Expert Systems With Applications 257 (2024) 125097

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Towards explainable automatic punctuation restoration for Portuguese using transformers

Tiago Barbosa de Lima [a], Vitor Rolim [c], André C.A. Nascimento [a], Péricles Miranda [a], Valmir Macario [a], Luiz Rodrigues [b,*], Elyda Freitas [c], Dragan Gašević [d], Rafael Ferreira Mello [a]

[a] Federal Rural University of Pernambuco, Recife, Brazil
[b] Federal University of Alagoas, Maceió, Brazil
[c] Federal University of Pernambuco, Recife, Brazil
[d] Monash University, Clayton, Australia

ARTICLE INFO

Keywords:
Punctuation restoration
Transformers
Explainable AI

ABSTRACT

Accurate punctuation in written text enables unambiguous communication, minimizing the risk of misunderstandings. Conversely, faulty punctuation can confuse the intended meaning, posing challenges for the author. The existing literature offers a collection of systems and algorithms to assist users in writing tasks. However, those focusing on English tend to exhibit higher accuracy. Furthermore, most models for punctuation restoration yield results without offering insight into their decision-making processes. Therefore, this study evaluated state-of-the-art punctuation restoration models specifically for Brazilian Portuguese and incorporated the principles of explainable artificial intelligence to clarify their predictions transparently. The findings indicate that the models assessed achieved an accuracy comparable to those of their English-language counterparts.

1. Introduction

Punctuation is key in crafting written texts, enabling authors to articulate their intended messages with precision and clarity (Suliman, 2019). It indicates pauses, dictates voice inflexions, and demarcates distinct expressions, thereby facilitating enhanced comprehension of a sentence (Lenza & Martino, 2021). Conversely, improper punctuation usage can distort the intended meaning and signal a deficiency in language proficiency (Suliman, 2019).

In Natural Language Processing (NLP), the task of analysing texts to predict punctuation – most often, commas and periods – is called punctuation restoration (Klejch, Bell, & Renals, 2017; Lima et al., 2022; Tilk & Alumäe, 2016). It involves predicting sentence punctuation based solely on the text of the sentence itself (Klejch et al., 2017). Various techniques exist for this task, ranging from named entity recognition to sequence-to-sequence approaches (Klejch et al., 2017; Tilk & Alumäe, 2016). In general, deep learning models, such as Bidirectional Encoder Representation (BERT) and Text-to-Text Transfer Transformer (T5), outperforms traditional machine learning models like Conditional Random Fields (CRF) for this task (Lima et al., 2022; Nagy, Bial, & Ács, 2021; Nielsen, 2015; Tilk & Alumäe, 2016).

The majority of the previous work focused on punctuation restoration to support Automatic Speech Recognition tasks (Courtland, Faulkner, & McElvain, 2020; Lima et al., 2022; Makhija, Ho, & Chng, 2019; Pan, García-Díaz, & Valencia-García, 2023; Păiş & Tufiş, 2021; Tilk & Alumäe, 2016). However, this is a relevant task for other contexts as well. For instance, the capacity to employ punctuation accurately in formal writing is a crucial component of the assessment and correction criteria for student compositions (Kurup, Joshi, & Shekhokar, 2016). Moreover, automatically assessing punctuation can provide students with constructive feedback for refining their writing skills while also facilitating the identification of primary challenges students encounter in mastering a language (Suliman, 2019). Another concern is that most of the models proposed in the literature focus on analysing English texts (Nagy et al., 2021).

Furthermore, the predictions made by black-box algorithms, such as BERT and T5, fail to provide information about the prediction process effectively (Khosravi et al., 2022). In this scenario, Explainable AI (XAI) holds a promise as a valuable tool that can unpack the models' outcomes. Explainability in AI refers to making a model's prediction process understandable to humans (Kim, Khanna, & Koyejo,

[*] Correspondence to: Av. Lourival Melo Mota, S/n - Tabuleiro do Martins, Maceió AL, Brazil.
E-mail addresses: tiago.blima@ufrpe.br (T.B. de Lima), vbr@cin.ufpe.br (V. Rolim), andre.camara@ufrpe.br (A.C.A. Nascimento), pericles.miranda@ufrpe.br (P. Miranda), valmir.macario@ufrpe.br (V. Macario), luiz.rodrigues@nees.ufal.br (L. Rodrigues), elyda.freitas@upe.br (E. Freitas), dragan.gasevic@monash.edu (D. Gašević), rafael.mello@ufrpe.br (R.F. Mello).

https://doi.org/10.1016/j.eswa.2024.125097
Received 19 March 2024; Received in revised form 26 June 2024; Accepted 11 August 2024
Available online 14 August 2024

2016; Miller, 2019), helping identify crucial attributes in predictions, detect biases, and highlight features causing incorrect predictions (Arrieta et al., 2020). Due to deep learning's complexity, Explainable AI (XAI) models, using local and global explainability approaches, have been developed (Ribeiro, Singh, & Guestrin, 2016; Valenzuela-Escárcega, Nagesh, & Surdeanu, 2018; Yang, Rangarajan, & Ranka, 2018). Techniques like SHAP and Captum assign importance to each input feature (Lundberg & Lee, 2017; Ribeiro et al., 2016), with Captum leveraging transformers to visualize contributing tokens (Kumar & Boulanger, 2020). For sequence-to-sequence models, Inseq assesses attribute significance affecting predictions (Atanasova, Simonsen, Lioma, & Augenstein, 2020; Sarti, Feldhus, Sickert, & van der Wal, 2023). Hence, XAI contributes to ensuring reliable and transparent AI models (Khosravi et al., 2022; Tjoa & Guan, 2020), which can be used in practice, for instance, to aid educators in providing insightful feedback to students (Cavalcanti et al., 2021).

Therefore, the current study was conducted with the aim to evaluate a range of state-of-the-art punctuation restoration models for different types of texts written in Brazilian Portuguese. Initially focusing on the Automatic Speech Recognition task—as recommended in existing literature, we extended our evaluation to encompass various educational documents, including books and student essays. Moreover, we introduced XAI to analyse punctuation restoration models. To reach our goals, we investigated different punctuation restoration algorithms: BERT, T5, *Bidirectional Long Short Term Memory* (BLSTM), and Conditional Random Fields (CRF), which have been employed extensively in the field of punctuation restoration (Courtland et al., 2020; Lima et al., 2022; Makhija et al., 2019; Tilk & Alumäe, 2016). Moreover, we introduced, for the first time, the analysis of GPT-4 model in this task.

To assess the proposed models, we first employed the TEDTALK2012 dataset (Federico, Cettolo, Bentivogli, Michael, & Sebastian, 2012), containing lectures from Ted Talks 2012, widely used in previous studies. Additionally, we compiled a set of educational texts from the Linguistics and Computing Centre of the University of São Paulo (NILC) (Gazzola, Evaldo Leal, & Aluisio, 2019) to fine-tune the models in an educational context. Finally, the models were applied to restore the punctuation of essays composed by Brazilian middle school students. In summary, this article's contributions are:

- Evidence on how the performance of traditional and sequence-to-sequence models for punctuation restoration compare;
- Using XAI to analyse T5 model predictions on the essay dataset, illustrating alignment with Brazilian Portuguese grammatical norms, which could support automated feedback systems for educators.
- Demonstrating that BERT and T5 models outperformed traditional NER methods (CRF and BLSTM), indicating superior ability to capture word relationships and punctuation contexts.
- Expanding single dataset analyses by demonstrating our model generalizability across TEDTALK2012 and NILC datasets, with T5 achieving an F-score up to 0.883.
- Highlighting the impact of prompt variety on model performance based on GPT models

2. Background

This section presents background information on punctuation restoration, XAI, and GPT models.

2.1. Punctuation restoration

Punctuation restoration is usually modelled as a sequence labelling task, in which the predicted labels are punctuation marks such as periods, commas, and question marks. Sequence labelling is commonly employed to classify tokens in tasks like Part of Speech Tagging (POS), Named Entity Recognition (NER), and identifying complex words (Gooding & Kochmar, 2019). In the context of punctuation restoration, sequence labelling can assign punctuation to words, an approach adopted by multiple studies, such as those proposed in Tilk and Alumäe (2016) and Nagy et al. (2021).

