UNIVERSIDADE FEDERAL RURAL DE PERNAMBUCO

DEPARTAMENTO DE ESTATÍSTICA E INFORMÁTICA

PROGRAMA DE PÓS–GRADUAÇÃO EM INFORMÁTICA APLICADA

TIAGO BARBOSA DE LIMA

RESTAURAÇÃO DE PONTUAÇÃO EM PORTUGUÊS UTILIZANDO APRENDIZADO PROFUNDO E EXPLICABILIDADE

RECIFE – PE

2023

TIAGO BARBOSA DE LIMA

RESTAURAÇÃO DE PONTUAÇÃO EM PORTUGUÊS UTILIZANDO APRENDIZADO PROFUNDO E EXPLICABILIDADE

Dissertação submetida à Coordenação do Programa de Pós-Graduação em Informática Aplicada do Departamento de Estatística e Informática da Universidade Federal Rural de Pernambuco, como parte dos requisitos necessários para obtenção do grau de Mestre em Informática Aplicada.

ORIENTADOR: Prof. Dr. Rafael Ferreira Leite de Mello

RECIFE – PE

2023

Dados Internacionais de Catalogação na Publicação
Universidade Federal Rural de Pernambuco
Sistema Integrado de Bibliotecas
Gerada automaticamente, mediante os dados fornecidos pelo(a) autor(a)

D278r DE LIMA, TIAGO
Restauração de pontuação em Português utilizando aprendizado profundo e explicabilidade / TIAGO DE LIMA. - 2023.
146 f. : il.

Orientador: Rafael Ferreira Mello.
Inclui referências e apêndice(s).

Dissertação (Mestrado) - Universidade Federal Rural de Pernambuco, Programa de Pós-Graduação em Informática Aplicada, Recife, 2023.

1. Correção Automática de Pontuação. 2. Explicabilidade na IA. 3. Processamento de Linguagem Natural. 4. Regras de Pontuação. 5. Escrita. I. Mello, Rafael Ferreira, orient. II. Título

CDD 004

Primeiramente, dedico este trabalho ao bom Deus por me permitir chegar até aqui. Dedico também aos meus pais por todo apoio prestado, aos meus professores que ajudaram durante o processo de descoberta do conhecimento e aos meus amigos e parentes que me motivaram durante essa jornada.

Agradecimentos

Primeiramente, gostaria de agradecer a Deus por ter me concedido o conhecimento, força e energia necessária para chegar até aqui.

Meu agradecimento em especial aos meus pais por todo apoio e dedicação desde da minha infância e vida adulta, sempre me apoiando e incentivando nos estudos fornecendo aquilo que eu precisava para trilhar minha carreira acadêmica e profissional até aqui.

Gostaria de agradecer ao meu professor, o Dr. Rafael Mello, por todo apoio em empenho durante o curso da orientação, fornecendo todo suporte necessário para a realização de parte da pesquisa das mais diversas formas possíveis. Além disso, gostaria de agradecer toda a equipe do AiBoxLab que contribuíram significativamente com as publicações realizadas por mim durante este período. Em especial a Vitor Rolim pela disponibilização de APIs para realização de parte dos experimentos necessários para realização do trabalho aqui apresentado, bem como, direcionamentos importantes na condução dos experimentos. Além disso, gostaria de agradecer a Elyda Freitas pelas inúmeras correções e revisões do texto final dos artigos e da dissertação. Por fim, gostaria de mencionar o agradecimento aos amigos, colegas e familiares que torceram por mim durante esse período até a conclusão do curso.

Mas, como está escrito: As coisas que o olho não viu, e o ouvido não ouviu, e não subiram ao coração do homem, são as que Deus preparou para os que o amam.

(1 Coríntios 2:9)

Resumo

A utilização correta da pontuação, além de mostrar o domínio da língua escrita, evita ambiguidades em diferentes idiomas, como o português e inglês. No entanto, avaliar algoritmos de verificação automática de pontuação para português brasileiro, bem como, técnicas capazes de gerar uma explicação da predição de algoritmos é um tema até então pouco explorado.

A restauração de pontuação tem por objetivo melhorar a legibilidade de textos gerados automaticamente em atividades como Reconhecimento de Fala. Neste contexto, a rotulagem de sequência permite atribuirmos a uma palavra ou sub-palavra rótulos que possam ser preditos por modelos de aprendizado de máquina. Isso permite que possamos utilizar modelos de IA para predizer as pontuações de cada palavra caso seja necessário.

Modelos pré-treinados são capazes de se adaptar facilmente a um novo domínio textual com poucas épocas de aprendizado. Assim, modelos como BERT e T5 podem ser adaptados para recuperar pontuações através da rotulagem de sequência.

Apesar disso, desenvolver aplicações utilizando IA em contextos como educacionais requer atender às legislações relacionadas de confiabilidade no uso desta vigentes em diversos países. Isso levou o desenvolvimento de técnicas que permitissem avaliar as razões de uma determinada predição através das características de entrada do modelo tornando-os mais transparentes.

Dessa forma, este trabalho explora diferentes algoritmos de restauração de pontuação em Português Brasileiro. Os resultados obtidos chegam a mais de 80% de f1-score tanto em conjunto de dados tradicionais como IWLST2012 Tedtalk assim como em conjunto de dados de textos educacionais como o desenvolvido pelo núcleo interdisciplinar de linguística (NILC) da universidade de São Paulo. Além disso, adicionamos a nossa analise a avaliação em textos redigidos por alunos do ensino fundamental. Por fim, a utilização de IA explicada (XAI, sigla em inglês) a fim de avaliar a pontuação de maneira mais assertiva e se as predições dos modelos de IA estão de acordo com as regras de pontuação. Palavras-chave: Correção automática de pontuação, IA explicativa, Processamento de Linguagem Natural, Regras de Pontuação o, Escrita.

Palavras-chave: Correção automática de pontuação, *Explanaible AI*, Processamento de Linguagem Natural, Regras de Pontuação, Escrita.

Abstract

The correct use of punctuation not only shows mastery of the written language, but also avoids ambiguities in different languages, such as Portuguese and English. However, evaluating automatic punctuation checking algorithms for Brazilian Portuguese, as well as techniques capable of generating an explanation of algorithm prediction, is a topic that has not been explored to date.

Punctuation restoration aims to improve the readability of automatically generated texts in activities such as Automatic Speech Recognition. In this context, sequence labeling allows us to assign labels to a word or sub-word that can be predicted by machine learning models. This allows us to use AI models to predict the scores of each word if necessary.

Pre-trained models are able to adapt easily to a new textual domain with a few learning epochs. Thus, models such as BERT and T5 can be adapted to retrieve punctuation.

Despite this, developing applications using AI in contexts such as education requires complying with the related reliability legislation in force in various countries. This has led to the development of techniques that make it possible to evaluate the reasons for a given prediction through the model's input characteristics, making them more transparent.

This paper explores different algorithms for restoring punctuation in Brazilian Portuguese. The results obtained reach more than 80% of f1-score in both traditional datasets such as IWLST2012 Tedtalk and educational text datasets curated by the Núcleo de Lingística e Computação (NICL) from University of São Paulo. In addition, we added to our analysis the evaluation of da essays dataset written by elementary school students. Finally, Explainable AI (XAI) was used in order to evaluate the score more accurately and the models predictions were analyzed the predictions of the AI models according to the scoring rules.

Keywords: Automatic Punctuation Correction, *Explanaible AI*, Natural Language Processing, Punctuation Rules, Writing.

Lista de Figuras

Figura 1 – Consumo de recursos de GPU pelos algoritmos BI-LSTM +CRF and BERT. 26
Figura 2 – Exemplo de como a regra de pontuação 17 é entendida pelo modelo de restauração de pontuação usando o modelo do T5 base treinado com o conjunto de dados TEDTALK2012. 40
Figura 3 – Exemplo de como a regra de pontuação é entendida pelo modelo de restauração de pontuação usando o modelo T5 BASE treinado com o conjunto de dados TEDTALK201217. 41

Lista de tabelas

Tabela 1 – Exemplo do conjunto de dados TEDTALK2012: 'Então, decidi ir para minha casa.'. 16
Tabela 2 – Número de rótulos e sentenças no conjunto de dados IWSLT2012 TEDTALK. 22
Tabela 3 – Número de sentenças, palavras e rótulos no conjunto de obras literárias em Português. 23
Tabela 4 – Resultados da Precisão, Revocação e F1-Score para todas as classes de pontuação, juntamente com as métricas de média micro avaliados no conjunto de dados IWLST2012. 24
Tabela 5 – Resultados da avaliação (precisão, revocação e pontuação F1) do modelo BERT-BASE no conjunto de dados OBRAS. 25
Tabela 6 – Número final de textos, sentenças e rótulos após o pré-processamento dos conjuntos de dados NILC e redações dos alunos.. 29
Tabela 7 – Hiperparâmetros do modelo para os modelos BERT e T5. 30
Tabela 8 – A tabela mostra o resultado de todos os modelos e avaliações com o conjunto de dados NILC para Precisão (P), Revocação (R) e F1-score (F). 31
Tabela 9 – Tabela mostra o resultado para todos os modelos e medidas avaliadas no conjunto de dados MEC com as medidas Precisão (P), Revocação (R) e Pontuação F1 (F). 31
Tabela 10 – Número de exemplos em cada caso avaliado. 32
Tabela 11 – Quantidade de rótulos e o total de palavras para no conjunto de dados TEDTALK2012 e NILC. 35
Tabela 12 – Características extraídas e usadas como entrada do algoritmo CRF para a palavra "casa". A biblioteca *spacy* foi usada com o modelo para aquisição da classe morfossintática de cada palavra. 36
Tabela 13 – Os parâmetros de treinamento de cada modelo. 37
Tabela 14 – Resultados dos experimentos do conjunto de teste IWLST2012 TEDTALK, para Precisão (P), Revocação (R) e F1-score (F). 38
Tabela 15 – Resultados dos experimentos do conjunto de teste NILC, para Precisão (P), Revocação (R) e F1-score (F). 38

