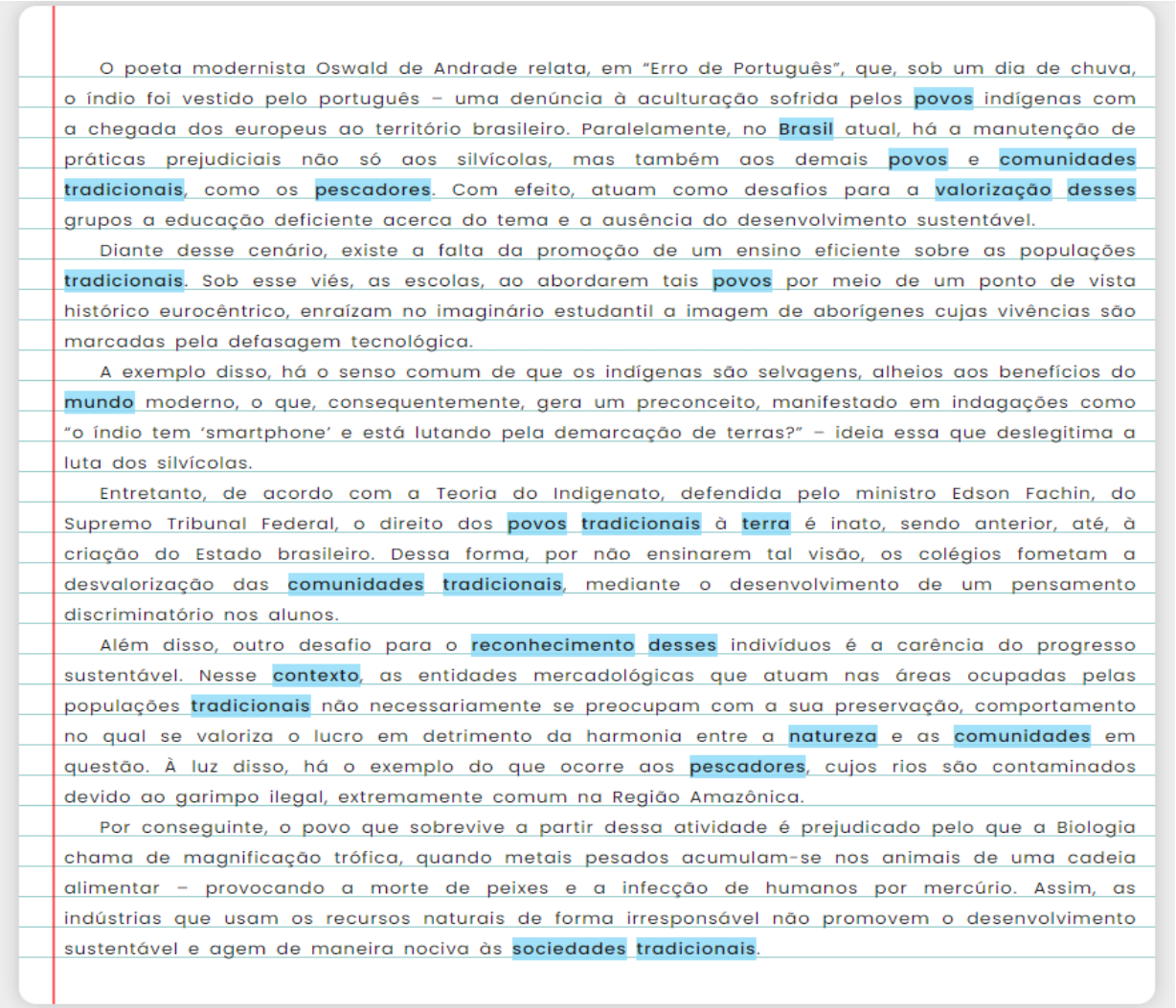
Source: Created by the authors.

When analyzing the results of the two essays shown in Figure 6, which have the same theme addressed in Figure 5, it was possible to identify that, in addition to there being many words related to the motivating text, such words have a greater relationship with the theme, such as: “decriminalization”, “illegals”, “drugs”, “politics” and “trafficking”, among other occurrences.

These results could be made available to the teacher together with an indication of the probability of avoiding the topic, which will be displayed in the classifiers for the next topics. Thus, in

addition to indicating the percentage of alignment with the essay proposal, the teacher would also indicate the words that most adhered to the essay proposal. This part of the experiments is already available for use on the developed CRIA platform. Figure 7 shows how the student and the teacher can identify the words that adhere to the theme proposed in the content of an elaborate essay.

Figure 7 – Identification of sticky words in an essay



O poeta modernista Oswald de Andrade relata, em "Erro de Português", que, sob um dia de chuva, o índio foi vestido pelo português – uma denúncia à aculturação sofrida pelos povos indígenas com a chegada dos europeus ao território brasileiro. Paralelamente, no Brasil atual, há a manutenção de práticas prejudiciais não só aos silvícolas, mas também aos demais povos e comunidades tradicionais, como os pescadores. Com efeito, atuam como desafios para a valorização desses grupos a educação deficiente acerca do tema e a ausência do desenvolvimento sustentável.

Diante desse cenário, existe a falta da promoção de um ensino eficiente sobre as populações tradicionais. Sob esse viés, as escolas, ao abordarem tais povos por meio de um ponto de vista histórico eurocêntrico, enraízam no imaginário estudantil a imagem de aborígenes cujas vivências são marcadas pela defasagem tecnológica.

A exemplo disso, há o senso comum de que os indígenas são selvagens, alheios aos benefícios do mundo moderno, o que, consequentemente, gera um preconceito, manifestado em indagações como "o índio tem 'smartphone' e está lutando pela demarcação de terras?" – ideia essa que deslegitima a luta dos silvícolas.

Entretanto, de acordo com a Teoria do Indigenato, defendida pelo ministro Edson Fachin, do Supremo Tribunal Federal, o direito dos povos tradicionais à terra é inato, sendo anterior, até, à criação do Estado brasileiro. Dessa forma, por não ensinarem tal visão, os colégios fometam a desvalorização das comunidades tradicionais, mediante o desenvolvimento de um pensamento discriminatório nos alunos.

Além disso, outro desafio para o reconhecimento desses indivíduos é a carência do progresso sustentável. Nesse contexto, as entidades mercadológicas que atuam nas áreas ocupadas pelas populações tradicionais não necessariamente se preocupam com a sua preservação, comportamento no qual se valoriza o lucro em detrimento da harmonia entre a natureza e as comunidades em questão. À luz disso, há o exemplo do que ocorre aos pescadores, cujos rios são contaminados devido ao garimpo ilegal, extremamente comum na Região Amazônica.

Por conseguinte, o povo que sobrevive a partir dessa atividade é prejudicado pelo que a Biologia chama de magnificação trófica, quando metais pesados acumulam-se nos animais de uma cadeia alimentar – provocando a morte de peixes e a infecção de humanos por mercúrio. Assim, as indústrias que usam os recursos naturais de forma irresponsável não promovem o desenvolvimento sustentável e agem de maneira nociva às sociedades tradicionais.