Table 1
Example of the TEDTALK2012 dataset translated from Portuguese to English.

Sentence:	Then	I	just	started	to	play
Annotation:	I-COMMA	O	O	O	O	I-PERIOD

An initial motivation behind punctuation restoration was to enhance the readability of text generated by speech recognition algorithms (Courtland et al., 2020; Devlin, Chang, Lee, & Toutanova, 2018; Gooding & Kochmar, 2019; Nagy et al., 2021). A prominent dataset used in many of these studies is TEDTALK (Federico et al., 2012), which comprises audio descriptions of lectures. These lectures are available in multiple versions, differentiated by year and language, including English and Portuguese (Federico et al., 2012).

In general, the annotation procedures in punctuation restoration are (Păiş & Tufiş, 2021):

- Words without punctuation are labelled as 'O' (or [Other]).
- Words that require a comma are labelled as 'I-COMMA'.
- Words that need a period are assigned the 'I-PERIOD' label.

An illustrative example of this annotation can be found in the TEDTALK 2012 dataset, as displayed in Table 1.

Initial approaches to address this task often employed CRF. Recognized for their efficacy in NER, CRFs can be harnessed with linguistic features or integrated into a neural network architecture (Lafferty, McCallum, & Pereira, 2001; Lima et al., 2022; Lu & Ng, 2010; Manning, Surdeanu, Bauer, Finkel, Bethard, & McClosky, 2014). For instance, Lu and Ng (Lu & Ng, 2010) introduced factorial-CRF to restore the punctuation symbols in English and Chinese transcriptions. Their study utilized the BTEC (Basic Travel Expression Corpus) dataset to assess the model's performance in English. Additionally, they used the Challenge Task (CT) dataset, encompassing travel dialogues from the IWLST2009 corpus (Paul, Federico, & Stüker, 2010). Their findings revealed the benefit of employing factorial CRF over the conventional linear CRF model. For instance, in the English evaluation using the BTEC dataset, the accuracy achieved was 88%. However, it is worth noting that their study did not explore neural network-based models. Also, strategies to eliminate superfluous words from the text, akin to the approach presented in Baldwin, Cook, Lui, MacKinlay, and Wang (2013), were not assessed. It is important to highlight that previous works also adopted phonetic features from speech recordings to optimize the restoration model (Tilk & Alumäe, 2016).

In addition to utilizing shallow features, recent research has adopted word embedding techniques (Pennington, Socher, & Manning, 2014). The deployment of Global Vectors for Word Representation (GloVe) has led to considerable enhancements in tasks reliant on contextual word representation, such as NER (Pennington et al., 2014). In this direction, Tilk and Alumäe (2016) further extended this advancement by integrating BLSTM with 300-dimensional pre-trained GloVe vectors, topped with a final layer of CRF for punctuation restoration. When evaluated on English datasets, their method demonstrated an improvement of up to 8.9% compared to conventional approaches.

More recent research has expanded the capabilities of punctuation restoration through language models. Makhija et al. (2019) leveraged BERT to directly predict punctuation marks, achieving an impressive absolute improvement of 17 points in the overall F1-score over (Tilk & Alumäe, 2016) previous work. Furthering this line of investigation, Nagy et al. (2021) applied the BERT model combined with multilingual versions such as Hubert and mBERT. They obtained an F1-score of 0.798 for English and 0.822 for Hungarian. In comparing different techniques, the base BERT pre-trained model outperformed other models like the 300-dimensional CRF and BLSTM+skip-gram embedding, registering the highest F1 result (Nagy et al., 2021). Courtland et al. (2020) took a different approach, exploring sequence-to-sequence methods for punctuation restoration. Their evaluation included a wide range of pre-trained models such as XLNet-base, T5-base, BERT-base, ALBERT-base, DistilRoBERTa, and RoBERTa-large. Among these, RoBERTa-large stood out with remarkable improvements, boasting 48.7% in relative gains and 15.3% in absolute gains compared to the state-of-the-art. Recently, Vandeghinste and Guhr (2023) fine-tuned the RobBERT model to predict punctuation for the Dutch language. The results reached up to 0.789 for F1-score.

The literature identified three studies regarding works focusing on punctuation restoration for Portuguese. Initially, the work reported by Lima et al. (2022) utilized the TEDTALK2012 dataset, comprising 139,653 sentences for training, 1570 for validation, and 887 considering three graphic punctuation marks: period (.), comma (,), and question mark (?) with other punctuation marks being converted to a full stop (Federico et al., 2012). Three different approaches were evaluated by Lima et al. (2022): CRF (Lu & Ng, 2010), BLSTM+skipgram with 300-dimensional embeddings (Tilk & Alumäe, 2016), and BERT (Nagy et al., 2021). The findings of the Lima et al. study demonstrated that the best approach was BERT, achieving an F-score of 0.771. de Lima, Rodrigues, Macario, Freitas, and Mello (2023) extended the evaluation of pre-trained models by adding the BERT-large and T5 base to the analysis performed earlier in Lima et al. (2022). After five training epochs, the T5 model obtained an average F1-score of 0.88 and 0.89 for the base and large versions, respectively. The BERT-base and large models achieved average f1-score values of 0.882 and 0.87. Finally, the work proposed by Pan et al. (2023) investigated the application of punctuation restoration for Portuguese and Spanish. Pan et al. also included the word capitalization in the prediction. In this study, five punctuations were considered for Portuguese: full stop (.), question mark (?), colon (:), and exclamation mark (!), with tokens indicating uppercase (u) and lowercase (l). In Spanish, six graphic punctuation marks were added to the dataset: ¡!u, ¿l, ¡u, ¿l, ¿?l, ¡!u. The evaluation of various transformer architecture models, including BETO, DistilBETO, MarinAI, BERTIN, and XLM-R for Spanish, and BertTimbau, BR_BETO, and XLM-R for Portuguese, revealed that the pre-trained BETO model achieved an F1-score of 93.873 for Spanish and the XLM-R model with F1-score of 93.663 for Portuguese.

The above three studies underline a predominance in the use of pre-trained models as a solution for the punctuation marks problem, similar to English works (Courtland et al., 2020; Nagy et al., 2021). The current study followed the approach adopted in Lima et al. (2022), incorporating the evaluation of larger models (T5, BERT-large, and GPT) in different types of documents, not only for speech recognition.

2.2. Explainable artificial intelligence

Explainability refers to the capacity to render the prediction process of a model comprehensible to humans (Kim et al., 2016; Miller, 2019). This ability aids in identifying relevant attributes used in predictions, detecting biases, and pinpointing features leading to incorrect predictions (Arrieta et al., 2020). In contexts such as education and healthcare, explainability is critical in fostering more reliable and transparent AI models (Khosravi et al., 2022; Tjoa & Guan, 2020).

It is well known that predictions of deep learning algorithms are often difficult to understand due to their intrinsic complexity formed by multiple layers of learning (Danilevsky, et al., 2020). To address this, XAI models have been developed. XAI employs two primary approaches known as local and global explainability (Ribeiro et al., 2016; Valenzuela-Escárcega et al., 2018; Yang et al., 2018). *Local* explainability emphasizes visualizing the input attributes of a specific instance, highlighting features that led to a specific prediction. On the other hand, *global* explainability seeks patterns within the model as a whole, listing the most predictive features (Doshi-Velez & Kim, 2017). Techniques like *SHapley Additive exPlanations (SHAP)* and Captum assign an influence level to each input element regarding the intended label's prediction (Khosravi et al., 2022; Ribeiro et al., 2016). SHAP estimates complex model outputs via a simplified, explainable model, allocating an importance value to each input feature (Lundberg & Lee, 2017). These values, when combined, approximate the original model's output (Lundberg & Lee, 2017; Ribeiro et al., 2016). The methodology maintains consistency so that the importance assigned to a given input does not decrease as its contribution increases or remains the same. It assesses how the model output is influenced when conditioned to a specific input characteristic, starting from a base expected value (Lundberg & Lee, 2017). The SHAPE method is reported in the paper (Oliveira et al., 2023). The work evaluates BERT and other machine learning algorithms in cohesion score task both in English and Portuguese Essays (Oliveira et al., 2023). One of the challenges related to Essay cohesion prediction is the mismatch between human-generated and machine-generated features used for cohesion prediction (Oliveira et al., 2023). Then, the authors explore the relationship between the top 15 most important features related to cohesion prediction using CatBoost Regressor and their influence on the model's output. Moreover, the paper highlights the word tokens most relevant to BERT prediction score (Oliveira et al., 2023). With the provided analysis, it was possible to comprehend that human raters assign a higher score for longer Essays in the English context while Brazilian raters favour more lexical diverse text.