Tabela 16 – Resultado dos experimentos para todos os modelos treinados com ambos os conjuntos de dados e testados no conjunto de dados do MEC, considerando as métricas de Precisão (P), Revocação (R) e F1-score (F). (Os resultados foram convertidos para percentagem para fins comparativos). 39
Tabela 17 – Exemplo de predição de pontuação usando o modelo T5 base treinado com o conjunto de dados TEDTALK2012 para os exemplos relacionados a regras de pontuação. 42
Tabela 18 – A tabela mostra o número de redações em cada uma das base de dados citadas. 47

Lista de Siglas

IA	*Inteligência Artificial*
AM	*Aprendizado de Máquina*
PLN	*Processamento de Linguagem Natural*
BERT	*Bidirectional Encoder Representation*
BLSTM	*Bidirectional Long Short Term Memory*
GPT	*Generative Pretrained Transformer*
CRF	*Conditional Random Fields*
XAI	*eXplanaible AI*
PP	*Pergunta de Pesquisa*
QP	*Questão de Pesquisa*
ENEM	*Exame Nacional do Ensino Médio*
NILC	*Núcleo Interinstitucional de Linguística Computacional*
BLEU	*Bilingual Evaluation Understudy*
GPU	*Graphics Processing Unit*
IWLST	*The International Conference on Spoken Language Translation*
LLM	*Large Language Model*

Sumário

1 Introdução 14

1.1 Caracterização do Problema 15

1.2 Justificativa 17

1.3 Perguntas de Pesquisa 18

1.4 Objetivos 18

1.4.1 Objetivo Geral 18

1.4.2 Objetivos Específicos 19

1.5 Organização do Trabalho 19

2 Algoritmos de Rotulagem de Sequências para Restauração de Pontuação em Textos em Português Brasileiro 21

2.1 Design Experimental 22

2.2 Resultados 23

2.2.1 PP1: Com que precisão os modelos de restauração de pontuação conseguem predizer corretamente a pontuação de uma sentença? 23

2.2.2 PP2: Qual o desempenho do modelo em um conjunto de dados de domínio diferente daquele no qual ele foi treinado? 24

2.2.3 PP3: Qual o melhor modelo em termos de performance de treinamento? 25

2.3 Resumo 25

3 Verificação automática de pontuação em redações de estudantes em Português . 28

3.1 Design Experimental 28

3.2 Resultados 29

3.3 PP1: Com que precisão os modelos BERT e T5 conseguem predizer corretamente a pontuação de uma sentença? 30

3.4 PP2: Até que ponto o BERT e o T5 podem estimar com precisão os erros de pontuação nas produções textuais dos estudantes? 30

3.5 Resumo 32

4 Explorando a Correção Automática de Pontuação em Textos de Alunos 34

4.1 Design Experimental 35

4.2 Resultados 37

4.2.1 PP1) Com que precisão os modelos de restauração de pontuação conseguem predizer corretamente a pontuação de uma sentença de acordo com as regras da norma padrão da língua portuguesa? 37
4.2.2 PP2) Os modelos XAI são capazes de fornecer um retorno (*feedback*, termo em inglês) adequado sobre a predição de pontuação? 40
4.2.3 Análise de Erros dos Modelos de Acordo com a Regra gramatical . . . 41
4.3 Resumo 42
5 Avaliação Automática de Redação: Uma revisão sistemática 44
5.1 Design Experimental 45
5.2 Resultados 46
5.3 Resumo 48
6 Considerações Finais 49
6.1 Artigos aceitos 50
6.2 Limitações da pesquisa 50
6.3 Trabalhos Futuros 50
Referências 52
7 Apêndices 56

1 Introdução

Por meio do uso da pontuação na língua escrita o autor é capaz de repassar efetivamente a real intenção daquilo que pretende comunicar (LENZA; MARTINO, 2021). Ademais, a pontuação permite sinalizar pausas, inflexões da voz, separar expressões, entre outras funcionalidades da escrita culta, de maneira que a má utilização da pontuação demonstra a falta de domínio do idioma em que se escreve (LENZA; MARTINO, 2021). Assim, existe um amplo esforço na literatura em desenvolver modelos de aprendizado de máquina para predição automática de pontuação, a fim de melhorar a legibilidade de textos gerados artificialmente por algoritmos de reconhecimento de fala, por exemplo (PĂIŞ; TUFIŞ, 2021). Os primeiros trabalhos da área se concentravam na utilização de características do áudio para predição de pontuação (CHRISTENSEN et al., 2001). Técnicas mais recentes, no entanto, dispensam a utilização de características do áudio e utilizam apenas textos para realizar as predição de pontuação ao longo de uma sentença obtendo um resultado comparativo melhor (TILK; ALUMÄE, 2016; COURTLAND et al., 2020). Por fim, a restauração de pontuação pode ser utilizada na avaliação de textos compostos por alunos como forma de agregar a uma análise mais aprofundada como proposto por (LIMA et al., 2023b). Dessa forma, a restauração de pontuação é um processo crucial em atividades de geração textual como reconhecimento de fala. Ademais, abordagens mais recentes de predição de pontuação através apenas do texto. Isso permite a aplicação de restauração de pontuação de maneira mais ampla na correção de redações, caso abordado neste trabalho.

Para que seja possível corrigir automaticamente a pontuação em textos diversos, uma das alternativas é a utilização da predição ou restauração de pontuação. No contexto de Processamento de Linguagem Natural (PLN), a restauração de pontuação é explorada em diversos trabalhos, por exemplo, para o reconhecimento de fala e tradução da língua falada (MATUSOV et al., 2006), embora ainda não tenha sido aplicada amplamente em idiomas como Português (TILK; ALUMÄE, 2016; NAGY et al., 2021; COURTLAND et al., 2020; LIMA et al., 2023a). Além disso, a restauração de pontuação também pode ser útil na verificação automática da pontuação de textos escritos por estudantes, com a predição realizada por um modelo de restauração de pontuação, apesar de algumas limitações terem sido encontradas (LIMA et al., 2023b). Diversos algoritmos já foram utilizados na restauração de pontuação, como *Conditional Random Fields* (CRF), modelos pré-treinados como *Deep Bidirectional Transformers for Language*

Understanding (BERT) e T5, entre outros (TILK; ALUMÄE, 2016; NAGY et al., 2021; COURTLAND et al., 2020; LIMA et al., 2023a).

Contudo, os algoritmos de aprendizagem profunda utilizados na restauração de pontuação, em geral, não são transparentes devido à complexidade destes como , por exemplo, os modelos BERT e T5, mencionados anteriormente, o que torna a sua utilização um obstáculo devido aos requisitos regulatórios em diversos países (TJOA; GUAN, 2020; DANILEVSKY et al., 2020; KHOSRAVI et al., 2022). Isso tem impulsionado o desenvolvimento de técnicas de IA explicada (XAI, sigla em inglês), capazes de determinar as características das entradas que contribuíram para uma determinada predição (conhecido como explicabilidade local) ou a explicação de todo um modelo de Inteligência Artificial (IA) chamada de explicabilidade global (DANILEVSKY et al., 2020). Ferramentas de XAI como *Captum* ajudam na avaliação de diferentes níveis de coesão em textos de redação em português e inglês (OLIVEIRA et al., 2023) e verificam quais os aspectos textuais que mais influenciam na nota relacionada a esse requisito. Portanto, o aprimoramento de técnicas de IA aplicadas na área educacional deve ser acompanhado de métodos que tragam maior transparência a modelos de aprendizagem profunda.

Dessa forma, a restauração de pontuação é uma técnica amplamente utilizada a fim de melhorar a legibilidade de textos gerados por modelos generativos como em atividades de reconhecimento de fala. Diferentes técnicas têm sido propostas para endereçar o problema, com a principal delas utilizando apenas o próprio texto para fazer as predições, o que permite a restauração de pontuação ser aplicada em diferentes contextos. Contudo, os melhores algoritmos para restauração de pontuação não são transparentes na forma como realizam as predições, o que pode ser um impeditivo para aplicações mais ampla desses modelos no contexto de restauração de pontuação. Assim, este trabalho aborda de forma ampla e detalhada, diferentes algoritmos de restauração de pontuação aplicados ao português brasileiro. Por fim, utilizou-se IA explicativa para trazer mais transparência aos modelos de aprendizagem profunda aplicados à restauração de pontuação.

1.1 Caracterização do Problema

Os símbolos de pontuação permitem uma melhor compreensão do texto escrito e facilitam o pós-processamento em algoritmos de reconhecimento de fala, por exemplo (TILK; ALUMÄE, 2016). Assim, diferentes trabalhos da área como (LIMA et al., 2022; LIMA et al., 2023b; PAN

et al., 2023) utilizam técnicas capazes de recuperar a pontuação do texto predito. O problema consiste em classificar se uma determinada palavra deve ser seguida de um determinado símbolo de pontuação como ponto final (.), vírgula (,) e interrogação (?).

De modo geral, os trabalhos da área tentam resolver este problema por meio da rotulação de sequência, técnica bastante utilizada em diferentes atividades de PLN, como reconhecimento de entidades nomeadas, identificação de palavras complexas, entre outras (GOODING; KOCHMAR, 2019; DEVLIN et al., 2018; NAGY et al., 2021; COURTLAND et al., 2020). Em geral, a restauração de pontuação consiste em rotular as palavras que antecedem cada pontuação a fim de indicar a presença ou ausência de pontuação (TILK; ALUMÄE, 2016; NAGY et al., 2021; LU; NG, 2010). Dessa forma, os rótulos mais comumente utilizados são ponto final ([PERIOD] ou I-PERIOD), vírgula ([COMMA] ou I-COMMA) e ([QUESTION] ou I-QUESTION). A não pontuação é indicada por O ou [Other]. A Tabela 1 mostram um exemplo de anotação tanto para o modelo BERT como T5.