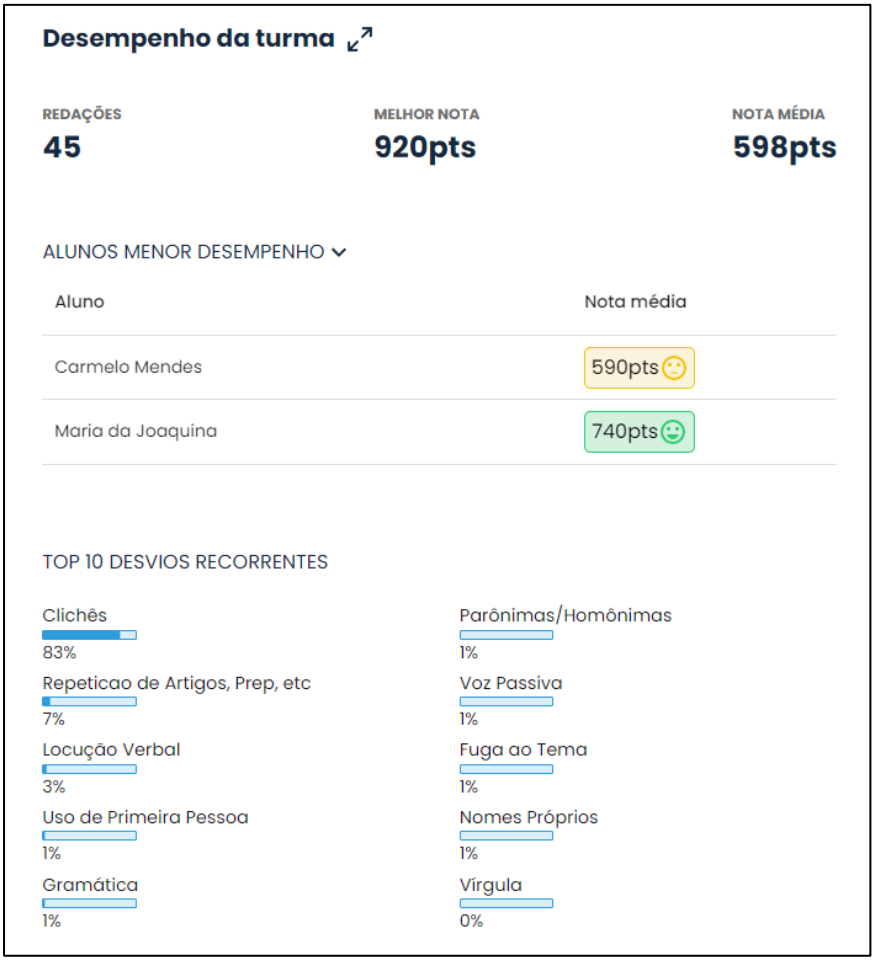
Source: CRIA Plataform (2023).
Note: The figure shows the text originally written in Portuguese.

The platform provides the quantity and distribution of these words per paragraph. The essay shown in Figure 7 has the theme: "Challenges for valuing traditional communities and peoples in Brazil". It is possible to verify that the text presented in the example contained 25 words identified by their similarity to the proposed theme.

This first phase of the experiments allowed the teacher to generate relevant information about the performance of their students and classes since the platform makes students' recurring errors available as well as those mistakes that are most frequently made in writing. In this way, the teacher identifies those students who have the greatest difficulties in textual production, as well as the themes that are most difficult for students to assimilate. This solution (CRIA) has already been implemented and has been providing benefits to students, who can rewrite their essays before even sending them to the teacher. This method is especially valid when the students realize that they may have touched on the topic proposed for the essay. In addition, there is also a benefit for the teacher, as he/she receives an essay previously analyzed by the student and, even so, can interfere if he/she does not agree with the notes indicated by the CRIA platform.

The solution developed also provides the teacher with access to an evaluation panel after the correction is carried out by the intelligent technique. In it, the teacher visualizes the performance of each class, as well as the exposure of recurring errors in that class, as shown in Figure 8. It is also available to view the essays assigned to him/her for correction. The screens shown in the Figure 8 allow the teacher to evaluate the class's biggest mistakes to work on in the classroom, as well as validate the corrections coming from intelligent technical correction.

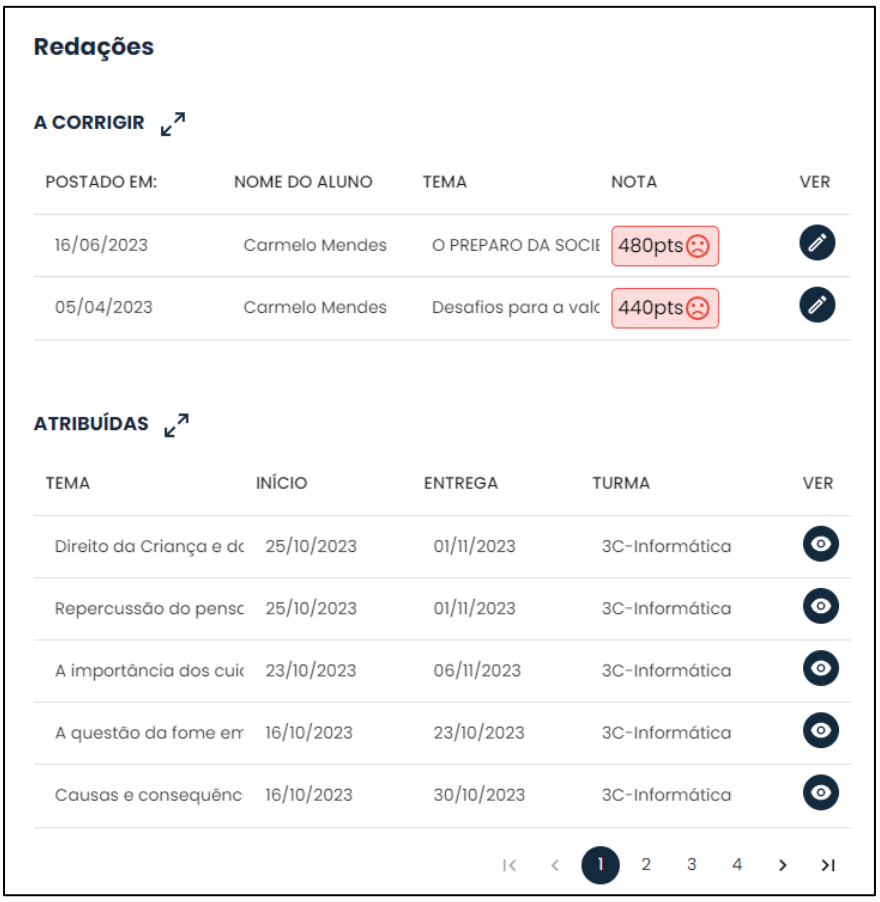
Figure 8 – Teacher Panel after correction of the intelligent technique for monitoring and evaluating class performance



Source: CRIA Plataform (2023).

Note: The names of the indicated students are fictitious. This screen of the CRIA platform is only available in Portuguese.

Figura 8 – Teacher Panel after correction of the intelligent technique for class evaluation class performance (continuation)



Source: CRIA Plataform (2023).
Note: The names of the indicated students are fictitious. This screen of the CRIA platform is only available in Portuguese.

Preparation of the base for the classifiers

To apply the classifiers selected in this experiment, a new stage was initiated. After the process of standardizing the essays, it was still necessary to adapt the texts to the classifiers, as these do not recognize texts. Therefore, the text padding and vectorization process was applied before starting the training of the developed model.

The size of the vectors created in this stage of the experiment was 510 words, which means that, after the normalization procedure carried out, the largest essay or motivating text found had 510 words. Thus, the algorithm completes the vectors with zero to leave them all in the same dimension before starting to train the next models.

Once the data was processed, the next step was to apply the different classification models selected for this experiment. To apply the classifiers, the models were generated, applying the two separate test bases, as occurred with the Convolutional Neural Network. Next, the evaluation metrics used (Accuracy and Confusion Matrix) generated for each model in question were applied.