Captum uses a transformers architecture to facilitate the visualization of tokens that contribute most and least to the predicted label (Khosravi et al., 2022; Vaswani et al., 2017). It has been employed for automatic essay correction and extracts features related to sophistication, cohesion, and text complexity (Kumar & Boulanger, 2020).

In the context of sequence-to-sequence models, Inseq (Sarti et al., 2023) extracts the importance of attributes considering their influence on a prediction. It assigns a score that denotes the significance of each attribute for the model's final output (Atanasova et al., 2020; Sarti et al., 2023). Attributes of great importance notably alter the model's result when removed. Also, a low correlation with attributes is observed when the model lacks confidence in a prediction.

2.3. GPT models

The language models have rapidly advanced in recent years. These include well-known *encoder-only* models such as BERT, DEBERTa, and RoBERTa; *decoder-only* models like GPT-2 and BART; and *sequence-to-sequence* models such as T5 (Devlin et al., 2018; He, Liu, Gao, & Chen, 2020; Lewis et al., 2020; Radford et al., 2019; Raffel et al., 2020). These pre-trained models have achieved state-of-the-art performance in various NLP tasks, including question answering, textual similarity, NER, and text translation (Devlin et al., 2018; He et al., 2020; Lewis et al., 2020; Radford et al., 2019; Raffel et al., 2020). Initially, these models could only perform a single task when fine-tuned with thousands of examples, losing the generalization capacity of a language model (Brown et al., 2020). However, research by Raffel et al. (2020) demonstrated that language models can adapt to different tasks without the need to train new models, thus enabling knowledge transfer between activities (Alzubaidi et al., 2023; Brown et al., 2020). Despite this versatility, the computational expense of training for each new task and the frequent lack of properly labelled datasets can still pose significant limitations.

Several recent studies reported the results related to the use of different versions of GPT models. The study by Brown et al. (2020) revealed that language models like GPT-3 could learn new tasks with few or zero examples. GPT-3 demonstrated proficiency in over 20 different tasks, including zero-shot scenarios (with no examples), few-shot learning (with only a few examples), and in-context learning (with activity context provided up to the input boundary of the model). The

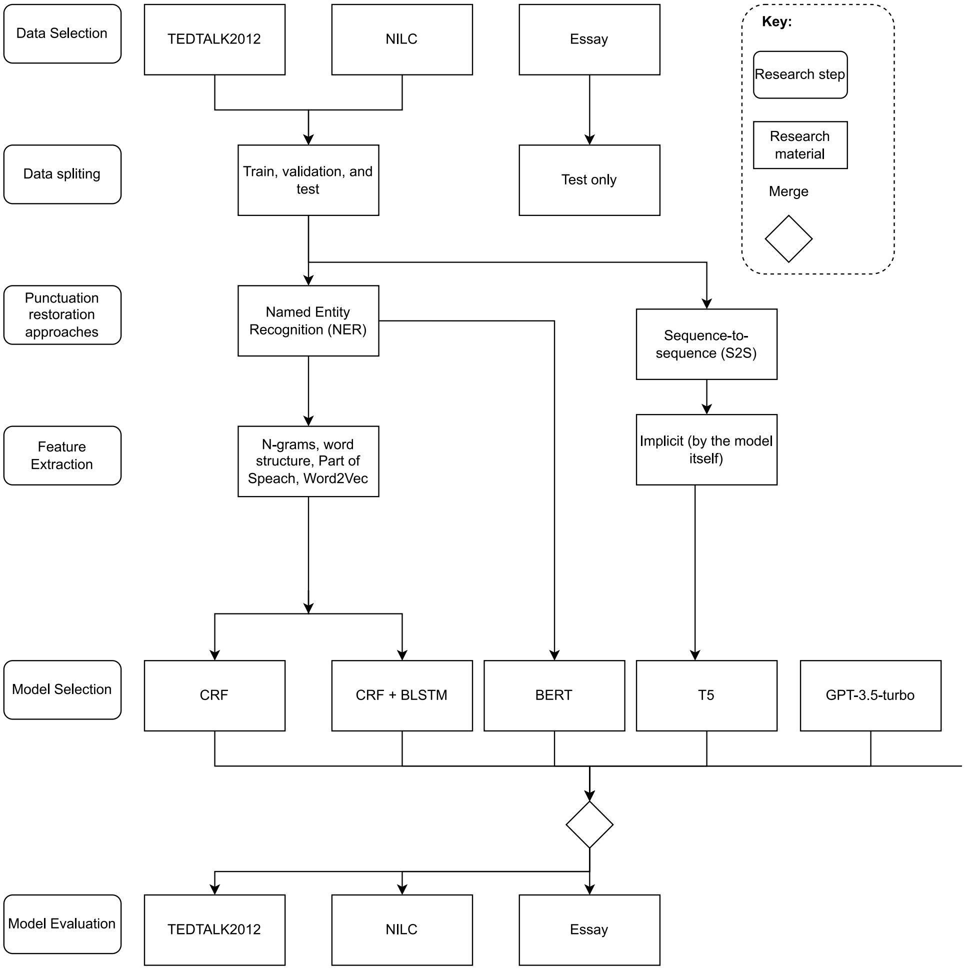


Fig. 1. Research method summary.

GPT-3 outperformed the previously leading T5 11B model in tasks like question-answering using the TriviaQA dataset (Brown et al., 2020). For this, the GPT-3 received five text sets comprising billions of non-proportionally balanced tokens, wherein some datasets were inputted multiple times depending on their quality. More recently, OpenAI released the GPT-4 language model, achieving even better results than its predecessor, GPT-3, across various NLP tasks (OpenAI, 2023). While GPT-3 ranked among the 10% worst results in some tasks, GPT-4 placed in the top 10%, even when tackling questions from professional exams like the Uniform Bar Exam in the United States and exams related to physics and psychology (OpenAI, 2023). However, GPT-4 retains some of GPT-3's limitations, such as hallucinations, context window restrictions, and the inability to learn from historical context (OpenAI, 2023).

In light of the substantial advancements in GPT models, particularly their demonstrated ability to adapt to various tasks, their increased proficiency in handling complex questions, and their overall improvements in performance, we chose to utilize these models in our study of the punctuation restoration task.

3. Research questions

The first goal of our study was to investigate the viability of using machine and deep learning methods to restore punctuation in Brazilian Portuguese. To this end, we evaluated various models, including traditional CRF and recent sequence-to-sequence models such as T5. We utilized a set of features and word embeddings to develop ten machine and deep learning models and compare their performance by adopting the same data set used by previous work (de Lima et al., 2023). Thus, the first research question in the current study was:

> **Research Question 1 (RQ1):** *To what extent can machine/deep learning models accurately predict punctuation restoration for Brazilian Portuguese texts?*

Subsequently, to assess the generalizability of our proposed models in broader contexts, we tested them with data from different domain documents in the training and testing steps. Specifically, we undertook a comprehensive cross-dataset experiment, whereby each model underwent training on a designated dataset and subsequently faced evaluation on another dataset, and vice versa. Moreover, we expanded our assessment to encompass an entirely distinct task by evaluating the models on a new dataset of student essays developed in this study. It is important to mention that we also evaluated the GPT models in this case. This comprehensive approach led us to formulate the second research question:

> **Research Question 2 (RQ2):** *To what extent are the machine/deep learning models generalizable in the punctuation restoration task for Brazilian Portuguese texts?*

Finally, the third goal of our study aimed to delve into the influence of different tokens on punctuation restoration. To our knowledge, no previous works have investigated the importance of the features for the punctuation restoration task, for the Portuguese language. In pursuing this goal, we leveraged the INSeq explainable AI technique (Sarti et al., 2023). To articulate our focus clearly, we formulated the following research question for this phase:

Research Question 3 (RQ3): *To what extent are XAI effective in highlighting relevant tokens for the punctuation restoration task in essays written in Portuguese?*

4. Method

Fig. 1 summarize this article's method, which is detailed next.