Tabela 1 – Exemplo do conjunto de dados TEDTALK2012: 'Então, decidi ir para minha casa.'.

Sentença:	**Então**	**decidi**	**ir**	**para**	**minha**	**casa**
Anotação BERT :	I-COMMA	O	O	O	O	I-PERIOD
Anotação T5:	**Então** [I-COMMA]	**decidir**	**ir**	**para minha** [Other]	**casa**	[I-PERIOD]

O conjunto de dados mais amplamente utilizado é conhecido como IWSLT2012 TEDTALK proposto por (FEDERICO et al., 2012) na qual consiste de transcrições de palestras do TEDTALK. Utilizado para treinar e avaliar diferentes modelos em diferentes trabalhos, o conjunto dados possui 142.110 sentenças. Adicionalmente, neste trabalho foi utilizado um conjunto de dados de textos gerados pelo Núcleo de Linguística e Computação da Universidade de São Paulo (NILC) (GAZZOLA SIDNEY EVALDO LEAL, 2019). Originalmente o conjunto de dados é composto por textos distribuídos em diferentes níveis de legibilidade de acordo com níveis educacionais deste do ensino fundamental I até o ensino superior. Os textos consistem em histórias para crianças e textos de ciências adaptados para o público infantil. Foram selecionados, no entanto, apenas textos referentes ao ensino fundamental I e II que contém um total de 13016 sentenças. Por fim, os conjuntos de dados IWSLT2012 e NILC foram utilizados tanto treinamento e avaliação dos modelos enquanto os demais foram utilizados apenas para propósitos de teste.

Além dos conjuntos de dados citados, textos relacionados a obras literárias em português foram utilizados para avaliação dos modelos de restauração de pontuação no capítulo 2. O

conjunto de textos com um total de 193.236 sentenças foi proposto no artigo (LIMA et al., 2022) com objetivo de avaliar a capacidade de avaliação fora do domínio em que foi treinado. Ademais, redações de alunos do ensino fundamental foram utilizadas nos capítulo 3 com 2004 sentenças e 2,168 no capítulo 4 apenas para testes devido a baixa qualidade e quantidade dos dados. Sendo assim, foram utilizados diferentes conjuntos de dados desde dos mais tradicionais como IWLST2012 Tedtalk, bem como, conjuntos de dados que permitissem uma avaliação fora do domínio de treinamento para avaliar a capacidade de generalização do modelo. Dessa forma, é possível obter-se uma avaliação dos modelos de restauração de pontuação em perspectivas mais diversas e não apenas naquelas que são tradicionalmente avaliadas.

Os algoritmos empregados podem ser tanto modelos estatísticos como CRF, mas modelos de aprendizagem profunda têm ganhado maior proeminência nos últimos anos. Os trabalhos que abordam a restauração de pontuação vão desde algoritmos mais simples, como *Conditional Ranom Fields* (CRF), até algoritmos de aprendizagem profunda, como BERT e T5 (LU; NG, 2010; COURTLAND et al., 2020). Assim sendo, a restauração de pontuação consiste em tentar prever a pontuação em um texto por meio da rotulação das palavras de acordo com a pontuação a ser empregada em cada uma delas (PĂIŞ; TUFIŞ, 2021).

1.2 Justificativa

A restauração de pontuação consiste na rotulação e subsequente predição automática de sinais de pontuação, no entanto, a maior parte dos trabalhos está direcionada à língua inglesa, com poucos trabalhos em outros idiomas como português (LIMA et al., 2022; LIMA et al., 2023b; PAN et al., 2023). Além disso, normalmente os trabalhos não investigam a capacidade dos modelos de aderirem corretamente a regras de pontuação, o que pode ser utilizado para futuras melhorias dos processos de treinamentos dos modelos, bem como da aplicação de XAI para tornar os modelos mais transparentes. (TILK; ALUMÄE, 2016; NAGY et al., 2021; LU; NG, 2010; LIMA et al., 2022; LIMA et al., 2023b).

Além disso, o desenvolvimento de modelos de inteligência artificial transparentes é um dos grandes desafios na área de Aprendizado de Máquina (SUNDARARAJAN et al., 2017). Modelos de Aprendizagem profunda, em geral, obtém os melhores resultados que os tradicionais, no entanto, identificar o processo de predição dos algoritmos pode ser um impeditivo para o uso destes em diversas áreas como educação, medicina e etc (SUNDARARAJAN et al., 2017).

Dessa forma, este trabalho, busca desenvolver um modelo de predição de pontuação confiável e transparente que pode ser utilizado como parte do pós processamento de fala em português, bem como, em outras aplicações como correção textual.

1.3 Perguntas de Pesquisa

A análise da pontuação é um processo fundamental para corrigir textos gerados automaticamente por algoritmos de reconhecimento de fala, por exemplo. Destarte, as seguintes Perguntas de Pesquisa (PP) serão endereçadas ao longo desse trabalho:

PERGUNTA DE PESQUISA 1 (PP1):
(PP1) Com que precisão os modelos de restauração de pontuação conseguem predizer corretamente a pontuação de uma sentença de acordo com as regras da norma padrão da língua portuguesa?

O objetivo da PP1 é identificar qual o melhor modelo de restauração de pontuação, baseando-se em métricas quantitativas (por exemplo, f1-score, revocação e precisão) e se eles são capazes de seguir as regras da norma culta do português.

PERGUNTA DE PESQUISA 2 (PP2):
PP2) Os modelos XAI são capazes de fornecer um retorno (feedback, termo em inglês) adequado sobre a predição de pontuação?

O objetivo da pergunta PP2 é responder se é possível obter um retorno que mostre a razão pela qual aquele rótulo foi predito. Isso seria útil particularmente em um cenário educacional onde a correção da pontuação precisa ser justificada para o estudante. Investigar também a possibilidade do fornecimento de retorno sobre a predição dos algoritmos de maneira automática.

1.4 Objetivos

1.4.1 Objetivo Geral

Predizer automaticamente a pontuação em textos em língua portuguesa de maneira transparente e explicável, através da utilização de algoritmos de aprendizagem profunda e XAI.

1.4.2 Objetivos Específicos

Para atingir o objetivo geral foram definidos os seguintes objetivos específicos:

- Treinar os modelos do estado da arte (i.e CRF, BLSTM, BERT e T5) para restauração de pontuação em textos em português.
- Avaliar os modelos com textos de língua portuguesa com métricas quantitativas (i.e f1-score, precisão e revocação).
- Criar relatório de desempenho dos algoritmos comparando a performance de cada um deles.
- Avaliar grandes modelos de linguagem (LLM, sigla em inglês) como GPT-3.5 turbo e GPT-4 com textos gerados por alunos do ensino fundamental em uma abordagem *zero-shot*.
- Utilizar IA explicativa para avaliação da predição do modelo em relação às regras de pontuação.

1.5 Organização do Trabalho

Esta dissertação foi estruturada em formato de artigos, dessa forma, cada capítulo apresenta um breve resumo da metodologia empregada para responder as perguntas de pesquisa, bem como, os principais resultados. Segue-se, portanto, a estrutura do presente trabalho: O capítulo 2 introduz o problema de restauração de pontuação em português brasileiro e aborda a restauração de pontuação em língua portuguesa, uma lacuna da área até então. Nele, os algoritmos de aprendizado de máquina, BERT base, *Bidirectional Long Short Memory* (BLSTM), e CRF são treinados e avaliados utilizando o conjunto de dados TEDTALK2012, usado amplamente na área de restauração de pontuação para treinamento e avaliação dos modelos. Adicionalmente, o trabalho avalia a utilização de recurso computacional por parte de cada um dos modelos, bem como, avalia o modelo BERT base treinado com o conjunto de dados TEDTALK2012 em um conjunto de dados de obras literárias brasileiras.

O capítulo 3 apresenta a aplicação de restauração de pontuação para corrigir os textos dos alunos de maneira automática; no trabalho, os modelos são treinados usando o conjunto de dados de textos educacionais dos níveis fundamental I e II e posteriormente testados no conjunto de dados de alunos de mesma faixa etária. A investigação mostrou que os modelos são robustos quando avaliados em sentenças bem estruturadas, contudo são suscetíveis a uma maior probabilidade de errarem em sentença mal estruturadas.

Em seguida, no Capítulo 4 efetua-se duas avaliações que não haviam sido realizadas anteriormente, a primeira está relacionada à capacidade do modelo de prever corretamente a pontuação em relação às regras de pontuação existentes. Isso é fundamental, caso o modelo seja aplicado em contextos mais sensíveis como educacional, onde a precisão do modelo em relação a um conjunto de regras é indispensável. A segunda avaliação consistiu na utilização de técnicas de IA explicada para verificar como está acontecendo a predição das pontuações. O trabalho mostrou que técnicas de IA explicativa podem ser utilizadas para esclarecer os resultados das predições dos modelos, bem como, que eles, em alguns casos, se adaptam às regras de pontuação. Ambas as avaliações, quando usadas em conjunto, podem ser aplicadas para criação de métodos de verificação automática de redação quando aliados a outros corretores gramaticais da língua portuguesa.

Além dos trabalhos mais relacionados diretamente ao tema, o capítulo 5 apresenta uma revisão da literatura relacionada à utilização da IA na correção de redação, os principais algoritmos utilizados, as métricas, os bancos de dados mais utilizados, e existe alguma evidência relacionada ao emprego da correção automática no mundo real.