Avoidance of topic classification using convolutional neural networks

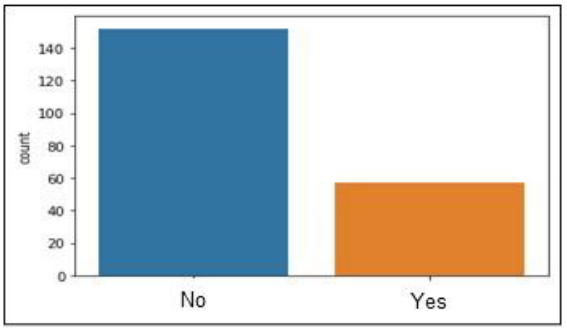
The results of the Convolutional Neural Network were arranged separately from the subsequent classifiers, as it uses other techniques for its application, using the Spacy library for this purpose, in addition to using a vectorization method different from the Scikit-Learn classifiers. In the third phase of the experiments, after carrying out the entire process of pre-processing the essays through normalization, removal of stop words, removal of special characters, stemming and lemmatization, and word padding, as explained in the method and materials topic research, the database was finally separated into the training set and test set.

For training the Convolutional Neural Network, 628 essays were used, which is equivalent to 47% of the essays in the database under analysis. The choice of essays for training and testing the model considered the actual grade assigned by teachers to each essay. The best results were obtained when the essays had grades above 499 points. This criterion also considered that this is normally the margin that universities use as a criterion for eliminating candidates (Universidades [...], 2020).

The configuration of the designed model used the TensorFlow keras library, with the following parameters: emb_dim = 200; nb_filters = 700; ffn_units = 1000; batch_size = 32; dropout_rate = 0.2; nb_epochs = 40. Initial parameter values were provided by Granatyr (2020) on the IA Expert Platform. In the following experiments, the parameters were adjusted according to the results presented.

After the training process and model generation, the next step was testing with the remaining 209 essays, which are equivalent to 33% of the training base. The balance between essays that deviated from the topic and those that did not can be seen in Figure 9. The essays that deviated from the topic were equivalent to 17% of the total analyzed. In the training phase, the same percentage of essays that strayed from the topic was also considered.

Figure 9 – Test base distribution



Source: Created by the authors.

After applying the test base to the model, one of the metrics used to evaluate the results and compare the different algorithms tested was the Confusion Matrix. In Table 2, the Confusion Matrix of the experiment carried out is presented, with the results of the Convolutional Neural Network for the first experiment with the test base, demonstrating the errors and successes of the model.

Table 2 – Confusion Matrix – First test base

0 (Royal Category)	143	9
1(Royal Category)	29	28
	0 (Prevision)	1 (Prevision)

Source: Created by the authors.

When analyzing the results shown in Table 2, it was identified that, of the 152 essays that did not escape the topic, the model classified nine essays as leakage, indicating 5.9% error, a value that represents the false positives. In the second line of the Confusion Matrix, 57 essays avoided the topic. Of these, the technique classified 28 essays as leaks, which is equivalent to 49.0% accuracy, a value that represents the true positives, that is, the class escaping the topic was determined as a positive class. This hit rate may have been due to the limitation of the database, which contained a few examples of deviations from the topic. This limitation has been overcome over time with the extension of the application of the experiments in this research on the CRIA platform, a service already implemented with schools and teachers. In this first test, the accuracy was approximately 81.8%, comparing the results of the solution implemented in the experiments of this research with the real results arising from the corrections made by the teachers.

A second test was carried out with essays with scores less than or equal to 499, that is, normally those that are disregarded by most universities for candidates' approval. The results can be seen in Table 3, which shows the Confusion Matrix generated for the other 540 essays that were not used in

the training phase. In the distribution of this database, there were 483 essays without escaping the topic and 57 essays with escape.

Table 3 – Confusion Matrix – Second test base

0 (Royal Category)	455	28
1(Royal Category)	29	28
	0 (Prevision)	1 (Prevision)

Source: Created by the authors.

In the second test carried out with essays with lower scores, an accuracy of 89.4% was obtained. Another analysis was carried out based on the predictions to find out if it was possible to extract any information from the false positives incurred, as the system reported that 28 essays were classified as out of topic, but they were not out of topic. This is because the error rate for false positives in this second test base was 5.7%, and the false negative rate was 94.3%. In the case of true negatives, the same result as in the previous experiment was obtained.

Table 4 consolidates the results obtained in the Convolutional Neural Network classification process. Despite having obtained accuracy results greater than 80% in both databases, the greatest attention is paid to false positives, that is when the system states that the essay mistakenly avoided the topic. The error rate was between 5.9% and 5.7% in the classification of escape from the topic for both bases, when in fact there was none. It is important to assess whether these essays may have partially deviated from the proposed theme, which will also lower the student's grade.

Table 4 – Consolidated results of the confusion matrix – Convolutional Neural Network

First test base with essays with scores above 500 and 57 essays that deviate from the topic					
RNC Classifier	Accuracy	TP (True Positive)	FN (False Negative)	FP (False positive)	TN (True Negative)
	81.8%	49.1%	50.9%	5.9%	94.1%
Second test base with essays with grades lower than 500 and 57 essays that deviate from the topic					
RNC Classifier	Accuracy	TP (True Positive)	FN (False Negative)	FP (False positive)	TN (True Negative)
	89.4%	49.1%	50.9%	5.7%	94.3%

Source: Created by the authors.

The results of Precision, Recall and F1-score, based on the results of applying the Convolutional Neural Network, are presented in Table 5, considering "1" for escaping the topic and "0" for not escaping the topic.

Table 5 – Consolidated precision, recall and F1Score results – Convolutional Neural Network

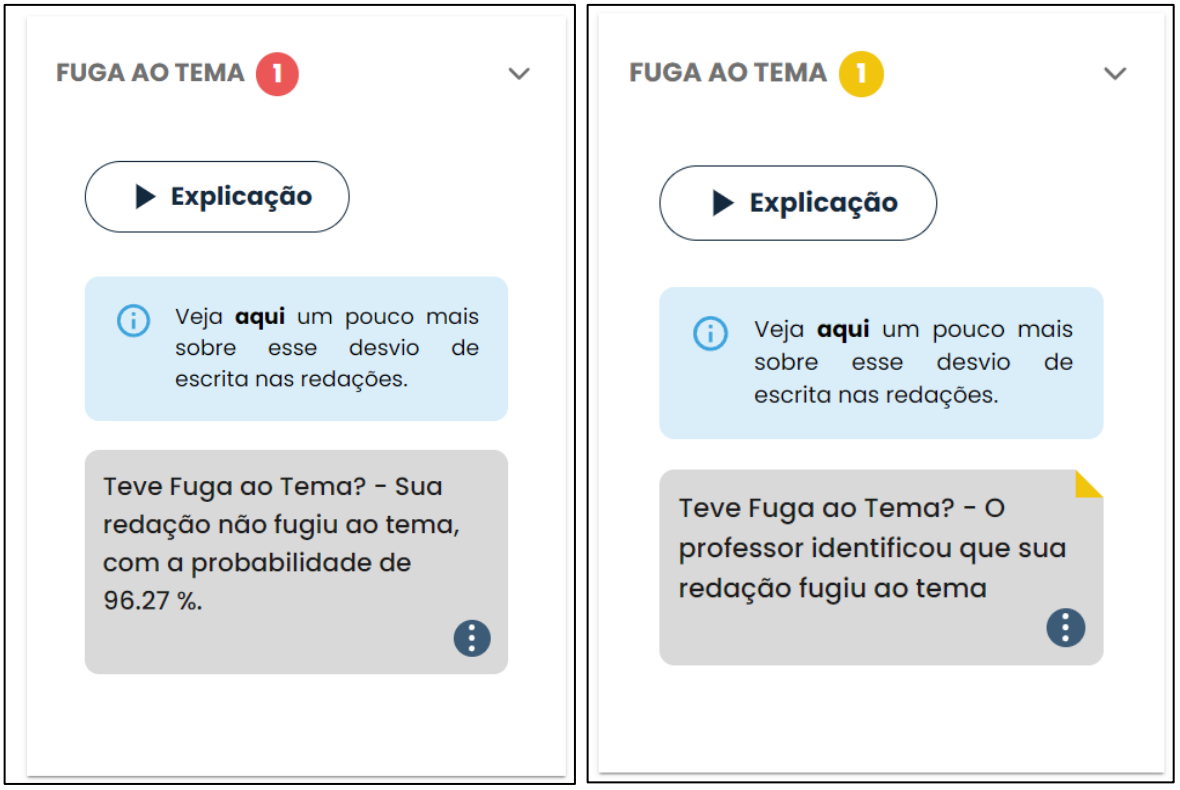
First test base with essays with grades above 500 and 57 essays that deviate from the topic				
		Precision	Recall	F1-Score
RNC Classifier	0	94%	94%	94%
	1	50%	49%	50%
Second test base with essays with grades lower than 500 and 57 essays that deviate from the topic				
		Precision	Recall	F1-Score
RNC Classifier	0	94%	94%	94%
	1	50%	49%	50%