4.1. Data description

This study used TEDTALK2012 (Federico et al., 2012) and NILC (Gazzola et al., 2019) datasets for the model training process, including model training, validation, and testing. Moreover, we leveraged a third dataset comprising students' essays to evaluate these models. Given the limited number of examples, this third dataset was not employed in the model training stage, but only in the testing stage. This research focused on evaluating two primary punctuation marks: commas and periods. Other terminal punctuation marks, such as question marks and exclamation marks, were classified under the category of periods in this study, such as in Federico et al. (2012), Tilk and Alumäe (2016).

The first dataset investigated was the Portuguese version of the IWLST TEDTALK2012 dataset (Federico et al., 2012), consisting of 3482 texts with 155,787 training sentences, 1048 validation sentences, and 1886 testing sentences (details in Table 2). This database, comprising both audios and transcripts of TEDTALK2012 lectures, is used in various applications such as Automatic Speech Recognition, Machine Translation, and punctuation restoration (Federico et al., 2012; Tilk & Alumäe, 2016). This dataset is a benchmark for comparing various state-of-the-art works regarding punctuation restoration (Courtland et al., 2020; Makhija et al., 2019; Tilk & Alumäe, 2016). For this study, we chose to adhere to the original division of the IWLST TEDTALK2012 dataset, in alignment with previous literature (Tilk & Alumäe, 2016).

The second dataset, NILC data, comprises educational texts from various sources (e.g., textbooks and exam scripts), and it is categorized into four levels: Elementary School I and II, High School, and Higher Education (Gazzola et al., 2019). We selected relevant texts for Elementary Schools I and II due to the age group of the students whose compositions are being evaluated in the third dataset (Essay). At the Elementary School I level, we used news texts tailored for children between 8 and 11 years old, also utilized in Scarton and Aluísio (2010) for automatic classification of texts suitable for children and adults. For Elementary School II, the utilized texts include textbooks, exam scripts from the Brazilian Basic Education Assessment System (SAEB), E-Books, and educational resources from the National Campaign for Community Schools (CNEC). This dataset encompasses 1697 texts and 13,016 sentences with an average length of 300 to 596 words (Gazzola et al., 2019).

The last dataset for this study, referred hereafter as the Essay dataset, was obtained through an analysis of 385 essays written by middle school students in Brazilian public schools, comprising 2168 sentences. The essays were annotated/revised by two school teachers who worked at the same level as the students who produced the texts. The final agreement between the coders was 0.67 of Cohen's kappa (Cohen, 1960), and a third teacher annotated/revised the divergent essays. Table 2 synthesizes the number of instances in each dataset.

4.2. Feature extraction

As mentioned before, this study evaluated two main approaches for punctuation restoration: NER and the Sequence-to-Sequence (S2S) approach used by generative AI models. In this section, we describe the features used in the NER task, as the S2S approach did not require an explicit extraction of features. Initially, we focused on the traditional strategy of employing n-grams and word structure features (e.g., case formatting and checking for numbers in words) proposed by Lu and Ng (2010). We also incorporated Part of Speech (PoS) attributes, significantly enhancing the score restoration, as demonstrated in Shi, Wang, Wang, Li, Liu, and Lin (2021). Table 3 exemplifies a subset of these features. In contrast, our second approach to feature extraction pertained to the utilization of word embeddings. We specifically employed the Portuguese variant of Word2Vec (W2V) Embeddings with a 300-dimensional model, as outlined by Hartmann, et al. (2017). The rationale behind choosing W2V was its commendable performance in past studies (Che, Wang, Yang, & Meinel, 2016; Lima et al., 2022). Additionally, incorporating BERT solely in the embedding layer did not enhance the accuracy of the punctuation restoration models, as noted in Courtland et al. (2020), Nagy et al. (2021). Nevertheless, an exclusively BERT-based architecture was also evaluated in our study.[1]

Table 2
Number of instances per class in the datasets.

		I-COMMA	I-PERIOD	Instances
TEDTALK2012	Validation	1,068	981	2,049
	Test	2,155	1,797	3,952
	Train	157 486	151,496	308,982
	Total	160,709	154,274	314,983
NILC	Validation	1,424	1,044	18,596
	Test	3,335	2,621	44,161
	Train	11,961	9,424	44,161
	Total	16,720	13,089	106,918
Essay dataset	Test	2,215	1,434	3,649
	Total	2,215	1,434	3,649

Table 3
Example of traditional features extracted.

Name	Value
bias	1.0
word.lower()	'casa'
word[-3]	a
word[-2]	s
word.isupper()	FALSE
word.istitle()	FALSE
word.isdigit()	FALSE
postag	NOUN
postag[2]	U
word.islower()	TRUE
word[0].isupper()	FALSE
word[0].islower()	FALSE
not word[0].isalnum()	FALSE
not word.isalnum()	FALSE
word.isalpha()	TRUE

4.3. Model selection and evaluation

To the analysis of punctuation restoration as a NER task, we evaluated three approaches: (i) the CRF algorithm for sequence labelling; (ii) the BLSTM+CRF model; (iii) pre-trained models of BERT (BASE and LARGE). We opted for the CRF algorithm motivated by its successful implementation in various studies, including those in Portuguese (Lima et al., 2022), and significant outcomes in previous studies (Lu & Ng, 2010). Combined with the traditional features described earlier, the CRF is considered our *baseline*. We trained the model with 100 iterations following the methodology in Lima et al. (2022). Our second approach incorporated the BLSTM model, previously employed for punctuation restoration in English and Portuguese (Lima et al., 2022; Tilk & Alumäe, 2016). Research indicates that the most effective utilization of BLSTM involves running it for 100 epochs, coupled with a final layer of CRF (Lima et al., 2022; Tilk & Alumäe, 2016). Finally, we adopted BERT embedding in combination with an attention-based neural network model (Raffel et al., 2020), a combination that has

[1] Refer to Section 4.3 for more details.

Table 4
Prompt for GPT 3.5 and GPT 4.

Element	Text
Instruction	Act like punctuation corrector to include ‘period’ and ‘comma’ punctuation marks
Context	for Brazilian Portuguese
Output format	the following text without any other corrections: ‘text’

yielded significant results for Named Entity Recognition (NER) and text comprehension tasks (Raffel et al., 2020). In alignment with prior research, we scrutinized BERT’s BASE and LARGE variants (Lima et al., 2022; Nagy et al., 2021).

In the direction of using S2S models, we evaluated the T5 model (Carmo, Piau, Campiotti, Nogueira, & Lotufo, 2020; Wolf et al., 2020) as well as the GPT-3.5-turbo and GPT-4 language model versions released by OpenAI (Brown et al., 2020; OpenAI, 2023). The T5 model followed the same approach proposed by Courtland et al. (2020) to perform punctuation restoration using the IWLST2012 Tedtalk dataset. It was fine-tuned in TEDTALK2012 and NILC using five epochs (Makhija et al., 2019). The GPT models were used to evaluate only the Essay dataset and analysed with a zero-shot approach (Brown et al., 2020) due to the lack of resources available to run these models. Therefore, the prompt was based on the instructions for human evaluators for the Essay dataset on punctuation analysis. Following the instructions proposed in Giray (2023), we design the prompt to have (i) clear instructions, (ii) delimitation of the context, and (iii) indication of output format, as presented in Table 4. We use the OpenAI API to access the GPT-3.5-turbo models and the GPT-4 model.[2]

The assessment metrics adopted in our study include precision (P), recall (R), and F1-score (F1), measures extensively utilized in the evaluation of sequence labelling models in various research efforts, ranging from those employing CRF (Lu & Ng, 2010) to those leveraging pre-trained models (Courtland et al., 2020; Lima et al., 2022). Precision quantifies the proportion of true positives against the total, while recall gauges the fraction of the desired outcomes that the predictive model successfully identified. F1-score is a harmonic mean of these two measures. For a comprehensive evaluation, we selected the micro-average of the predictions rendered by each model, following the methodology employed in Lima et al. (2022).