Por fim o Capítulo 6 encerra o trabalho trazendo as considerações finais, limitações e trabalhos futuros.

2 Algoritmos de Rotulagem de Sequências para Restauração de Pontuação em Textos em Português Brasileiro

Considerando que a restauração de pontuação em uma sentença é uma técnica bastante aplicada na literatura, principalmente em língua inglesa, com poucos trabalhos em outros idiomas, decidiu-se investigar diferentes algoritmos aplicados na literatura que pudessem ser aplicados em língua Portuguesa (PĂIŞ; TUFIŞ, 2021). Neste trabalho avaliou-se diferentes abordagens, como a utilizada no trabalho (LU; NG, 2010), que consiste na utilização do algoritmo *Conditional Random Fields* (CRF, sigla em inglês), baseada na extração de características do texto para predição de rótulos em sequências de texto. Em seguida, avaliamos também o modelo *Bidirectional Long Short Memory* (BLSTM) utilizado no trabalho de (TILK; ALUMÄE, 2016) para predição de pontuação, considerando apenas o texto escrito sem qualquer outra extração de características. Por fim, utilizamos o modelo pré-treinado *Bidirectional Encoder Representation* (BERT-BASE), utilizado em diversas atividades de classificação e rotulação de sequência para restauração de pontuação em inglês (NAGY et al., 2021). As perguntas de pesquisa (PP) respondidas foram:

PP1: Com que precisão os modelos de restauração de pontuação conseguem predizer corretamente a pontuação de uma sentença?

Como essa pergunta busca-se encontrar o modelo que melhor se adéqua à resolução do problema de restauração de pontuação em Português, dentre as diferentes abordagens já utilizadas em inglês e outros idiomas, trazendo um melhor direcionamento para trabalhos futuros na área.

PP2: Qual o desempenho do modelo em um conjunto de dados de domínio diferente daquele no qual ele foi treinado?

Com essa pergunta, avaliou-se a capacidade do modelo de generalizar e predizer a pontuação corretamente em um conjunto de dados diferente daquele no qual foi treinado. Isso poderá ajudar na aplicação de restauração de pontuação nos mais diferentes contextos trazendo as principais vantagens e limitações existentes.

PP3: Qual o melhor modelo em termos de performance de treinamento?

Através dessa pergunta verificar-se qual modelo obtém a melhor performance de treinamento e dessa forma mais se adequa a um cenário de restrição de recurso computacional.

2.1 Design Experimental

Os modelos foram treinados com o conjunto de dados IWLST2012 TEDTALK2012 que é utilizado em diversos trabalhos, para treinamento e avaliação de modelos de restauração de pontuação (TILK; ALUMÄE, 2016; NAGY et al., 2021; COURTLAND et al., 2020). Nós o dividimos em 139.653 sentenças para treinamento, 1.570 para validação e 887 para teste, como apresentado na Tabela 2. Utilizou-se também o conjunto de dados de obras literárias em português brasileiro para verificar a capacidade do modelo de prever em um domínio diferente do originalmente treinado (ver tabela 3). Apesar de existir um desbalanceamento entre as classes utilizados no treinamentos, os modelos obtiveram um resultados significativo. Ademais, até onde foi nossa pesquisa, não encontramos abordagens capazes de solucionar efetivamente tal problema.

Tabela 2 – Número de rótulos e sentenças no conjunto de dados IWSLT2012 TEDTALK.

Labels	**TRAIN**	**DEV**	**TEST**
O	1.929.873	14.069	22.208
VÍRGULA	169.384	1.169	2.270
PONTO FINAL	147.379	935	1.721
INTERROGAÇÃO	11.595	87	152
Sentenças	139.653	1.570	887
Palavras	2.258.231	16.260	26.351

Para o modelo CRF, escolheu-se as características como sufixo, palavras anteriores e posteriores, bem como, etiquetagem morfossintática. Também executou-se o modelo BERT-BASE por 12 épocas e BLSTM por 100 iterações com parada antecipada (*early-stopping*, termo em inglês), assim como realizado em outros trabalhos da literatura (TILK; ALUMÄE, 2016; CHE et al., 2016). Utilizamos uma representação vetorial pré-treinados para melhorar o nosso resultado ainda mais. Diferentemente de outros trabalhos que usaram o Vetores Globais para Representação de Palavras (glove, sigla em inglês), utilizou-se um word2vec pré-treinado skip-gram com 300 dimensões em conjunto com o modelo BLSTM, a fim de melhorar ainda mais nossos resultados. As métricas de avaliação escolhidas foram f1-score, precisão e revocação, aplicadas também em outros trabalhos como (MAKHIJA et al., 2019; TILK; ALUMÄE, 2016;

Tabela 3 – Número de sentenças, palavras e rótulos no conjunto de obras literárias em Português.

Rótulos	**Quantidade**
O	2.298.811
Vírgula	303.424
Ponto Final	202.573
Interrogação	15.380
Sentenças	193.236
Palavras	2.820.188

COURTLAND et al., 2020). Para avaliação do recurso computacional gasto durante o processo de treinamento utilizaremos a ferramenta Wandb (BIEWALD, 2020).

2.2 Resultados

Os resultados foram promissores e mostraram que o modelo pode, em última instância, ser aplicado em um contexto diferente e manter uma performance adequada. Os experimentos do modelo pré-treinado BERT-BASE pode obteve 81% do f1-score quando avaliado no conjunto de teste do TEDTALK2012, obtendo 73.5% quando avaliado no conjunto de dados de obras literárias.

2.2.1 PP1: Com que precisão os modelos de restauração de pontuação conseguem predizer corretamente a pontuação de uma sentença?

Apesar de no geral o modelo BERT-BASE obter os melhores resultados, ele é superado pelo BLSTM+CRF em relação a precisão do ponto final e é comparável em relação à vírgula. Apesar de ser superado por ambos os modelos, BERT-BASE e BLSTM+CRF, o algoritmo CRF obteve resultado comparável em relação à precisão da predição da vírgula (ver tabela 4).

Os resultados mostram que os algoritmos baseados em redes neurais são superiores ao algoritmo CRF, *baseline*, em relação à restauração de pontuação. O algoritmo CRF mostra um resultado aquém em relação à utilização da vírgula (0,39 f1-score), e mais ainda em relação à interrogação (0,08 f1-score), um padrão também apresentado nas outras abordagens em menor

Tabela 4 – Resultados da Precisão, Revocação e F1-Score para todas as classes de pontuação, juntamente com as métricas de média micro avaliados no conjunto de dados IWLST2012.

Modelo	VÍRGULA			PONTO FINAL			INTERROGAÇÃO			Média Micro
	P	R	F1	P	R	F1	P	R	F1	F1
CRF	0,556	0,306	0,395	0,869	0,836	0,852	0,318	0,046	0,080	0,614
BI-LSTM+Skip$_s$300	0,670	0,530	0,592	**0,924**	0,842	0,881	0,750	0,572	0,649	0,724
BERT-BASE	**0,770**	**0,719**	**0,744**	0,911	**0,887**	**0,899**	**0,844**	**0,711**	**0,771**	**0,810**

grau. Um dos aparentes motivos para isso é o fato de termos mais exemplos para ponto final do que para vírgula e menos de um décimo de exemplos para interrogação, como apresentado na Tabela 2.

Uma das formas de mitigar esse problema é permitindo que mais dados sejam deixados para treinamento em uma avaliação futura. Quanto à abordagem utilizando redes neurais artificiais, observou-se que o modelo BERT-BASE obtém um resultado superior ao modelo BLSTM, mesmo com a utilização de incorporação (*embeddings*, termo em inglês) pré-treinados. Isso também já foi observado em outros trabalhos, como os de (COURTLAND et al., 2020), dado que o modelo BERT consegue representar eficientemente relações semânticas entre as palavras (DEVLIN et al., 2018).

2.2.2 PP2: Qual o desempenho do modelo em um conjunto de dados de domínio diferente daquele no qual ele foi treinado?

A Tabela 5 mostra os resultados do modelo treinado no conjunto de obras literárias em português. Apesar de uma queda na revocação do ponto final, o modelo mantém um resultado médio de 0,735, mantendo um resultado de f1-score de 0.871 em relação à classe ponto final.

Por fim, o modelo BERT-BASE avaliado usando o conjunto de dados de obras literárias em português mostrou resultados satisfatórios mesmo não tendo sido treinado para o conjunto de dados especifico, o que indica um potencial de utilização para outras atividades, como correção textual.

Tabela 5 – Resultados da avaliação (precisão, revocação e pontuação F1) do modelo BERT-BASE no conjunto de dados OBRAS.

Classe	Precisão	Revocação	F1
VÍRGULA	0,697	0,608	0,649
PONTO FINAL	0,877	0,865	0,871
INTERROGAÇÃO	0,626	0,427	0,508
Médias Micro	0,771	0,703	0,735

2.2.3 PP3: Qual o melhor modelo em termos de performance de treinamento?

A Figura 1 mostra como os algoritmos se comportam durante o treinamento. O algoritmo BERT-BASE também tem um alto consumo de memória e tempo de acesso a esta. É possível ver que o modelo BERT-BASE consome muito mais recurso computacional em termos de GPU do que os demais modelos BLSTM e o algoritmo CRF, fazendo com que estes últimos sejam mais adequados à restrição de uso de GPU, por exemplo.