Source: Created by the authors.

The results presented regarding the model's accuracy indicate all the positive class classifications that the model made, and how many are correct. Recall among all positive class situations as an expected value indicates how many are correct. Finally, the F1-Score makes a harmonic average between the other two. These metrics also indicate a greater degree of accuracy for the negative class of essays, that is, those that did not escape the topic.

Figure 10 shows a practical example of the presentation of the probability of avoiding the topic, as available on the CRIA platform. The student and teacher receive an indication of the probability of escaping the topic. If the teacher does not agree, he or she can change the indication of the correction made by Artificial Intelligence using the "do not agree" option. This way, the system will continue to learn from the teacher's assessment profile, while the model is being retrained. On the right side of Figure 10, the student can identify whether the teacher interfered in the assessment of Artificial Intelligence through the indicative text.

Figure 10 – Practical example of breakout probability return



Source: CRIA Plataform (2023).
Note: The names of the indicated students are fictitious. This screen of the CRIA platform is only available in Portuguese.

The percentage of probability of avoiding the topic presented by the platform to the student and teacher, which in the example shown in Figure 10 was 99.61%, is an important indicator of the adherence of the written essay to the proposed topic. Thus, the lower the percentage presented, the lower the indication of an escape from the topic. Based on the percentage presented by the platform, the teacher can also associate this percentage with the result of the first experiment of this research, focused on words that adhere to the proposed theme.

In this research, after the classification process using Convolutional Neural Networks, the next step was to test other classifiers.

Results of escape-to-topic classification applying other selected classifiers

The results relating to the Machine Learning classifiers from the Scikit-Learn library are presented below, after a new normalization of the essays. The selected Scikit-Learn classifiers were explained in the classifiers topic, these being: MLPClassifier, DecisionTreeClassifier, RandomForestClassifier, SGDClassifier, SVM (SVC), GradientBoostingClassifier, AdaBoostClassifier.

For all classifiers, the same Convolutional Neural Network base was used. The essays were divided into training and testing, using the same division criteria stipulated for the Convolutional Neural Network, enabling comparison with the results obtained. The same metrics used in the experiments carried out with the Convolutional Neural Network were also applied: Confusion Matrix; Accuracy; Recall and F1-score.

Table 6 shows the results of each classifier, using the same criteria applied to the Convolutional Neural Network, that is, the scores assigned by the human evaluator equal to or greater than 500 points.

Table 6 – Scikit Learn Classifiers – Consolidated Confusion Matrix

First test base with essays with grades above 500 and 57 essays that deviate from the topic					
Classifier	Accuracy	TP (True Positive)	FN (False Negative)	FP (False Positive)	TN (True Negative)
MLPClassifier	78%	33%	67%	4.6%	95.4%
DecisionTreeClassifier	74.6%	14%	86%	2.6%	97.4%
RandomForestClassifier	72.7%	0%	100%	0%	100%
SGDClassifier	78%	47%	53%	11%	89%
SVM (SVC)	72.2%	0%	100%	0%	100%
GradientBoostingClassifier	74.6%	51%	49%	16%	84%
AdaBoostClassifier	77%	44%	56%	11%	89%

Source: Created by the authors.

To identify the best results from the first test base, the accuracy combined with the True Positives and True Negatives were considered, in addition to the classification that presented the smallest error of the False Positives, that is, the results that indicated that the writing had leaked to the topic, when in fact there was none. This is because this item is what would cancel the student's test and should therefore have the lowest error rate.

Thus, taking this information into account, the classifier that presented the best results was the GradientBoostingClassifier, with the highest escape-to-theme (TP) accuracy of 51%, FP error rate of 16%, and accuracy of 74.6%. Other classifiers that performed well were SGDClassifier and MLPClassifier, with the following error rates: 11% and 4.6%, respectively, presenting an accuracy of 86% for both classifiers.

Table 7 presents the results of the second test base, that is, those essays that had scores assigned by the human evaluator below 500 points.

Table 7 – Scikit-Learn Classifier Results – Second Test Base

Second test base with essays with grades lower than 500 and 57 essays that deviate from the topic					
CLASSIFIER	ACCURACY	TP (True Positive)	FN (False Negative)	FP (False Positive)	TN (True Negative)
MLPClassifier	91%	33%	67%	3,3%	96,7%
DecisionTreeClassifier	82%	14%	86%	9%	91%
RandomForestClassifier	89,4%	0%	100%	0%	100%
SGDClassifier	86,6%	47%	53%	8,7%	91,3%
SVM (SVC)	89,4%	0%	100%	0%	100%
GradientBoostingClassifier	72,9%	51%	49%	24%	76%
AdaBoostClassifier	71,2%	44%	56%	25%	75%

Source: Created by the authors.

The results relating to those essays with grades below 500 points shown in Table 5 showed the same hierarchy order as the classifiers with the best performance. The results in Table 7, despite dealing with essays with lower grades, obtained similar results to the previous database. The biggest

difference was seen in False Positives, which, in this case, obtained the MLPClassifier and SGDClassifier as classifiers with the lowest error in False Positives, with respectively 3.3% and 8.7%.

Table 8 shows the results of the other evaluation metrics to assist in understanding the results of the developed model.

Table 8 – Precision, Recall and F1-Score metrics – Consolidated Scikit-Learn classifiers

MLPClassifier	Precision	Recall	F1-Score
Class 0 (without escape)	79%	95%	87%
Class 1 (Escaping the topic)	73%	33%	46%
DecisionTreeClassifier	Precision	Recall	F1-Score
Class 0 (without escape)	75%	97%	85%
Class 1 (Escaping the topic)	67%	14%	23%
RandomForestClassifier	Precision	Recall	F1-Score
Class 0 (without escape)	*73%*	*100%*	*84%*
Class 1 (Escaping the topic)	*0%*	*0%*	*0%*
SGDClassifier	Precision	Recall	F1-Score
Class 0 (without escape)	*82%*	*89%*	*85%*
Class 1 (Escaping the topic)	*61%*	*47%*	*53%*
SVM (SVC)	Precision	Recall	F1-Score
Class 0 (without escape)	*73%*	*100%*	*84%*
Class 1 (Escaping the topic)	*0%*	*0%*	*0%*
GradientBoostingClassifier	Precision	Recall	F1-Score
Class 0 (without escape)	*82%*	*84%*	*83%*
Class 1 (Escaping the topic)	*54%*	*51%*	*52%*
AdaBoostClassifier	Precision	Recall	F1-Score
Class 0 (without escape)	*81%*	*89%*	*85%*
Class 1 (Escaping the topic)	*61%*	*44%*	*51%*

Source: Created by the authors.