To summarize, to address Research Question 1, we implemented the CRF, BLSTM+CRF, BERT (BASE and LARGE), and T5 models on the TEDTALK2012 and NILC datasets, employing them in training, validation, and testing scenarios. For Research Question 2, we evaluated the generalizability of the models when trained and evaluated in different datasets (TEDTALK2012 and NILC), and we evaluated the performance of the models, including the GPT 3.5 and 4, on the Essays dataset.

4.4. Explainable artificial intelligence

To answer research question 3, we selected the Inseq technique (Sundararajan, Taly, & Yan, 2017) because it provides insight into the input attributes that most impact the model’s output, making interpretation more straightforward in a sequence-to-sequence approach (Sundararajan et al., 2017). We demonstrate which tokens influence the model’s score restoration process most when achieving a correct prediction. More specifically, in this study, we adopted the *Integrated Gradients* approach, as outlined in Sundararajan et al. (2017), because it provides a clear and straightforward way to understand the impact of each input feature on the output generated by the model, enhancing the interpretability of the model’s decisions, especially in the context of the sequence-to-sequence approach used in the study. Consequently, leveraging this technique enabled us to highlight the words that predominantly affect the prediction process, particularly in instances of correct predictions. In our specific context, a corresponding *score* is assigned to every generated *token* to the resultant outputs. The tokens are generated such that they are followed by a punctuation label whenever it is necessary. An example for the prediction of punctuation in the text “Hello I am fine” is: Hello [COMMA] I am [O] fine [PERIOD]. [COMMA] stands for comma mark, [O] means no punctuation and [PERIOD] means final period. This scoring mechanism serves to convey the relevance of each token for a specific prediction.

5. Results

This section presents the results for each of our RQs.

5.1. RQ1: To what extent can transform models accurately predict punctuation restoration for portuguese texts?

Table 5 presents the assessment outcomes using the TEDTALK2012 dataset. Among the models assessed, the BERT LARGE model outperformed the rest regarding recall and F-score for both COMMA and PERIOD categories. In contrast, the T5 model excelled in precision. There was not a significant disparity in the performance across all models when extracting the PERIOD category. Furthermore, in the COMMA category, deep learning models, namely BERT and T5, demonstrated comparable results. These results were notably superior to traditional models, such as CRF and BILSTM+CRF.

Table 6 presents the analysis of the NILC dataset. The results showed a similar trend to that given in Table 5. The BERT LARGE and T5 models emerged as the best models evaluated, consistently delivering superior outcomes compared to the traditional approaches. When examining the performance metrics associated with the PERIOD category, the results were consistent across all the models, reaching the best-case F-score of 0.995. This uniformity underscored the robustness of each model in this particular category. Conversely, a more differentiated pattern was observed within the COMMA category. Here, the deep learning models, especially BERT LARGE and T5, displayed a notable edge, outstripping the traditional methodologies regarding the measures evaluated.

5.2. RQ2: To what extent are the transform models generalizable in the punctuation restoration task for Brazilian portuguese texts?

A comparative analysis was carried out to evaluate the generalizability of the proposed models. For this analysis, we focused on the BERT and T5 models, which reached the best results in RQ1. Initially, we evaluated the training performance on TEDTALK2012, testing performance on NILC and vice-versa. Then, we assessed the performance of models trained on the TEDTALK2012 and NILC datasets to ascertain their accuracy in predicting punctuation within the Essay dataset. Beyond these initial models, we expanded our scope to include the capabilities of GPT 3.5 and GPT 4 in undertaking the same punctuation restoration task on the Essay dataset.[3]

Table 7 presents the results of cross-domain assessments involving diverse language models against two distinct datasets: TEDTALK2012 and NILC. As mentioned, these models underwent training on one dataset and were subsequently subjected to testing on the other. In the context of training on TEDTALK2012 and then evaluating on NILC, the T5 BASE model surpassed the performance of the BERT-BASE and BERT-LARGE models in the COMMA category, achieving an F1-score of 0.787. Additionally, its overall performance attained an F1-score of 0.883, underscoring its efficacy. On the other hand, when training

[2] OPENAI API: https://platform.openai.com/docs/api-reference/authentication.

[3] Due to the costs to use GPT API we did not evaluate them in the TEDTALK2012 and NILC datasets.

Table 5
Performance of the proposed models in the TEDTALK2012 dataset in terms of precision (P), recall (R), and F-score (F).

	IWLST2012 TEDTALK								
	COMMA			PERIOD			General		
	P	R	F	P	R	F	P	R	F
CRF	0.592	0.305	0.402	**0.971**	0.988	**0.979**	0.836	0.630	0.718
BLSTM	**0.741**	0.524	0.614	0.966	**0.990**	0.978	0.869	0.746	0.803
BERT BASE	0.688	0.611	0.647	0.970	0.984	0.977	0.831	0.789	0.809
BERT LARGE	0.715	**0.634**	**0.672**	0.969	0.984	0.976	0.844	**0.801**	**0.822**
T5 BASE	0.831	0.501	0.625	0.969	0.989	**0.979**	**0.915**	0.733	0.814

Table 6
Performance of the proposed models in the NILC dataset in terms of precision (P), recall (R), and F-score (F).

	NILC								
	COMMA			PERIOD			General		
	P	R	F	P	R	F	P	R	F
CRF	0.596	0.352	0.443	**0.998**	0.991	0.995	0.832	0.645	0.727
BLSTM	0.717	0.614	0.662	0.994	0.992	0.993	0.854	0.787	0.820
BERT BASE	0.802	0.772	0.787	0.997	**0.993**	**0.995**	0.893	0.873	0.883
BERT LARGE	0.810	**0.784**	**0.797**	0.996	**0.993**	0.994	0.896	**0.880**	**0.888**
T5 BASE	**0.831**	0.747	0.787	0.993	0.991	0.992	**0.909**	0.858	0.883

Table 7
Cross-domain Analyses: Evaluation of models trained on TEDTALK2012 dataset in the NILC test dataset and vice-versa in terms of precision (P), recall (R), and F-score (F).

	Training on TEDTALK2012 and evaluating on NILC								
	COMMA			**PERIOD**			**General**		
	P	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**
BERT-BASE	0.681	0.686	0.683	**0.998**	0.990	**0.994**	0.825	0.825	0.825
BERT-LARGE	0.681	0.686	0.683	**0.998**	0.990	**0.994**	0.825	0.825	0.825
T5 BASE	**0.831**	**0.747**	**0.787**	0.995	**0.989**	0.992	**0.910**	**0.858**	**0.883**
	Training on NILC and evaluating on TEDTALK2012								
	COMMA			**PERIOD**			**General**		
	P	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**
BERT-BASE	0.688	0.611	0.647	0.970	0.984	0.977	0.831	0.789	0.809
BERT-LARGE	0.681	**0.686**	**0.683**	**0.998**	**0.990**	**0.994**	**0.825**	**0.825**	**0.825**
T5 BASE	**0.714**	0.574	0.637	0.967	0.978	0.972	0.849	0.766	0.805

on NILC and evaluating on TEDTALK2012, the BERT-LARGE model emerged as the best model, achieving an F1-score of 0.825. It is important to emphasize that these models' performance varied significantly across diverse datasets, indicating the challenge of generalizing across different text domains. Moreover, the COMMA category consistently registers comparatively lower values throughout the analysis.

The second analysis aimed to evaluate the generalizability of the models was to assess the results on the Essay dataset. As outlined in Table 8, the T5 model achieved performance with an F1-score of 0.607 and 0.546 when trained on the TEDTALK2012 and NILC dataset, respectively. Moreover, it is important to highlight the underperformance of all the models when predicting the COMMA category. Even the T5 model, despite its relative superiority, only managed to reach an F1-score of 0.205. These results, particularly for the COMMA category, demonstrate the intricate nature and inherent complexity of the texts in the Essay dataset and the difficulty of generalizing these results for Portuguese. Finally, our experiments indicate that the GPT variants underperformed, registering the least impressive results across all evaluation measures, reaching up to 0.363 for F1-score. It is important to mention that we have not evaluated different prompts, which could influence the results of these models.

5.3. RQ3: To what extent are XAI effective in highlighting relevant tokens for the punctuation restoration task in essays written in Brazilian portuguese?