2.3 Resumo

Nesse capítulo, investigou-se três modelos principais para restauração de pontuação em Português brasileiro, CRF, BLSTM+CRF e BERT-BASE, como um trabalho inicial de pesquisa. Através dos resultados de F1-score, precisão e revocação podemos dizer que o modelo BERT-BASE obtém os melhores resultados na avaliação do conjunto de teste proposto e um resultado consistente quando avaliado fora do domínio treinado, obtendo uma média de f1-score de quase 9 em relação ao modelo BLSTM+CRF. Isso indica que pode-se obter resultados significativos mesmo avaliando os modelos em outro domínio, o que é fundamental para aplicação no contexto educacional onde dados rotulados nem sempre são encontrados. Além disso, ao avaliar-se o custo computacional empregado por cada modelo fica evidente que o modelo BERT-BASE consome mais recurso do que os demais. Melhorias podem ser realizadas no futuro como avaliação de uma gama mais ampla de modelos pre-treinados como BERT GRANDE e T5-BASE e GRANDE. Dessa forma, o próximo capítulo dessa dissertação abordar um conjunto mais ampla de modelos como BERT GRANDE e T5 BASE e GRANDE além do BERT base usado neste trabalho. Também, este trabalho não avalia os modelos em um conjunto educacional e nem em textos

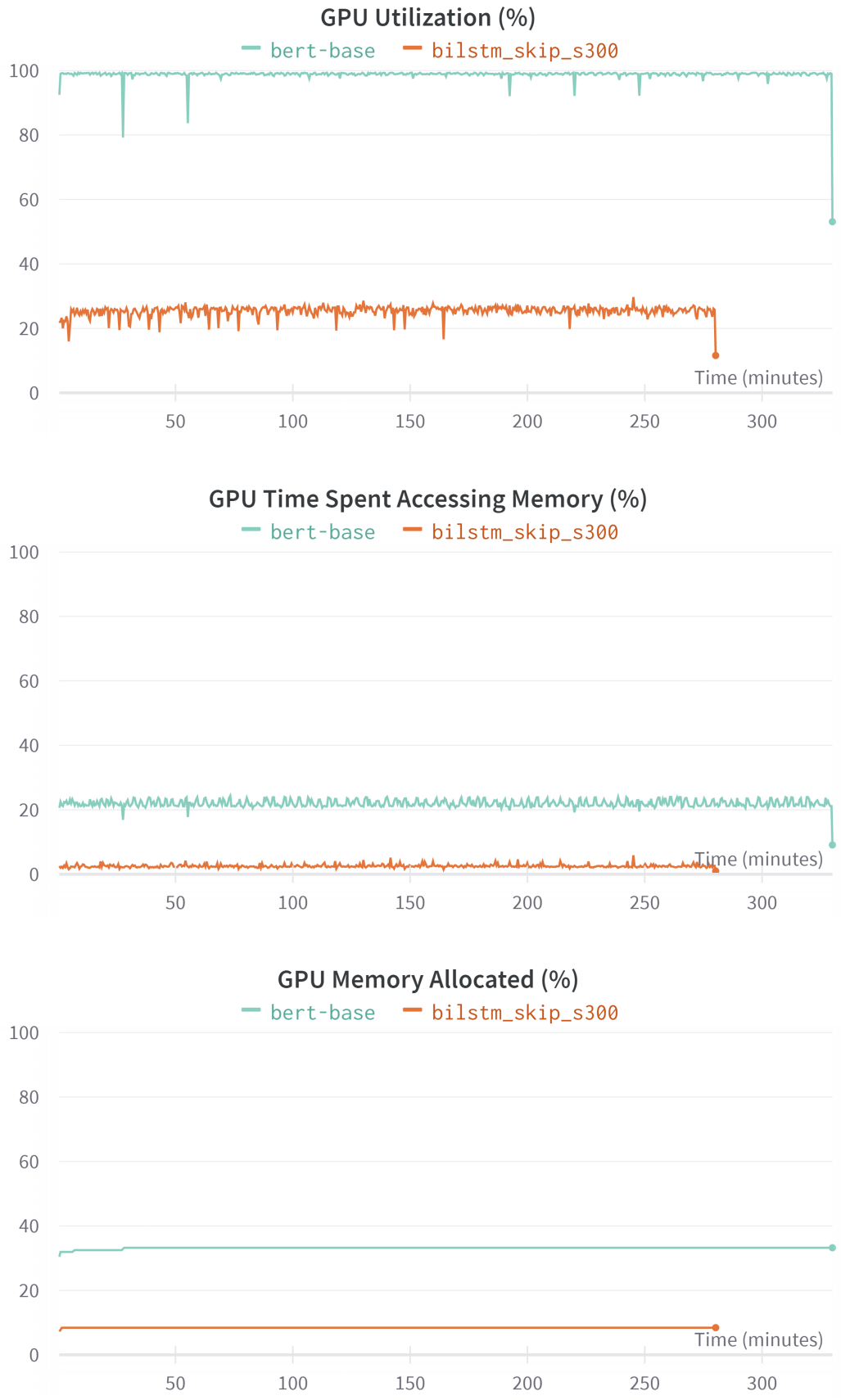

Figura 1 – Consumo de recursos de GPU pelos algoritmos BI-LSTM +CRF and BERT.

gerados por alunos o que será realizado no próximo capítulo. Por fim, como forma de aplicação, usaremos textos de alunos para avaliação da escrita em relação a pontuação.

3 Verificação automática de pontuação em redações de estudantes em Português

Neste capítulo, aprofundou-se a discussão em relação à aplicação da restauração de pontuação no contexto educacional, avaliando mais amplamente o impacto da aplicação de modelos pré-treinados na restauração de pontuação em textos de alunos do ensino fundamental. O objetivo, portanto, consiste em avaliar qual algoritmo tem melhor desempenho quando avaliado em domínio diferente daquele no qual foi treinado, assim como realizado no trabalho (LIMA et al., 2022), porém, avaliando o desempenho do modelo em textos escritos por alunos do ensino fundamental. Por isso, foi utilizado o conjunto de dados de textos de histórias infantis para diversas faixas etárias, desde do ensino fundamental I até o ensino superior, organizados pelo Núcleo Interinstitucional de Linguística Computacional (NILC). Assim sendo, nossas perguntas de pesquisa são:

PP1: Com que precisão os modelos BERT e T5 conseguem predizer corretamente a pontuação de uma sentença?

Ao responder essa pergunta, ficará mais evidente qual dos modelos pré-treinados podem prever corretamente a pontuação, avaliando em termos de precisão, revocação e f1-score.

PP2: Até que ponto os algoritmos BERT e o T5 podem estimar com precisão os erros de pontuação nas produções textuais dos estudantes?

Na segunda pergunta (PP2), avalia-se a capacidade dos modelos de serem utilizados como corretores gramaticais de maneira eficiente em textos de alunos do ensino fundamental.

3.1 Design Experimental

No trabalho foi utilizado os modelos BERT na versão BASE e GRANDE e os modelos T5 na versão BASE e GRANDE.

O primeiro conjunto de dados utilizado consiste em 1.695 textos separados em 13.016 sentenças de pequenas histórias, reportagens adaptadas para crianças ou textos científicos descritos em diferentes faixas etárias, originalmente utilizado para realização da classificação do nível de legibilidade para fins educacionais (GAZZOLA SIDNEY EVALDO LEAL, 2019) (ver tabela 6). Os textos vão desde do nível fundamental I até o nível médio; foram escolhidos, no entanto, para os experimentos, apenas com o conjunto de dados do nível fundamental I e II

pelo fato de abranger as faixas etárias do conjunto de textos gerados por alunos usados como conjunto de teste. Para permitir uma posterior avaliação com o conjunto de testes de redação dos alunos os símbolos como ';', '!', '?' foram considerados como sendo ponto final. Este conjunto de dados foi escolhido devido à categorização deste em diversos níveis educacionais, o que torna possível selecionarmos textos mais próximos em termos de legibilidade daqueles redigidos por crianças que elaboraram os textos que iremos utilizar na avaliação (GAZZOLA SIDNEY EVALDO LEAL, 2019) .

O segundo conjunto de dados utilizado é composto por 256 redações separadas em 2004 sentenças escritas por alunos do ensino fundamental, anotadas em relação ao uso do ponto fina e da vírgula, devido ao pouco número de exemplos, deixamos esse conjunto de dados apenas para testes veja a tabela 6.

Tabela 6 – Número final de textos, sentenças e rótulos após o pré-processamento dos conjuntos de dados NILC e redações dos alunos..

divisão	Número de Textos	Número de Sentenças	Sentenças Elementar I	Sentenças Elementar II	I-PONTO	I-VÍRGULA
treino	613	9371	4898	4473	11961	9424
teste	597	2604	1361	1243	2621	3335
validação	485	1041	544	497	1424	1044
Total	1695	13016	6803	6213	16006	13803
MEC	256	2004	-	-	2004	1082

Os modelos foram treinados na plataforma *Google Colab* com uma T4 Tesla GPU. Os hiperparâmetros para treinamento podem ser vistos na tabela 7.

As métricas de avaliação foram as mesmas utilizadas no trabalho anterior (LIMA et al., 2022): f1-score, precisão e revocação, com acréscimo da métrica BLEU score (bilingual evaluation understudy) para avaliação do modelo T5 no conjunto de validação durante o processo de treinamento (PAPINENI et al., 2002).

3.2 Resultados

Nesta seção, apresentou-se os resultados dos modelos treinados, validados e testados com o conjunto de dados do NILC, e posteriormente apenas testados com o conjunto de dados

Tabela 7 – Hiperparâmetros do modelo para os modelos BERT e T5.

Parâmetro	**BERT**	**T5**
Taxa de aprendizado	5,00e-5	5,00e-5
Tamanho do batch de treino	8	2
Tamanho do batch de avaliação	8	2
Semente	42	42
Otimizador	Adam com betas=(0,9,0,999) epsilon=1e-08	Adam com betas=(0,9,0,999) epsilon=1e-08
Tipo de programador de LR	linear	linear
Número de épocas	5	5

das redações dos alunos.