When evaluating the results in Table 8, the classifiers with the best results and which maintained similar values in the precision, Recall, and F1-Score metrics were the SGDClassifier and GradientBoostingClassifier. This is because the results of the negative class (non-escape) were above 80%, and the successes of the positive class (escape from the topic) maintained an average above 50%.

The MLPClassifier classifier had accuracy in both classes (positive and negative) above 70%. In the Recall item, for the positive class, only 33% was correct. The F1-Score made a harmonic average between the results of the previous two and, for the positive class, it also remained below 50%.

Evaluation and discussion of results

Considering the experiments carried out in this research, starting with those who have not yet used the classifiers, it has already been possible to identify some possibilities for evaluating students' writing and/or argumentation in essays. Based on this principle, it is understood that it is possible to measure whether the writing adheres to the informed theme proposal, which can bring important knowledge to the evaluator or teacher in the students' evolution in textual production, with paragraph-by-paragraph markings.

When evaluating the results obtained from the Convolutional Neural Network classifiers, a greater gain was identified to the Scikit-Learn classifiers, both about accuracy and false positive results, Precision, Recall, and F1-Score metrics. Although the best results were obtained with the Convolutional Neural Network, the good results obtained with the Scikit-Learn classifiers should be highlighted.

When evaluating to designate the best model, it is important to understand how the different models would be applied in a real situation. In this way, the false positive (FP) error percentage must be minimal, as well as the true positive (TP) percentage must be high, which means that the model more adequately identifies the escape from the topic, the central proposal of this research.

Therefore, to analyze the results, the models that were unable to identify escape from the topic were disregarded, that is, those that presented a true negative rate (TN) of 100%, meaning that they

were unable to achieve the objective proposed in this article. research, namely: Random Forest Classifier and SVM (SVC).

Next, the model that had the best accuracy was the Convolutional Neural Network, with results of up to 89% accuracy and a false positive rate (FP) of just 5.7%. However, if the true positive rate (TP) is evaluated, the one in which the model got the topic right, the best results occurred with the GradientBoostingClassifier, with 51% of hits in the positive class; however, its false positive rate (FP) was 20%, on average. Another classifier that achieved better results concerning false positives (FP) was MLPClassifier, with a maximum error of 4.6%, a true positive rate (TP) of 33%, and an accuracy between 78% and 90%.

However, such metrics are likely to achieve better results when applied to a larger database with more examples of essays classified as off-topic, especially the Convolutional Neural Network. This is because, according to Rodrigues (2018), the Convolutional Neural Network obtains better results with a large base of examples. However, based on the results presented in this research, constant experiments continued to be carried out seeking to improve accuracy and increase the database. The studies carried out in this research constituted the basis for putting the CRIA Platform into production.

Since October 2022, five schools have tested and validated the CRIA platform before its official launch. Teachers who used it reported time savings, as the student already received an instant assessment. Thus, the student makes a critical assessment of their writing based on the indications of writing deviations made available by the platform, including avoidance of the topic, in addition to the indication of explanatory articles for reading. The objective is to make the student understand "what" and "why" he/she made a mistake or why he/she should not make a certain mistake, helping him/her to understand the Enem rules for writing an essay. The teacher receives for evaluation an essay with fewer errors, thus providing more effort and time to carry out a more contributory evaluation, with the indication of correction points that the platform was unable to detect in an automated way.

The increased use of the CRIA platform has made it possible to increase the database for new training. This has been achieved by receiving essays from different schools, age groups, and the social diversity of students. The diversity of essays has made it possible to reduce biases in assessment, as well as expand the database with texts already corrected by teachers.

In the classroom context, such application has provided gains concerning the time spent on correction and less teacher strain when evaluating texts. For the student, the provision of faster feedback has proven to be positive, in addition to the guarantee of subsequent evaluation by the subject teacher. Thus, the solution validated in these experiments contributes to reducing the time and resources used in the process of evaluating texts produced by students, without discarding the role of the teacher as the protagonist of this process.

In the research carried out, no other authors were found who used the same methods and techniques applied in this research to evaluate the "avoidance of the topic" criterion in essays. The closest research to this work was the study carried out by Passero (2018), which specifically analyzed escape from the topic. The author obtained excellent results with an accuracy of 96.76% and false positives (FP) of 4.24%. However, the study did not make available the results of the true positive rate (VP), that is, those results that identified an escape from the topic, a crucial factor in this research.

The CRIA platform developed by the University of São Paulo (USP, 2021) also aimed to automatically correct essays, however without checking for deviation from the topic. Thus, in the application that is available for use, there is a high rate of errors when evaluating essays with grades considered low. In the tests carried out, essays that should have received a grade of 0 (zero) for avoiding the topic were evaluated with grades higher than 400. Ramisch's research (2020) also evaluated essays, however, he proposed to find problems of syntactic deviations, achieving accuracy 75.6% accuracy.

Therefore, it is understood that the results presented in this research make important contributions to the evolution of the study of this area of academic research. Thus, based on the results presented here, it is possible to glimpse the first prerogatives of benefits of the solution now developed to aid the work of teachers and evaluators during the process of correcting texts produced by students or candidates. The solution now validated based on the experiments carried out in this research has been applied in a system to evaluate the evolution of students throughout their academic studies, thus allowing the teacher to understand the individual difficulties of students in a class.

CONCLUSIONS

The application of the techniques mentioned in the experiments carried out in this research sought to indicate which Artificial Intelligence techniques compared present better results for identifying avoidance of the topic in newsrooms. Thus, it is understood that this study achieved the established research objective, since, after applying different classifiers to formulate a model, it was possible to indicate those that bring better results when identifying escape from the topic in the newsrooms, indicating the percentage of success of the model designed in this research.

The experiments brought promising results, both in Convolutional Neural Networks and Scikit-Learn classifiers. The model that achieved the best accuracy was the Convolutional Neural Network, with results of up to 89.4% accuracy and a false positive rate (FP) of 5.7%. However, if the true positive rate (TP) is evaluated, the rate at which the algorithm got the topic right, the best results occurred with the GradientBoostingClassifier, with 51% of hits in the positive class, despite its false positive rate (FP) was, on average, 20%. Another classifier that achieved better results concerning False Positives (FP) was MLPClassifier, with a maximum error of 4.6%, a true positive rate (TP) of 33%, and an accuracy between 78% and 90%.

In this way, the proposed objective of comparing different Artificial Intelligence techniques for classifying missing topics in texts and identifying those that brought the best results to enable an intelligent essay correction system was achieved. The solution developed in this research makes it possible to generate useful information and knowledge for teachers and evaluators of educational texts in the task of identifying deviations in writing and possible escape from the proposed topic for writing, problems that, once incurred, lead to insufficient grades. to students.