To address the RQ3, we adopted the T5 model and provided a few examples of sentences using the Inseq tool, as described in the method section. We limited this analysis to the T5 model because its results excelled in precision throughout our previous, while still yielding a performance comparable to other models (e.g., CRF, BERT-base, and BERT-large) in F-score. Hence, we considered T5 able to represent the best modelling approach for this XAI analysis. This RQ aimed to identify the tokens from the input sentence that wield the most pronounced influence on the base T5 model's output.

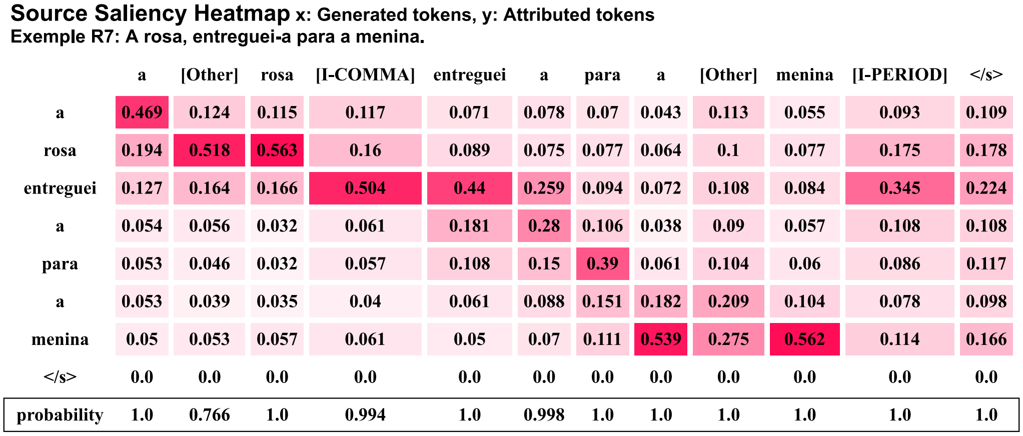
Fig. 2 presents the most relevant tokens for the prediction in stronger colours. The tags [I-COMMA] and [I-PERIOD] represent the inclusion of these tokens, and [Other] means that no punctuation was inserted in all the previous tokens. For instance, the token "entreguei" played an important role in determining the predictions for both the comma [I-COMMA] (0.504) and the period [I-PERIOD] (0.345). There was a substantial likelihood that a misspelling or omission of this word would lead the model to generate entirely distinct predictions. Moreover, it was possible to see that the token "a" is a good predictor of "menina".

Fig. 3 illustrates an example of a list of items sentence. In this case, the translation of the sentence would be "I bought rice (arroz), milk (leite), meat (carne), and chayote (chuchu)".. It is possible to see the influence surrounding tokens exerted on predicting the comma's placement. Notably, the comma's placement post the specified token was markedly influenced by the presence of the token "milk" and "meat". This phenomenon aligns with the Portuguese grammatical norms, as these words belong to the same grammatical class and, consequently, necessitate separation by a comma to adhere to standard syntactic rules.

Table 8
Results of the models evaluated on the Essay dataset in terms of precision (P), recall (R), and F-score (F).

Training on TEDTALK2012 and evaluating on Essay dataset									
		COMMA			PERIOD			GENERAL	
	P	R	F	P	R	F	P	R	F
BB	0.104	0.494	0.172	0.876	0.955	0.914	0.289	0.761	0.418
BL	0.104	**0.516**	0.173	0.878	0.955	0.915	0.284	**0.770**	0.415
T5	**0.154**	0.307	**0.205**	**0.947**	**0.967**	**0.957**	**0.523**	0.722	**0.607**
Training on NILC and evaluating on Essay dataset									
		COMMA			PERIOD			GENERAL	
	P	R	F	P	R	F	P	R	F
BB	0.124	**0.382**	0.187	0.806	0.961	0.877	0.360	0.717	0.479
BL	0.117	0.365	0.178	0.759	0.959	0.847	0.347	0.709	0.466
T5	**0.131**	0.385	**0.195**	**0.940**	**0.968**	**0.954**	**0.429**	**0.749**	**0.546**
GPT3.5 (3.5) and GPT-4 (4) models evaluating on Essay dataset									
		COMMA			PERIOD			GENERAL	
	P	R	F	P	R	F	P	R	F
3.5	0.068	0.316	0.112	0.235	0.626	0.341	0.152	0.514	0.234
4	**0.072**	**0.406**	**0.123**	**0.479**	**0.902**	**0.625**	**0.240**	**0.742**	**0.363**

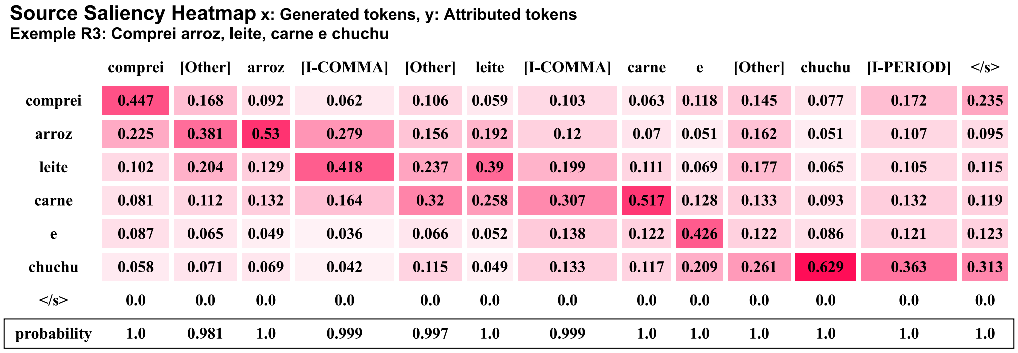
BB = BERT Base; BL = BERT Large.



Source Saliency Heatmap x: Generated tokens, y: Attributed tokens
Exemple R7: A rosa, entreguei-a para a menina.

	a	[Other]	rosa	[I-COMMA]	entreguei	a	para	a	[Other]	menina	[I-PERIOD]	</s>
a	0.469	0.124	0.115	0.117	0.071	0.078	0.07	0.043	0.113	0.055	0.093	0.109
rosa	0.194	0.518	0.563	0.16	0.089	0.075	0.077	0.064	0.1	0.077	0.175	0.178
entreguei	0.127	0.164	0.166	0.504	0.44	0.259	0.094	0.072	0.108	0.084	0.345	0.224
a	0.054	0.056	0.032	0.061	0.181	0.28	0.106	0.038	0.09	0.057	0.108	0.108
para	0.053	0.046	0.032	0.057	0.108	0.15	0.39	0.061	0.104	0.06	0.086	0.117
a	0.053	0.039	0.035	0.04	0.061	0.088	0.151	0.182	0.209	0.104	0.078	0.098
menina	0.05	0.053	0.057	0.061	0.05	0.07	0.111	0.539	0.275	0.562	0.114	0.166
</s>	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
probability	1.0	0.766	1.0	0.994	1.0	0.998	1.0	1.0	1.0	1.0	1.0	1.0

Fig. 2. Explanations of punctuation restoration with T5 model.



Source Saliency Heatmap x: Generated tokens, y: Attributed tokens
Exemple R3: Comprei arroz, leite, carne e chuchu

	comprei	[Other]	arroz	[I-COMMA]	[Other]	leite	[I-COMMA]	carne	e	[Other]	chuchu	[I-PERIOD]	</s>
comprei	0.447	0.168	0.092	0.062	0.106	0.059	0.103	0.063	0.118	0.145	0.077	0.172	0.235
arroz	0.225	0.381	0.53	0.279	0.156	0.192	0.12	0.07	0.051	0.162	0.051	0.107	0.095
leite	0.102	0.204	0.129	0.418	0.237	0.39	0.199	0.111	0.069	0.177	0.065	0.105	0.115
carne	0.081	0.112	0.132	0.164	0.32	0.258	0.307	0.517	0.128	0.133	0.093	0.132	0.119
e	0.087	0.065	0.049	0.036	0.066	0.052	0.138	0.122	0.426	0.122	0.086	0.121	0.123
chuchu	0.058	0.071	0.069	0.042	0.115	0.049	0.133	0.117	0.209	0.261	0.629	0.363	0.313
</s>	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
probability	1.0	0.981	1.0	0.999	0.997	1.0	0.999	1.0	1.0	1.0	1.0	1.0	1.0

Fig. 3. Explanations of punctuation restoration with T5 model in a sentence with a list of items.