As seguintes perguntas de pesquisa foram elaboradas para o desenvolvimento do trabalho:

3.3 PP1: Com que precisão os modelos BERT e T5 conseguem predizer corretamente a pontuação de uma sentença?

Nesta primeira avaliação, a Tabela 8 mostra que o modelo T5 GRANDE obteve o melhor resultado na média entre todos os modelos, ficando apenas um pouco acima do T5 BASE. Outro ponto de destaque na Tabela 8 é o modelo BERT GRANDE, que obteve um resultado média similar ao T5 BASE. Por fim, o modelo T5 tanto BASE como GRANDE tem uma melhor previsão do que os modelos BERT BASE e GRANDE, enquanto estes têm um melhor resultado médio de revocação. Portanto, ambos os modelos apresentam um resultado equilibrado com uma pequena vantagem para o T5.

3.4 PP2: Até que ponto o BERT e o T5 podem estimar com precisão os erros de pontuação nas produções textuais dos estudantes?

Essa segunda pergunta de pesquisa responde sobre a viabilidade de utilizar modelos de restauração de pontuação para correção de textos de alunos. Na Tabela 9 podemos ver que o modelo BERT-GRANDE obtém o melhor resultado para avaliação de textos de alunos quando comparado com o modelo T5, em ambos os rótulos, ponto e vírgula. Apesar do modelo BERT-GRANDE obter um valor médio de 0,727, o resultado para vírgula é apenas de 0,186, o

Tabela 8 – A tabela mostra o resultado de todos os modelos e avaliações com o conjunto de dados NILC para Precisão (P), Revocação (R) e F1-score (F).

	BERT BASE			**BERT GRANDE**		
	P	**R**	**F**	**P**	**R**	**F**
COMMA	0.802	0.772	0.787	0.81	0.784	0.797
PERIOD	**0.997**	0.993	**0.995**	0.996	0.993	0.994
AVG	0.891	0.873	0.882	0.895	0.88	0.887
	T5 BASE			**T5 GRANDE**		
	P	**R**	**F**	**P**	**R**	**F**
COMMA	0.831	0.747	0.787	0.842	0.762	0.8
PERIOD	0.995	**0.989**	0.992	**0.998**	**0.994**	**0.996**
AVG	0.906	0.858	0.88	0.914	0.868	0.89

que mostra uma certa dificultado quanto à predição do rotulo em relação ao conjunto de dados anteriormente testado. O padrão se repete tanto para o modelo BERT-BASE quanto para os demais modelos T5 BASE e GRANDE.

Tabela 9 – Tabela mostra o resultado para todos os modelos e medidas avaliadas no conjunto de dados MEC com as medidas Precisão (P), Revocação (R) e Pontuação F1 (F).

	BERT BASE			**BERT GRANDE**		
	P	**R**	**F**	**P**	**R**	**F**
VÍRGULA	0,12	0,368	0,181	0,123	0,381	0,186
PONTO	**0,984**	**0,999**	**0,991**	0,97	0,996	0,983
MÉDIA	0,707	0,797	0,732	0,698	0,799	0,727
	T5 BASE			**T5 GRANDE**		
	P	**R**	**F**	**P**	**R**	**F**
VÍRGULA	0,049	0,126	0,07	0,047	0,139	0,07
PONTO	**0,8**	0,009	0,018	**0,697**	0,011	0,021
MÉDIA	0,603	**0,04**	**0,032**	0,527	**0,044**	**0,034**

Dessa forma, irá-se realizar uma avaliação quantitativa através das métricas tradicionais, bem como, uma avaliação qualitativa em 3 casos de teste: 1) quando o modelo prevê uma quantidade errada de número de rótulos e coloca a vírgula ou ponto no lugar incorreto; 2) quando

o modelo prevê a mesma quantidade de pontos e vírgulas esperadas na sentença, mas coloca no lugar errado; por fim 3) quando o modelo prevê corretamente tanto a quantidade de rótulos quando lugar de cada um deles. Os resultados foram reunidos e estão resumidos na Tabela 10.

Tabela 10 – Número de exemplos em cada caso avaliado.

Caso de Teste	**Número de Pontuações**	**Proporção**
1	237	54,11%
2	15	3,42%
3	186	42,47%
Total	438	100%

Na Tabela 10 podemos ver que para maioria dos casos o modelo prevê uma quantidade incorreta de rótulos e em um lugar incorreto.

3.5 Resumo

Neste trabalho, foi investigada a possibilidade da utilização de restauração de pontuação em textos em Português Brasileiro. Para isso, o modelo foi treinado e avaliado em um conjunto de dados educacional com objetivo de ter um modelo ajustado ao tipo de teste mais provável a ser corrigido. Os modelos que apresentaram melhores resultados foram o T5 BASE e GRANDE, que obteve um f1-score médio de 0,996. Apesar disso, o modelo tem um desempenho aquém quando avaliado com o conjunto de dados de textos produzidos pelos alunos. Neste quesito, o modelo BERT-GRANDE obteve o melhor resultado, com média de 0,727. Apesar disso, o resultado é influenciado principalmente pelo resultado do ponto final, dado que o resultado referente a vírgula foi de 0,186. Um dos motivos por trás disso pode ser o fato dos textos dos alunos conterem erros gramaticais. Uma das formas de mitigar o problema é tornar o processo mais transparente através do uso da IA explicada, por exemplo. Portanto, o modelo pode ser utilizado para fins educacionais, contudo, é necessário uma investigação mais aprofundada de como mitigar os possíveis erros do modelo, devido à estrutura pobre dos textos, e torná-los robustos aos erros gramaticais mais graves. Assim sendo, a fim de endereçar as limitações desse capítulo, o próximo trabalho sumarizado nas próximas seções aborda técnicas como Inteligência Artificial Explicada para entender melhor o processo de predição dos melhores modelos e

fornecer possíveis conclusões em relação as predições dos modelos. Além disso, será o melhor modelo avaliado através de métricas quantitativas será avaliado quanto as regras gramaticais apresentados em livros de língua portuguesa. Por fim, avaliaremos uma gama maior de modelos e conjuntos de dados avaliando todos os modelos apresentados no capítulo 1 dessa dissertação e com todos os modelos do presente capítulo com exceção do modelo T5 GRANDE e do conjunto de dados de obras literárias. Ademais, o conjunto de dados de redações será incrementados com novas sentenças ampliando assim o escopo de avaliação.

4 Explorando a Correção Automática de Pontuação em Textos de Alunos

Neste capítulo, procurou-se aprofundar o conhecimento sobre a utilização dos modelos de restauração de pontuação os quais não conseguem obter um resultado satisfatório quando avaliados em relação ao rótulo da vírgula em textos de alunos do ensino fundamental. Dessa forma, foram construídas estratégias para mitigar as predições incorretas do melhor modelo avaliado no conjunto de redações escritas por alunos através das pergntas de pesquisas (PP) a seguir:

PP1: Com que precisão os modelos de restauração de pontuação conseguem predizer corretamente a pontuação de uma sentença de acordo com as regras da norma padrão da língua portuguesa?

Através dessa pergunta de pesquisa responde-se o quão preciso são os modelos de IA, adicionando desta feita a avaliação das regras gramaticais de acordo com o proposto no livro (LENZA; MARTINO, 2021).

PP2: Os modelos XAI são capazes de fornecer um retorno (*feedback*, termo em inglês) adequado sobre a predição de pontuação?

Na segunda pergunta de pesquisa, avaliou-se o impacto da aplicação de IA explicativa no processo de predição dos modelos, a fim de melhorar a transparência dos mesmos, por meio de destaques adicionados às palavras que mais contribuem para cada rótulo.

Para responder a PP1 foi realizada uma avaliação tradicional, quantitativa, empregando as métricas utilizadas anteriormente, como f1-score, precisão e revocação. Nós acrescentamos mais modelos e conjunto de dados a serem utilizados no treinamento e validação destes (LIMA et al., 2022). Em resposta a PP2 o melhor modelo foi avaliado em um conjunto de exemplos relacionados a 8 regras gramaticas exibidas no livro (LENZA; MARTINO, 2021). Por fim, ferramentas de IA explicativa permitem que possamos obter uma maior transparência quanto ao rótulo predito, ajudando no processo de predição do modelo, o que pode contribuir para futuras melhorias, bem como, no processo de correção propriamente dito (KUMAR; BOULANGER, 2020).

4.1 Design Experimental

Os experimentos foram realizados com os conjuntos de dados já mencionados nos capítulos 2 e 3. O primeiro é o conjunto de dados IWLST2012 TEDTALK, que contém palestra de diversos temas propostos por (FEDERICO et al., 2012). O segundo é composto por textos organizados por (GAZZOLA SIDNEY EVALDO LEAL, 2019) para classificação de legibilidade de textos para fins educacionais mencionado nesse trabalho como conjunto de dados do NILC. Apesar do conjunto original de dados do NILC conter textos de histórias para crianças, adaptações de reportagens entre outros gêneros em diferentes níveis de legibilidade que vão do ensino fundamental I até o ensino superior, optamos por utilizar apenas os textos do ensino fundamental I e II, por serem mais adequados à faixa etária dos textos a serem avaliados. Por fim, utilizou-se o conjunto de dados de redações escrito pelos alunos empregado no capítulo 3, acrescido alguns exemplos por meio da coleta de mais redações anotadas, o que permitiu avaliar o potencial de cada modelo mais amplamente. Diferentemente dos demais conjuntos de dados que são utilizados para treinamento e teste, o conjunto de redações dos estudantes será utilizado apenas para teste, devido à baixa quantidade de exemplos e qualidade dos textos gerados. A quantidade de palavras para cada um dos conjuntos de dados pode ser vista na Tabela 11.

Tabela 11 – Quantidade de rótulos e o total de palavras para no conjunto de dados TEDTALK2012 e NILC.