The results of the experiments carried out here demonstrate an assertiveness of 89.4% accuracy for the Convolutional Neural Network classifier. This result made it possible to create an application to provide automatic feedback, as support for teachers or text evaluators, which helps to reduce the time required for correction, in addition to providing better assistance to institutions, teachers, and students. However, other classifiers can be applied in parallel, to enable a double-checking process, which will help to more efficiently signal the indication of a departure from the topic to the teacher/essay evaluator.

The solution now developed aims to reduce inequality in selection processes that use evaluation of essays prepared by candidates, offering greater learning opportunities regardless of the institution where the student studies. This is because this automated solution could also be applied in schools for training purposes in the process of preparing and correcting essays. In addition, the solution validated in this research, in addition to providing the possibility of training and improving writing quality, also enables faster returns for those involved in the teaching-learning process, that is, teachers and students.

For educational institutions with a high load of texts produced, the solution provided in this research can enable the teacher to move from an audit function of basic aspects in corrections to a function more focused on proving the effectiveness of student learning in correcting the text. Furthermore, the teacher's fatigue factor would be reduced, as currently the responsible professional corrects around 50 essays a day, in the case of evaluating texts produced in Enem. This is because, by adopting the solution validated in this research, the teacher could count on a system that indicates probable errors, which would greatly facilitate the work of evaluators and teachers. In other contexts, such contributions could even be beneficial to teachers who dedicate themselves exclusively to correcting texts. This is because such professionals could make use of the indications provided by the solution and then direct their efforts to other, more complex demands of the text correction activity.

As principais contribuições deste estudo buscam permitir ao avaliador, professor ou empresas que aplicam processos seletivos avaliarem as redações com menor esforço, otimizando assim o trabalho e reduzindo o tempo e o custo do processo de avaliação de textos dissertativos. A solução delineada nesta pesquisa pode, portanto, ser primordial na aplicação do Enem digital, proporcionando ao avaliador, assim, auxílio na identificação das falhas de escrita e minimizando interferências como fadiga e alteração de humor do avaliador, sintomas estes que podem afetar a correção de um texto dissertativo.

The main contributions of this study seek to allow evaluators, teachers, or companies that apply selection processes to evaluate essays with less effort, optimizing the work and reducing the time and cost of the dissertation text evaluation process. The solution outlined in this research can be essential in the application of the digital Enem, providing the evaluator with assistance in identifying writing flaws and minimizing interferences such as fatigue and changes in the evaluator's mood, symptoms that can affect the correction of a dissertation text.

From an academic perspective, the experiments carried out and the results presented can serve as a basis for studies that, once aligned with the knowledge of teaching professionals, can generate new approaches that enable students to write texts that are cohesive to the proposed topic. With this, the inequality between students from public and private schools can be reduced, since the existence of a platform that allows more frequent training and faster response can contribute to facilitating not only the teacher's work but also providing, in the future, clearer and more developed writing, thus enabling greater student maturity in this process.

A limiting issue for the results presented in this study refers to the database used in the experiments carried out. We intend to add more essays to this sample, notably with a greater occurrence of writing errors. For this purpose, mainly those deviations in writing that resulted in a zero grade for avoiding the topic, since in the corpus of this study there were only 230 essays that deviated from the topic diagnosed by the evaluator. Another limiting factor in this research is the selection of Artificial Intelligence techniques used in the experiments carried out in this study, under the responsibility of the authors.

For future research, it is recommended to expand the essay database, aiming to provide more effectiveness to experiments with Artificial Intelligence techniques. This fact is already very close to reality, since, through the CRIA platform, we seek to increase the base of essays and examples of escape from the topic. Furthermore, the possibility of adding other AI techniques, other than those used in this study, is also indicated.

REFERENCES

AFFONSO, Emmanuel T. F.; SILVA, Alisson M.; SILVA, Michel P.; RODRIGUES, Thiago M. D.; MOITA, Gray F. Uso de redes neurais multilayer perceptron (MLP) em sistema de bloqueio de websites baseado em conteúdo. *Mecánica Computacional*, v. XXIX, n. 93, p. 9075-9090, 2010.

BANERJEE, Dibyendu. Natural language processing (NLP) simplified: A step-by-step guide. *Data Science Foundation*, 2020, publicado em 14/4/2020. Disponível em: https://datascience.foundation/sciencewhitepaper/natural-language-processing-nlp-simplified-a-step-by-step-guide. Acesso em: 15 abr. 2021.

BIANCHI, Alexandre. As classificações dos algoritmos de machine learning. *Viceri-Seidor*, 2020, publicado em 27/5/2020. Disponível em: https://www.viceri.com.br/insights/as-classificacoes-dos-algoritmos-de-machine-learning. Acesso em: 9 maio 2021.

BITTENCOURT JÚNIOR, José A. S. *Avaliação automática de redação em língua portuguesa empregando redes neurais profundas*. 2020. 100 f. Dissertação (Mestrado em Ciência da Computação). Goiânia: Universidade Federal de Goiás, 2020.

BRASIL. Ministério da Educação e Cultura (MEC). MEC realiza conferência para discutir estratégias de alfabetização no Brasil. *Portal MEC*, 2019, publicado em 22/10/2019. Disponível em: http://portal.mec.gov.br/component/tags/tag/5?start=60. Acesso em: 15 abr. 2021.

BRITZ, Denny. Understanding convolutional neural networks for NLP. *Denny's Blog*, 2015, publicado em 7/11/2015. Disponível em: http://www.wildml.com/2015/11/understanding-convolutional-neural-networksfor-nlp/. Acesso em: 31 mar. 2021.

BROWNLEE, Jason. Boosting and adaboost for machine learning. *Machine Learning Mastery*, 2016. Disponível em: https://machinelearningmastery.com/boosting-and-adaboost-for-machine-learning/. Acesso em: 10 jun. 2021.

CÂNDIDO, Thiago G.; WEBBER, Carine G. Avaliação da Coesão Textual: Desafios para Automatizar a Correção de Redações. *RENOTE – Revista Novas Tecnologias na Educação,* v. 16, n. 1, p. 1-10, 2018.

CARNEIRO, Álvaro L. C. Redes neurais convolucionais para processamento de linguagem natural. *Medium*, 2020, publicado em 7/7/2020. Disponível em: https://medium.com/data-hackers/redes-neurais-convolucionais-para-processamento-de-linguagem-natural-935488d6901b. Acesso em: 31 mar. 2021.

CARVALHO, André C. P. F. de; FAIRHURST, Michael C.; BISSET, David L. An integrated boolean neural network for pattern classification. *Pattern Recognition Letters*, v. 15, p. 807-813, ago./1994.

CONEGLIAN, Caio S. *Recuperação da informação com abordagem semântica utilizando linguagem natural:* a inteligência artificial na ciência da informação. 195 f. Tese (Ciência da Informação – FFC). São Paulo: Universidade Estadual Paulista – UNESP, 2018.

COSTA DA SILVA, Josenildo. Aprendendo em uma floresta aleatória. *Medium*, 2018, publicado em 12/3/2018. Disponível em: https://medium.com/machina-sapiens/o-algoritmo-da-floresta-aleat%C3%B3ria-3545f6babdf8. Acesso em: 31 mar. 2021.

CRIA – Plataforma CRIA. *Plataforma de Correção automática de Redações*, 2023. Disponível em: https://web.cria.net.br. Acesso em: 25 ago. 2023.

DIANA, Daniela B. G. Os 16 maiores erros de redação cometidos pelos estudantes. *Toda Matéria*, 2021, publicado em 8/1/2021. Disponível em: https://www.todamateria.com.br/erros-de-redacao/. Acesso em: 29 mar. 2021.