6. Discussion

Our first research question aimed to verify the performance of traditional and sequence-to-sequence models for classifying punctuation restoration. Tables 5 and 6 compare the different models evaluated for the datasets TEDTALK2012 and NILC, respectively. Although a direct comparison with previous studies in the literature is not possible due to the differences in language and methodology, this study achieved results that stand competitively alongside other state-of-the-art contributions (Courtland et al., 2020; Lima et al., 2022; Nagy et al., 2021). Furthermore, the BERT and T5 models reached better results when compared to the traditional NER methods (CRF and BLSTM). These findings indicate that the embedding models can learn more during training and better capture relationships between words and the punctuation (Hartmann, et al., 2017; Nagy et al., 2021). It is also important to highlight that the PERIOD category consistently reached better values than the COMMA category. In the case of Brazilian Portuguese, it could be explained by the many grammatical rules to employ the comma compared to the final marks (Lenza & Martino, 2021), which could indicate that it would be required to have more data to generalize these rules.

Our second research question aimed to verify the generalizability of the assessed models. The findings demonstrated that a reciprocal analysis involving training and testing across the TEDTALK2012 and NILC datasets, and vice versa, sustained result consistency, with the peak performance of T5 reaching an F-score up to 0.883. While this angle of analysis is novel within the domain of punctuation restoration, our outcomes align closely with results from traditional single-dataset analyses (Courtland et al., 2020; Lima et al., 2022; Makhija et al., 2019; Pan et al., 2023; Păiş & Tufiş, 2021; Tilk & Alumäe, 2016). For example, the study by de Lima et al. (2023) explored various models for predicting punctuation in the TEDTALK2012 dataset and achieved an F1-score of 0.882, corroborating the competitiveness of our approach.

In an additional analysis for the second research question, we deployed models trained on the TEDTALK2012 and NILC datasets to assess punctuation restoration in a corpus of essays written by middle-school students. While results remained consistent for the 'PERIOD' category, they deviated for the 'COMMA' category. This outcome aligns with expectations, as pre-trained models do not universally generalize well to student-generated texts, which often contain errors that can influence the model's performance (Ferreira-Mello, André, Pinheiro, Costa, & Romero, 2019). As previously emphasized, Brazilian Portuguese encompasses complex grammatical rules governing comma usage (Lenza & Martino, 2021; Squarisi, 2021). Therefore, ill-structured sentences can consequently precipitate secondary errors such as punctuation misplace (Oliveira et al., 2023).

Furthermore, we expanded our evaluation in connection to the second research question to include GPT-based models specifically for the essay dataset. Neither GPT-3.5 Turbo nor GPT-4 yielded satisfactory outcomes, achieving only an F1-Score of 0.363. It should be noted that we did not explore a variety of prompts, a factor crucially linked to the output quality of this type of model (White et al., 2023). Consequently, exploring alternative prompts could substantially influence the final performance in our quantitative assessment of zero-shot approaches using GPT-3.5 Turbo and GPT-4. Additionally, a few-shot learning approach could potentially yield significant improvements in future results (Brown et al., 2020).

In conclusion, our third research question employed XAI to examine the predictions made by the T5 model on the essay dataset. Specifically, XAI elucidated how the model's outcomes correlated with Brazilian Portuguese grammatical norms. For example, Fig. 2 illustrates an instance where the verb was translocated to a different position within the sentence, and Fig. 3 reveals the alignment of terms sharing similar grammatical functions, such as nouns in this case. In both cases, the predictions are aligned with the Brazilian Portuguese grammatical norms (Lenza & Martino, 2021; Squarisi, 2021). This analysis can empower educators to offer targeted feedback on student compositions or even to develop automated systems capable of delivering timely, evaluative feedback on student essays (Cavalcanti et al., 2021).

7. Final remarks

This study evaluated a variety of machine learning and deep learning models tailored explicitly for the task of punctuation restoration in Brazilian Portuguese. It introduced the assessment of model generalizability, incorporating XAI techniques to shed light on the decision-making processes behind model predictions. Notably, GPT-based models were examined for the first time in the context of this task. Our findings suggest that the T5 model outperformed others in the evaluated scenarios.

Despite the encouraging outcomes, this study has limitations. First, following existing literature (Federico et al., 2012; Tilk & Alumäe, 2016), the task was constrained to focus on the COMMA and PERIOD categories, thereby omitting other types of punctuation equally pivotal to textual composition. Second, while the research concentrated on punctuation restoration in Brazilian Portuguese, the models could potentially be adapted for other languages to facilitate comparative analyses. Third, the study did not fully capitalize on the capabilities of large language models; it restricted its exploration to GPT-based models and did not investigate the impact of varying prompts (White et al., 2023). Lastly, another limitation lies in the absence of real-world validation of the trained models. To achieve a comprehensive research cycle, validating the tools with key stakeholders, such as teachers and students, who could employ punctuation restoration to enrich the feedback process (Cavalcanti et al., 2021).

In future research, we plan to rectify these limitations by broadening the scope of our model evaluations to include diverse datasets in content and linguistic contexts. Additionally, we aim to expand the assessment to incorporate a variety of prompts for large language models. Finally, we intend to integrate the findings into a learning analytics platform to offer automated, real-time feedback to students engaged in essay writing.

CRediT authorship contribution statement

Tiago Barbosa de Lima: Methodology, Implementation, Writing. **Vitor Rolim:** Methodology, Writing. **André C.A. Nascimento:** Writing – review & editing. **Péricles Miranda:** Writing – review & editing. **Valmir Macario:** Writing – review & editing. **Luiz Rodrigues:** Writing – review & editing. **Elyda Freitas:** Writing – review & editing. **Dragan Gašević:** Supervision, Writing – review & editing. **Rafael Ferreira Mello:** Supervision, Project administration, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgements

This research was partially funded by the Conselho Nacional de Desenvolvimento Científico e Tecnológico (310888/2021-2).

References

Alzubaidi, L., Bai, J., Al-Sabaawi, A., Santamaría, J., Albahri, A., Al-dabbagh, B. S. N., et al. (2023). A survey on deep learning tools dealing with data scarcity: definitions, challenges, solutions, tips, and applications. *Journal of Big Data*, *10*, 46.

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., et al. (2020). Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, *58*, 82–115.

Atanasova, P., Simonsen, J. G., Lioma, C., & Augenstein, I. (2020). A diagnostic study of explainability techniques for text classification. arXiv preprint arXiv:2009.13295.

Baldwin, T., Cook, P., Lui, M., MacKinlay, A., & Wang, L. (2013). How noisy social media text, how diffrnt social media sources? In *Proceedings of the sixth international joint conference on natural language processing* (pp. 356–364).

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., et al. (2020). Language models are few-shot learners. In *Advances in neural information processing systems: Vol. 33*, (pp. 1877–1901).

Carmo, D., Piau, M., Campiotti, I., Nogueira, R., & Lotufo, R. (2020). Ptt5: Pretraining and validating the t5 model on brazilian portuguese data. arXiv preprint arXiv:2008.09144.

Cavalcanti, A. P., Barbosa, A., Carvalho, R., Freitas, F., Tsai, Y. S., Gašević, D., et al. (2021). Automatic feedback in online learning environments: A systematic literature review. *Computers and Education: Artificial Intelligence*, *2*, Article 100027.

Che, X., Wang, C., Yang, H., & Meinel, C. (2016). Punctuation prediction for unsegmented transcript based on word vector. In *Proceedings of the tenth international conference on language resources and evaluation* (pp. 654–658).

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*, 37–46.

Courtland, M., Faulkner, A., & McElvain, G. (2020). Efficient automatic punctuation restoration using bidirectional transformers with robust inference. In *Proceedings of the 17th international conference on spoken language translation* (pp. 272–279).

Danilevsky, M., Qian, K., Aharonov, R., Katsis, Y., Kawas, B., & Sen, P. (2020). A survey of the state of explainable ai for natural language processing. arXiv preprint arXiv:2010.00711.

de Lima, T. B., Rodrigues, L., Macario, V., Freitas, E., & Mello, R. F. (2023). Automatic punctuation verification of school students' essay in portuguese. In *Anais do XX Encontro Nacional de Inteligência Artificial e Computacional, SBC* (pp. 58–70).