		I-COMMA	**I-PERIOD**	**Total palavras**
TEDTALK2012	validation	1068	981	2049
	test	2155	1797	3952
	train	157486	151496	308982
	TOTAL	160709	154274	314983
NILC	validation	1424	1044	18596
	test	3335	2621	44161
	train	11961	9424	44161
	TOTAL	16720	13089	106918
MEC	**TOTAL**	2215	1622	52694

Além disso, três tipos de modelos distintos foram empregados em nosso experimentos: o primeiro é o CRF, baseado em extração de características, que vão desde sufixo até o etiquetagem

morfossintática através da biblioteca *spacy* e da linguagem de programação *Python*; sendo utilizadas como características apresentadas na Tabela 12. O algoritmo foi iterado por 100 vezes, assim como realizado no trabalho (LIMA et al., 2022).

Tabela 12 – Características extraídas e usadas como entrada do algoritmo CRF para a palavra "casa". A biblioteca *spacy* foi usada com o modelo para aquisição da classe morfossintática de cada palavra.

Nome	**Valor**
bias	1.0
word.lower()	'casa'
word[-3]	a
word[-2]	s
word.isupper()	FALSO
word.istitle()	FALSO
word.isdigit()	Falso
postag	NOUN
postag[2]	U
word.islower()	VERDADEIRO
word[0].isupper()	FALSO
word[0].islower()	FALSO
not word[0].isalnum()	FALSO
not word.isalnum()	FALSO
word.isalpha()	VERDADEIRO

Em seguida, utilizou-se uma rede neural BLSTM, utilizando o word2vec como incorporação (*embedding*, termo em inglês) de 300 dimensões com uma camada final CRF, como no trabalho de (LIMA et al., 2022). Ademais, foram avaliados os mesmos modelos apresentados no capítulo 2 e 3, com exceção do T5 GRANDE. Logo, além dos modelos já citados, treinamos e avaliamos os modelos pré-treinados BERT BASE, BERT GRANDE e T5 BASE. Os parâmetros de treinamento de cada modelo podem ser vistos na Tabela 13.

Por fim, os modelos GPT-3.5 turbo e GPT-4 da openAI retornaram os resultados através de uma abordagem utilizando em que nenhum exemplo pe repassado para o modelo (*zero-shot*, termo em inglês), quando nenhum exemplo é apresentado para o modelo (BROWN et al., 2020). Elaboramos um *prompt* (termo em inglês) capaz de retornar apenas a sentença corrigida em

Tabela 13 – Os parâmetros de treinamento de cada modelo.

	NILC		**IWST2012 TEDTALK**	
	batch size	Num. Epochs	batch size	Num. Epochs
BERT BASE	8	5	8	1
BERT GRANDE	8	5	8	1
BLSTM+skipgram	4	100	2	100
T5 BASE	2	1	1	1

relação à pontuação, acrescentando apenas pontuação final e vírgula, agindo como um corretor de pontuação em português brasileiro:

prompt: *f"Act like punctuation corrector in brazilian portuguese: Place the 'period' and 'comma' punctuation marks in the following sentence without any other corrections: 'sentence' "*

4.2 Resultados

Esta seção apresenta os resultados de acordo com cada pergunta de pesquisa.

4.2.1 PP1) Com que precisão os modelos de restauração de pontuação conseguem predizer corretamente a pontuação de uma sentença de acordo com as regras da norma padrão da língua portuguesa?

Na Tabela 14 podemos ver que o modelo BERT-GRANDE obtém o melhor resultado de f1-score para vírgula (0,67), mesmo que não tenha o melhor resultado para precisão, com o modelo T5-BASE atingindo 0,83 neste quesito. Quanto à predição do ponto final, os modelos obtêm um resultado melhor no geral com CRF e T5 BASE, alcançando um resultado superior os demais modelos com 0,979. Por fim, na média geral, o modelo BERT-GRANDE obtém o melhor resultado para f1-score, com 0,822.

Quanto à Tabela 15, verificou-se que novamente que o modelo BERT-GRANDE obtém o melhor resultado para a predição da vírgula no conjunto de dados NILC, com f1-score de 0,797. Além disso, o modelo mantém um resultado bem próximo dos modelos BERT-BASE e do algoritmo CRF, que obtiveram o melhor resultado para ponto final com 0,995. Por fim, o modelo BERT-GRANDE obtém o melhor resultado em média para ambos os rótulos, com f1-score de

Tabela 14 – Resultados dos experimentos do conjunto de teste IWLST2012 TEDTALK, para Precisão (P), Revocação (R) e F1-score (F).

	IWLST2012 TEDTALK								
	VÍRGULA			**PONTO**			**Geral**		
	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+Skipgrams	**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 GRANDE	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

0,888.

Tabela 15 – Resultados dos experimentos do conjunto de teste NILC, para Precisão (P), Revocação (R) e F1-score (F).

	NILC								
	COMMA			**PERIOD**			**Geral**		
	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+Skipgrams	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 GRANDE	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

Em uma última análise dos dados quantitativos, pode-se ver na Tabela 16 que todos os modelos obtém um resultado em média abaixo de 0,25 em relação à predição da vírgula, com modelo T5-BASE chegando a 0,205. Os modelos obtêm um resultado semelhante aos apresentados nas Tabelas 15 e 14, com o modelo T5-BASE obtendo o melhor valor de f1-score, igual a 0,957. Assim, o modelo T5 BASE obtém o melhor resultado médio final para f1-score, com 0,607. Ademais, quando treinamos o modelo com o conjunto de dados NILC, o modelo T5-BASE continua obtendo o melhor resultado para utilização da vírgula, com 0,195. E, novamente, todos os modelos também obtêm um resultado aquém daquele obtido anteriormente quando avaliado no conjunto de teste do próprio NILC. Diferentemente de quando treinado com o conjunto de dados IWLST2012 TEDTALK, o algoritmo CRF obtém o melhor resultado

para predição da pontoação final, ficando acima do T5-BASE, com f1-score de 0,967. Assim como aconteceu anteriormente, os modelos apresentaram um resultado fraco na média, com o algoritmo CRF despontando com o melhor resultado, de 0,570. Por fim, também analisamos os modelos GPT-3.5-turbo e GPT-4 na avaliação da restauração de pontuação e podemos ver que nenhum deles obteve um resultado superior aos já obtidos anteriormente, com exceção da revocação na predição da vírgula pelo modelo GPT-4.

Tabela 16 – Resultado dos experimentos para todos os modelos treinados com ambos os conjuntos de dados e testados no conjunto de dados do MEC, considerando as métricas de Precisão (P), Revocação (R) e F1-score (F). (Os resultados foram convertidos para percentagem para fins comparativos).

MEC - TREINADO COM IWLST2012 TEDTALK									
	VÍRGULA			**PONTO**			**GERAL**		
	P	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**
CRF	0,086	0,179	0,116	0,912	0,974	0,942	0,470	0,679	0,556
BLSTM+Skipgrams	0,110	0,384	0,171	0,878	0,969	0,921	0,378	0,752	0,503
BERT BASE	0,104	0,494	0,172	0,876	0,955	0,914	0,289	0,761	0,418
BERT GRANDE	0,104	0,516	0,173	0,878	0,955	0,915	0,284	0,770	0,415
T5 BASE	0,154	0,307	0,205	0,947	0,967	0,957	0,523	0,722	0,607
MEC - TREINADO COM NILC									
	VÍRGULA			**PONTO**			**GERAL**		
	P	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**
CRF	0,083	0,164	0,111	0,960	0,974	0,967	0,494	0,674	0,570
BLSTM+Skipgrams	0,108	0,329	0,162	0,847	0,975	0,907	0,396	0,736	0,515
BERT BASE	0,124	0,382	0,187	0,806	0,961	0,877	0,360	0,717	0,479
BERT GRANDE	0,117	0,365	0,178	0,759	0,959	0,847	0,347	0,709	0,466
T5 BASE	0,131	0,385	0,195	0,940	0,968	0,954	0,429	0,749	0,546
	Zero Shot								
	VÍRGULA			**PONTO**			**GERAL**		
	P	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**
GPT-3.5-turbo	0,068	0,316	0,112	0,235	0,626	0,341	0,152	0,514	0,234
GPT-4	0,072	0,406	0,123	0,479	0,902	0,625	0,240	0,742	0,363

4.2.2 PP2) Os modelos XAI são capazes de fornecer um retorno (*feedback*, termo em inglês) adequado sobre a predição de pontuação?

Para analisar de maneira mais aprofundada as saídas dos modelos, utilizamos a ferramenta de IA explicada chamada *Inseq*, capaz de fornecer os atributos que mais contribuem para cada saída do nosso modelo (SARTI et al., 2023). A Figura 2 mostra como a predição da vírgula para elementos de mesma classe gramatical podem ser separados adequadamente e como eles influenciam diretamente na predição da pontuação na sentença: Comprei arroz, leite, carne e chuchu. Na coluna mais a esquerda (eixo 'y') estão as palavras utilizadas como entrada no modelo, enquanto na linha mais acima (eixo x) estão as palavras preditas pelo modelo. Dessa forma, o mapa de saliências mostra a influência de cada atributo de entrada na coluna esquerda, na predição de cada palavra do modelo na linha acima. Nesse sentido, quanto maior os valores e mais escurar a cor maior é a influência daquele atributo (neste contexto palavra) na saída daquela palavra por parte do modelo. Portanto, a vírgula entre as palavras arroz, leite e carne sofrem uma grande influencia das palavras ao redor (e.g arroz e leite) e menos da palavra comprei que não compartilha a mesma classe gramatical.

Na Figura 3, mostramos como o deslocamento do objeto indireto da sentença para o início é seguido corretamente de uma vírgula e como dos termos "a rosa" e "entreguei" influenciam fortemente na predição da vírgula.