EGGERS William D.; SCHATSKY, David; VIECHNICKI, Peter. AI-augmented government using cognitive technologies to redesign public sector work. *Deloitte*, 2017. Disponível em: https://www2.deloitte.com/content/dam/insights/us/articles/3832_AI-augmented-government/DUP_AI-augmented-government.pdf. Acesso em: 30 mar. 2021.

GOMES, Maria de F. C. A PNA e a unidade dialética afeto-cognição nos atos de ler e escrever. *Revista Brasileira de Alfabetização*, n. 10, edição especial, p. 122-124, 2020. https://doi.org/10.47249/rba.2019.v1.368.

GONÇALVES, Eduardo C. Mineração de texto – Conceitos e aplicações práticas. *SQL Magazine*, v. 105, p. 31-44, nov. 2012. Disponível em: https://www.researchgate.net/publication/317912973_Mineracao_de_texto_-_Conceitos_e_aplicacoes_praticas. Acesso em: 15 abr. 2021.

GOODFELLOW, Ian; YOSHUA Bengio. *Deep learning.* Cambridge: MIT, 2016.

GRANATYR, Jones. Processamento de Linguagem Natural com Deep Learning. *Expert Academy*, 2020, curso realizado em novembro de 2020. Disponível em: https://iaexpert.academy/courses/processamento-linguagem-natural-deep-learning-transformer/. Acesso em: 10 dez. 2020.

HAN, Jiawei; KAMBER, Micheline; PEI, Jian. *Data mining: concepts and techniques:* concepts and techniques. New York: Elsevier, 2011.

HARIRI, Reihaneh H.; FREDERICKS, Erick M.; BOWERS, Kate M. Uncertainty in big data analytics: survey, opportunities, and challenges. *Journal of Big Data*, v. 6, n. 1, p. 1-16, 2019.

INSTITUTO NACIONAL DE ESTUDOS E PESQUISAS EDUCACIONAIS ANÍSIO TEIXEIRA (INEP). *A redação no Enem 2022:* cartilha do participante. Brasília, 2022. Disponível em: https://download.inep.gov.br/download/enem/cartilha_do_participante_enem_2022.pdf. Acesso em: 19 out. 2023.

INSTITUTO NACIONAL DE ESTUDOS E PESQUISAS EDUCACIONAIS ANÍSIO TEIXEIRA (INEP). *Entenda como é calculada a nota do Enem*. 2020. Disponível em: http://portal.mec.gov.br/ultimas-noticias/418-enem-946573306/84461-entenda-como-e-calculada-a-nota-do-enem. Acesso em: 10 jun. 2020.

INSTITUTO NACIONAL DE ESTUDOS E PESQUISAS EDUCACIONAIS ANÍSIO TEIXEIRA (INEP). *Microdados Enem 2022*. 2023. Disponível em: https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/enem. Acesso em: 20 ago. 2023.

LEITE, Tiago M. Redes neurais, perceptron multicamadas e o algoritmo backpropagation. *Medium,* 2018, publicado em 10/5/2018. Disponível em: https://medium.com/ensina-ai/redes-neurais-perceptron-multicamadas-e-o-algoritmo-backpropagation-eaf89778f5b8. Acesso em: 10 maio 2021.

LESME, Adriano. Enem 2021: corretores podem corrigir até 200 redações por dia. *Brasil Escola – UOL*, 2021, publicado em: 1º/12/2021. Disponível em: https://vestibular.brasilescola.uol.com.br/enem/enem-2021-corretores-podem-corrigir-ate-200-redacoes-por-dia/351641.html#:~:text=Cada%20profissional%20ter%C3%A1%20que%20avaliar,de%2015%20a%2020%20dias.&text=Com%20o%20t%C3%A9rmino%20do%20Exame,a%20supervis%C3%A3o%20de%20216%20profissionais. Acesso em: 19 ago. 2023.

LUDERMIR, Teresa B. Inteligência artificial e aprendizado de máquina: estado atual e tendências. *Estudos Avançados*, v. 35, n. 101, p. 85-94, abr./2021.

MITCHELL, Tom M. *Machine learning*. New York: McGraw-Hill, 1997.

MORAIS, Edison A. M.; AMBRÓSIO, Ana P. L. *Mineração de textos*. Goiás: UFG, 2007.

MOREIRA, Sandro. Rede neural perceptron multicamadas. *Medium*, 2018, publicado em 24/12/2018. Disponível em: https://medium.com/ensina-ai/rede-neural-perceptron-multicamadas-f9de8471f1a9. Acesso em: 15 abr. 2021.

MÜLLER, Sarah; BERGANDE, Bianca; BRUNE, Philipp. Robot tutoring: on the feasibility of using cognitive systems as tutors in introductory programming education – A teaching experiment. *In*: EUROPEAN CONFERENCE OF SOFTWARE ENGINEERING EDUCATION (ECSEE'18), 3rd, 2018. *Anais* […] Association for Computing Machinery, New York, USA, p. 45-49.

MUYLAERT, Renata. Pandemia do novo coronavírus, Parte 6: inteligência artificial (NLP). *Sobrevivendo na Ciência*, 2020. Disponível em: https://marcoarmello.wordpress.com/2020/08/19/coronavirus6/. Acesso em: 27 jul. 2021.

NOBRE, João C. S.; PELLEGRINO, Sérgio R. M. ANAC: um analisador automático de coesão textual em redação. *In*: BRAZILIAN SYMPOSIUM ON COMPUTERS IN EDUCATION – SBIE, 2010. *Anais* [...] São Paulo: SBC, 2010, p. 1-12.

PASSERO, Guilherme. *Detecção de fuga ao tema em redações de língua portuguesa*. 145 f. Dissertação (Mestrado em Computação Aplicada). Itajaí: Universidade do Vale do Itajaí, 2018.

PESSANHA, Cínthia. Random Forest*:* como funciona um dos algoritmos mais populares de ML. *Medium,* 2019, publicado em 20/11/2019. Disponível em: https://medium.com/cinthiabpessanha/random-forest-como-funciona-um-dos-algoritmos-mais-populares-de-ml-cc1b8a58b3b4. Acesso em: 27 jul. 2021.

PINHO, Cíntia M. A.; VANIN, Anderson S.; BELAN, Peterson; NAPOLITANO, Domingos M. R. Uma ferramenta on-line para ensino de Redação, baseada nos critérios avaliativos do ENEM. *In*: KMBRASIL 2020 – CONGRESSO BRASILEIRO DE GESTÃO DO CONHECIMENTO, 15º, São Paulo. *Anais* [...] São Paulo: SBGC, 2020, p. 599-615.

PINTO, Álvaro V. *O conceito de tecnologia*. Rio de Janeiro: Contraponto, 2005.

PNL*:* entenda o que é o processamento de linguagem natural. *STEFANINI – Group*, 2019. Disponível em: https://stefanini.com/pt-br/trends/artigos/oque-e-processamento-de-linguagem-natural. Acesso em: 20 maio 2021.

PRATES, Wlademir R. Introdução ao processamento de linguagem natural (PLN). *Ciência e Negócios*, 2019, publicado em 1º/8/2019. Disponível em: https://cienciaenegocios.com/processamento-de-linguagem-natural-nlp/. Acesso em: 27 jul. 2021.

PREMLATHA, Karan R. What is AI? In a simple way. *AI Time Journal*, 2019, publicado em 5/2/2019. Disponível em: https://www.aitimejournal.com/@premlatha.kr/what-is-ai-in-a-simple-way. Acesso em: 15 abr. 2021.