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608.

Federico, M., Cettolo, M., Bentivogli, L., Michael, P., & Sebastian, S. (2012). Overview of the iwslt 2012 evaluation campaign. In *Proceedings of the international workshop on spoken language translation* (pp. 12–33).

Ferreira-Mello, R., André, M., Pinheiro, A., Costa, E., & Romero, C. (2019). Text mining in education. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *9*, Article e1332.

Gazzola, M., Evaldo Leal, S., & Aluisio, S. M. (2019). Predição da complexidade textual de recursos educacionais abertos em português. In *Proceedings of the Brazilian symposium in information and human language technology* (pp. 61–70).

Giray, L. (2023). Prompt engineering with chatgpt: A guide for academic writers. *Annals of Biomedical Engineering*, 1–5.

Gooding, S., & Kochmar, E. (2019). Complex word identification as a sequence labelling task. In *Proceedings of the 57th annual meeting of the association for computational linguistics* (pp. 1148–1153).

Hartmann, N. S., Fonseca, E. R., Shulby, C. D., Treviso, M. V., Rodrigues, J. S., & Aluísio, S. M. (2017). Portuguese word embeddings: Evaluating on word analogies and natural language tasks. In *Anais do XI Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*. 122–131.

He, P., Liu, X., Gao, J., & Chen, W. (2020). Deberta: Decoding-enhanced bert with disentangled attention. arXiv preprint arXiv:2006.03654.

Khosravi, H., Shum, S. B., Chen, G., Conati, C., Tsai, Y. S., Kay, J., et al. (2022). Explainable artificial intelligence in education. *Computers and Education: Artificial Intelligence*, *3*, Article 100074.

Kim, B., Khanna, R., & Koyejo, O. O. (2016). Examples are not enough, learn to criticize! criticism for interpretability. In *Advances in neural information processing systems*: *Vol. 29*.

Klejch, O., Bell, P., & Renals, S. (2017). Sequence-to-sequence models for punctuated transcription combining lexical and acoustic features. In *2017 IEEE international conference on acoustics, speech and signal processing* (pp. 5700–5704). IEEE.

Kumar, V., & Boulanger, D. (2020). Explainable automated essay scoring: Deep learning really has pedagogical value. In *Frontiers in education*. Frontiers Media SA, Article 572367.

Kurup, L., Joshi, A., & Shekhokar, N. (2016). Intelligent tutoring system for learning english punctuation. In *2016 international conference on computing communication control and automation* (pp. 1–6). IEEE.

Lafferty, J. D., McCallum, A., & Pereira, F. C. N. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the eighteenth international conference on machine learning* (pp. 282–289). CA, USA: Morgan Kaufmann Publishers Inc. San Francisco.

Lenza, P., & Martino, A. (2021). *Português Esquematizado*. Saraiva Educação S.A., URL: https://books.google.com.br/books?id=QHRIEAAAQBAJ.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., et al. (2020). BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th annual meeting of the association for computational linguistics, association for computational linguistics* (pp. 7871–7880). http://dx.doi.org/10.18653/v1/2020.acl-main.703, Online. URL: https://aclanthology.org/2020.acl-main.703.

Lima, T. B. D., Miranda, P., Mello, R. F., Wenceslau, M., Bittencourt, I. I., Cordeiro, T. D., et al. (2022). Sequence labeling algorithms for punctuation restoration in brazilian portuguese texts. In *Intelligent systems: 11th Brazilian conference, BRACIS 2022, campinas, Brazil, November 28–December 1 2022, proceedings, part II* (pp. 616–630). Springer.

Lu, W., & Ng, H. T. (2010). Better punctuation prediction with dynamic conditional random fields. In *Proceedings of the 2010 conference on empirical methods in natural language processing* (pp. 177–186).

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in neural information processing systems*: *Vol. 30*, Curran Associates, Inc., URL: http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf.

Makhija, K., Ho, T. N., & Chng, E. S. (2019). Transfer learning for punctuation prediction. In *2019 Asia-Pacific signal and information processing association annual summit and conference* (pp. 268–273). IEEE.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J. R., Bethard, S., & McClosky, D. (2014). The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations* (pp. 55–60).

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, *267*, 1–38.

Nagy, A., Bial, B., & Ács, J. (2021). Automatic punctuation restoration with bert models. arXiv preprint arXiv:2101.07343.

Nielsen, M. A. (2015). *Neural networks and deep learning, vol. 25*. CA, USA: Determination press San Francisco.

Oliveira, H., Ferreira Mello, R., Barreiros Rosa, B. A., Rakovic, M., Miranda, P., Cordeiro, T., et al. (2023). Towards explainable prediction of essay cohesion in portuguese and english. In *LAK23: 13th international learning analytics and knowledge conference* (pp. 509–519).

OpenAI (2023). Gpt-4 technical report. arXiv:2303.08774.

Pan, R., García-Díaz, J. A., & Valencia-García, R. (2023). Evaluation of transformer-based models for punctuation and capitalization restoration in spanish and portuguese. In *International conference on applications of natural language to information systems* (pp. 243–256). Springer.

Paul, M., Federico, M., & Stüker, S. (2010). Overview of the iwslt 2010 evaluation campaign. In *Proceedings of the 7th international workshop on spoken language translation: evaluation campaign* (pp. 3–27).

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing* (pp. 1532–1543).

Păiş, V., & Tufiş, D. (2021). Capitalization and punctuation restoration: a survey. *Artificial Intelligence Review*, 1–42.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. (2019). Language models are unsupervised multitask learners. *OpenAI Blog*, *1*(9).

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., et al. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, *21*, 5485–5551.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1135–1144).

Sarti, G., Feldhus, N., Sickert, L., & van der Wal, O. (2023). Inseq: An interpretability toolkit for sequence generation models. arXiv preprint arXiv:2302.13942.

Scarton, C. E., & Aluísio, S. M. (2010). Análise da inteligibilidade de textos via ferramentas de processamento de língua natural: adaptando as métricas do coh-metrix para o português. *Linguamática*, *2*, 45–61.

Shi, N., Wang, W., Wang, B., Li, J., Liu, X., & Lin, Z. (2021). Incorporating external pos tagger for punctuation restoration. arXiv preprint arXiv:2106.06731.

Squarisi, D. (2021). 50 Dicas para o uso da pontuação, disponível em: Minha biblioteca.

Suliman, F. (2019). Importance of punctuation marks for writing and reading comprehension skills. *Faculty of Arts Journal*, 29–53. http://dx.doi.org/10.36602/faj.2019.n13.06.

Sundararajan, M., Taly, A., & Yan, Q. (2017). Axiomatic attribution for deep networks. In *International conference on machine learning* (pp. 3319–3328). PMLR.

Tilk, O., & Alumäe, T. (2016). Bidirectional recurrent neural network with attention mechanism for punctuation restoration. In *Interspeech* (pp. 3047–3051).

Tjoa, E., & Guan, C. (2020). A survey on explainable artificial intelligence (xai): Toward medical xai. *IEEE Transactions on Neural Networks and Learning Systems*, *32*, 4793–4813.

Valenzuela-Escárcega, M. A., Nagesh, A., & Surdeanu, M. (2018). Lightly-supervised representation learning with global interpretability. arXiv preprint arXiv:1805.11545.

Vandeghinste, V., & Guhr, O. (2023). Fullstop: Punctuation and segmentation prediction for dutch with transformers. arXiv preprint arXiv:2301.03319.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). Attention is all you need. In *Advances in neural information processing systems*, (p. 30).

White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., Gilbert, H., et al. (2023). A prompt pattern catalog to enhance prompt engineering with chatgpt. arXiv preprint arXiv:2302.11382.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., et al. (2020). Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations, association for computational linguistics* (pp. 38–45). Online. URL: https://www.aclweb.org/anthology/2020.emnlp-demos.6.

Yang, C., Rangarajan, A., & Ranka, S. (2018). Global model interpretation via recursive partitioning. In *2018 IEEE 20th international conference on high performance computing and communications IEEE 16th international conference on smart city; IEEE 4th international conference on data science and systems* (pp. 1563–1570). IEEE.