PREUSS, Evandro; BARONE, Dante A. C.; HENRIQUES, Renato V. B. Uso de técnicas de inteligência artificial num sistema de mesa tangível. *In*: WORKSHOP DE INFORMÁTICA NA ESCOLA, 26º, 2020, *Anais* [...] Porto Alegre: SBC, 2020, p. 439-448.

RAMISCH, Renata. *Caracterização de desvios sintáticos em redações de estudantes do ensino médio:* subsídios para o processamento automático das línguas naturais. 156 f. Dissertação (Mestrado em Linguística). São Carlos: Universidade Federal de São Carlos, 2020.

RAMOS, Jorge L. C.; SILVA, João C. S.; PRADO, Leonardo C.; GOMES, Alex S.; RODRIGUES, Rodrigo L. Um estudo comparativo de classificadores na previsão da evasão de alunos em EAD. *In*: BRAZILIAN SYMPOSIUM ON COMPUTERS IN EDUCATION, VII, 2018. *Anais* […] São Paulo: SBC, 2018, p. 1463-1472.

RIEDO, Cassio R. F. *Avaliação qualitativa imediata de produções escritas em EaD*. 266 f. Tese (Doutorado em Educação). Campinas: Universidade Estadual de Campinas, 2020.

RIOLFI, Claudia R.; IGREJA, Suelen G. da. Ensinar a escrever no ensino médio: cadê a dissertação? *Educação e Pesquisa*, v. 36, n. 1, p. 311-324, abr./2010.

RODRIGUES, Diego A. R. *Deep learning e redes neurais convolucionais:* reconhecimento automático de caracteres em placas de licenciamento. 37 f. Monografia (Ciência da Computação). João Pessoa: Universidade Federal da Paraíba, 2018.

RODRIGUES, Vitor. Métricas de avaliação – Quais as diferenças? *Medium*, 2019, publicado em 12/4/2019. Disponível em: https://vitorborbarodrigues.medium.com/m%C3%A9tricas-de-avalia%C3%A7%C3%A3o-acur%C3%A1cia-precis%C3%A3o-recall-quais-as-diferen%C3%A7as-c8f05e0a513c. Acesso em: 11 maio 2021.

RUSSO, Inês F. D. *O impacto da inteligência artificial na sustentabilidade ambiental:* uma agricultura sustentável. 84 f. Dissertação (Mestrado em Gestão de Sistemas de Informação). Lisboa: Universidade de Lisboa, 2020.

SANTOS JÚNIOR, Jário J. dos. *Modelos e técnicas para melhorar a qualidade da avaliação automática para atividades escritas em língua portuguesa brasileira.* 76 f. Dissertação (Mestrado em Informática). Maceió: Universidade Federal de Alagoas, 2017.

SCIKIT-LEARN. *AdaBoost.* 2021a. Disponível em: https://scikit-learn.org/stable/modules/ensemble.html#adaboost. Acesso em: 11 maio 2021.

SCIKIT-LEARN. *Gradient tree boosting.* 2021b. Disponível em: https://scikit-learn.org/stable/modules/ensemble.html#gradient-boosting. Acesso em: 11 maio 2021.

SCIKIT-LEARN. *Stochastic gradient descent.* 2021c. Disponível em: https://scikit-learn.org/stable/modules/sgd.html#sgd. Acesso em: 11 maio 2021.

SCIKIT-LEARN. *Support vector machines.* 2021d. Disponível em: https://scikit-learn.org/stable/modules/svm.html#svm-classification. Acesso em: 11 maio 2021.

SCIKIT-LEARN. *Supervised learning.* 2021e. Disponível em: https://scikit-learn.org/stable/supervised_learning.html#supervised-learning. Acesso em: 11 maio 2021.

SILVA, Jonhy. Uma breve introdução ao algoritmo de machine learning gradient boosting utilizando a biblioteca Scikit-Learn. *Medium*, 2020, publicado em 22/6/2020. Disponível em: https://medium.com/equals-lab/uma-breve-introdu%C3%A7%C3%A3o-ao-algoritmo-de-machine-learning-gradient-boosting-utilizando-a-biblioteca-311285783099. Acesso em: 11 maio 2021.

SILVEIRA, Rita C. B. da; BARROS, Manoel J. F. de. Impacto da inteligência artificial na empregabilidade docente. *In:* COLÓQUIO INTERNACIONAL DE GESTÃO UNIVERSITÁRIA – GIGU, XX, Florianópolis. *Anais* [...] Florianópolis: UFSC, 2021, p. 1-17.

SOUZA, Vanessa F. de; PERRY, Gabriela T. Mineração de texto em moocs: análise da relevância temática de postagens em fóruns de discussão. *RENOTE – Revista Novas Tecnologias na Educação*, v. 17, n. 3, p. 204-2013, 2019.

SQUARISI, Dad; SALVADOR, Arlete. *A arte de escrever bem:* um guia para jornalistas e profissionais do texto. 9. ed. São Paulo: Contexto, 2020.

STARLLES, Wender. Confissões de uma corretora de redações do Enem. *Guia do Estudante*, 2022, atualizado em 8/8/2022. Disponível em: https://guiadoestudante.abril.com.br/enem/confissoes-de-uma-corretora-de-redacoes-do-enem/. Acesso em: 19 ago. 2023.

TOKARNIA, Mariana. Enem é um dos principais instrumentos de acesso ao ensino superior. *Agência Brasil*, 2019, publicado em 31/10/2019. Disponível em: https://agenciabrasil.ebc.com.br/educacao/noticia/2019-10/enem-e-um-dos-principais-intrumentos-de-acesso-ao-ensino-superior. Acesso em: 19 mar. 2021.

UNIVERSIA. *Entrevista com ex-corretor de redação.* 2015. Disponível em: https://www.universia.net/br/actualidad/orientacion-academica/corretorredaco-do-enem-leva-cerca-2-minutos-prova-diz-professor-1132810.html. Acesso em: 19 mar. 2021.

UNIVERSIDADE DE SÃO PAULO (USP). USP desenvolve ferramenta de correção automática de redações. *Portal USP São Carlos*, 2021, publicado em 10/3/2021. Disponível em: http://www.saocarlos.usp.br/usp-desenvolve-ferramenta-de-correcao-automatica-de-redacoes/. Acesso em: 29 mar. 2021.

UNIVERSIDADES privadas de SP adotam vestibular online e nota do Enem. *Jornal Cruzeiro do Sul*, 2020, publicado em 9/6/2020. Disponível em: https://www.jornalcruzeiro.com.br/brasil/universidades-privadas-de-sp-adotam-vestibular-online-e-nota-do-enem/. Acesso em: 29 mar. 2021.

WALTRICK, Camila. Machine learning – O que é, tipos de aprendizagem de máquina, algoritmos e aplicações. *Medium*, 2020, publicado em 7/5/2020. Disponível em: https://medium.com/camilawaltrick/introducao-machine-learning-o-que-e-tipos-de-aprendizado-de-maquina-445dcfb708f0. Acesso em: 29 mar. 2021.

Submitted : 05/18/2022
Preprint: 04/25/2022
Approved: 09/13/2023

AUTHORS' CONTRIBUTIONS

Author 1 - Project administration, formal analysis, conceptualization, data curation, writing – first version, investigation, methodology, resources, software, supervision, validation, and visualization.
Author 2 - Project administration, formal analysis, conceptualization, writing – first version, writing – review and editing, investigation, methodology, resources, supervision, validation, and visualization.
Author 3 - Conceptualization, writing – review and editing, methodology, resources, software, supervision, and validation.

DECLARATION OF CONFLICT OF INTEREST

The authors declare that there is no conflict of interest with this article.