Essa informação pode servir como aparato didático em uma plataforma de correção automática para explicar o porquê da necessidade da vírgula naquele contexto.

Figura 2 – Exemplo de como a regra de pontuação 17 é entendida pelo modelo de restauração de pontuação usando o modelo do T5 base treinado com o conjunto de dados TEDTALK2012.

	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

Figura 3 – Exemplo de como a regra de pontuação é entendida pelo modelo de restauração de pontuação usando o modelo T5 BASE treinado com o conjunto de dados TEDTALK201217.

	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

4.2.3 Análise de Erros dos Modelos de Acordo com a Regra gramatical

Em uma última análise, utilizamos os mais diferentes exemplos de sentenças que requerem pontuação de maneira obrigatória, analisando-as de acordo com 8 regras de pontuação encontradas nos livros (LENZA; MARTINO, 2021) e (SQUARISI, 2021) e agrupando os exemplos com as predições dos modelos no Quadro 17. As regras podem ser categorizadas em duas: a primeira é a utilização restritiva, que limita o sentido do precedente. Na segunda categoria ela é utilizada na explicação do termo anteriormente citado, como no caso da regra R1 (SQUARISI, 2021). No caso da regra R2, a vírgula isola o vocativo do restante da sentença. Além disso, na regra R3 analisamos a predição da vírgula no caso de orações coordenadas, onde os termos de mesma função são separados por vírgula (SQUARISI, 2021). As regras R4 e R5 permitem a ocultação do verbo (R4) ou de um termo de acordo com o contexto (R5) (SQUARISI, 2021), já a regra R6 separa os locais das datas. Ademais, a regra R7 estabelece que termos deslocados precisam ser isolados com vírgula. Neste caso, o termo é 'A rosa', e a regra R8 isola termos como 'por exemplo', 'minto', entre outros.

Como pode ser visto na Tabela 17, o modelo é apenas capaz de predizer corretamente as regras R6 e R8, errando as demais. Um das possíveis razões é a quantidade de exemplos para cada uma delas, dispostos de maneira desproporcional, bem como a utilização de diferentes contextos. A partir dessa tabela, portanto, o modelo pôde ser analisado qualitativamente e então demonstrar que os modelos de IA necessitam ainda de um processo de treinamento mais especifico com relação às regras gramaticais, requerendo uma melhor investigação sobre a aplicação em trabalhos futuros.

Tabela 17 – Exemplo de predição de pontuação usando o modelo T5 base treinado com o conjunto de dados TEDTALK2012 para os exemplos relacionados a regras de pontuação.

Regra		**Exemplos**
R1	T5 base	carlos meu vizinho bateu com o carro.
	referência	Carlos, meu vizinho, bateu com o carro.
R2	T5 base	pai, eu estou com fome.
	referência	Pai, eu estou com fome.
R3	T5 base	comprei arroz, leite, carne e chuchu.
	referência	Comprei arroz, leite, carne e chuchu.
R4	T5 base	uma flor essa menina.
	referência	Uma flor, essa menina.
R5	T5 base	o decreto regulamenta os casos gerais, a portaria os particulares.
	referência	O decreto regulamenta os casos gerais; a portaria, os particulares.
R6	T5 base	formoso 21 de novembro de 2004.
	referência	Formoso, 21 de novembro de 2004.
R7	T5 base	a rosa, entreguei a para a menina.
	referência	A rosa, entreguei-a para a menina.
R8	T5 base	erra se, por exemplo, ao colocar se vírgula entre sujeito e verbo.
	referência	Erra-se, por exemplo, ao colocar-se vírgula entre sujeito e verbo.

4.3 Resumo

Neste capítulo, avaliamos os modelos de pontuação BERT-BASE, BERT-GRANDE, T5-BASE, GPT3.5-turbo, GPT-4 e o algoritmo CRF na predição de pontuações para sentenças em português escritas por alunos. Quando avaliados em textos bem formados e dentro do contexto original no qual o modelo foi treinado, todos os modelos atingem um resultado satisfatório em termos de f1-score, precisão e recall, com destaque para o modelo BERT-GRANDE e T5-BASE. Porém, a avaliação com textos dos alunos provou ser um desafio, visto que as sentenças dos alunos podem conter erros gramaticais graves, dado que um resultado melhor é obtido quando o modelo BERT-BASE é avaliado em textos de obras literárias antigas do português do que quando treinado com ainda menos dados do conjunto de dados TEDTALK2012 (LIMA et al., 2022). Ainda assim, os modelos obtêm um resultado satisfatório na predição do ponto final, contudo falham significativamente quanto à predição da vírgula, possivelmente mais afetada

pelas inconsistências no texto. Além disso, os modelos GPT-3.5 turbo e GPT-4 mostram que apesar dos avanços em abordagens *zero-shot*, modelos treinados em uma atividade especifica ainda podem obter melhores resultados. Portanto, aplicação de restauração de pontuação no contexto educacional ainda possui bastantes desafios, mas que podem ser superados através de uma análise mais aprofundada dos textos dos alunos, tendo em vista os possíveis erros gramaticais que venham a impactar diretamente no resultado final. Ademais, abordagens como aprendizado com poucos exemplos podem melhorar significativamente os resultados aliadas a técnicas de engenharia de *prompt* (termo em inglês). Além disso, a análise de XAI mostrou que é possível atingir resultados que possam mostrar uma correlação entre a predição do modelo e as palavras de entrada que ativam um dos rótulos, neste caso a vírgula,quando o modelo segue as regras de pontuação pré-estabelecidas. Por fim, os trabalhos aqui apresentados foram contextualizados no âmbito da correção automática de redações escritas por alunos. Dessa forma, como forma de consolidar o conhecimento na área de avaliação de redações foi realizado um mapeamento sistemático da literatura avaliando as principais lacunas, bem como, investigando as principais abordagens relacionadas ao tema em português brasileiro.

5 Avaliação Automática de Redação: Uma revisão sistemática

O Exame Nacional do Ensino Médio (ENEM) permite o acesso a diversas universidades públicas do Brasil, porém tem custo excessivo tanto de tempo, bem como, recurso financeiro e humano envolvidos na correção de milhões de redações (MARINHO et al., 2021; MELLO et al., 2021). Neste contexto, a presente dissertação avaliou técnicas capazes de avaliar um aspecto em particular relacionado a avaliação automática de redações que é a correção de pontuação (virgula e ponto final). Apesar disso, escolheu-se desenvolver um mapeamento sistemático mais amplo capaz de mostrar as principais lacunas da área, bem como, os métodos mais recentes para mitiga-los. Assim, neste primeiro estudo, realizou-se um Mapeamento da Literatura sobre a correção automática de redação por meio do uso de técnicas de PLN. O estudo evidência os principais temas investigados na área em língua portuguesa, mostrando as lacunas e limitações dos trabalhos existentes, bem como, os métodos aplicados na condução dos estudos realizados. Elaborou-se uma pergunta de pesquisa mais ampla: **"Avaliação Automática de Redação: Uma revisão sistemática?"**, seguida de 6 outras subquestões de pesquisa, a fim de avaliar o status atual das pesquisas relacionadas à correção de redação do ENEM no Brasil.

Questão de Pesquisa 1 (Q1): *Quais são os principais objetivos da utilização de inteligência artificial na avaliação de redações?*

A primeira pergunta de pesquisa Q1 busca elucidar por que a inteligência artificial se tornou relevante na área e o objetivo de utilização desta.

Questão de Pesquisa 2 (Q2): *Quais os principais algoritmos de inteligência artificial que são utilizados para a avaliação de redações?*

A segunda pergunta (Q2), explora os principais algoritmos de IA utilizados na área, investigando quais deles são mais promissores neste contexto.

Questão de Pesquisa 3 (Q3): *Quais são as métricas mais utilizadas para validação?*

A avaliação dos algoritmos é parte essencial do processo de validação das propostas. Buscou-se através dessa pergunta de pesquisa (Q3) elucidar quais as principais métricas utilizadas em determinado contexto, o que pode ajudar a classificar e comparar os diferentes métodos utilizados em cada artigo.

Questão de Pesquisa 4 (Q4): *Quais são os bancos de dados mais utilizados para validação?*

Através dessa pergunta de pesquisa (Q4) podem-se organizar os principais conjuntos de dados utilizados na área para validação dos resultados apresentados pelos modelos, o que pode ser útil em uma comparação entre as diferentes abordagens no futuro.

Questão de Pesquisa 5 (Q5): *Existe alguma evidência de que a inteligência artificial auxilia na avaliação de redações?*

A pergunta de pesquisa Q5 endereça a necessidade de um olhar voltado para a prática: avaliando como modelos automáticos podem ser de fato empregados para ajudar alunos e professores no processo de ensino-aprendizagem.

Questão de Pesquisa 6 (Q6): *Quais os critérios utilizados na avaliação das redações?*

Os critérios de avaliados são importantes para saber quais os principais desafios no processo de avaliação e em quais aspectos se encontrar maior dificuldade, os quais podem ser endereçados em trabalhos futuros.

5.1 Design Experimental

Para responder essas perguntas de pesquisa foi realizada uma pesquisa sistemática, considerando um grupo de palavras-chave, tanto em português como em inglês, usando os conectivos lógicos AND e OR, relacionadas à área de educação de inteligência artificial.

- Educacional - redação (*essay*), tema (*prompt*), gramatica (*grammar*);
- Inteligência Artificial - aprendizado de máquina (*machine learning*), *deep learning*, processamento de linguagem natural (*natural language processing*);
- Aplicabilidade - avaliação (*evaluation/assessment*), pontuação (*scoring*), correção (*correction*), classificação (*grading*);
- Idioma de Aplicação - português (*portuguese*).

(“redação” OR “tema” OR “grámatica”)

AND

(“aprendizado de máquina” OR “deep learning” OR “processamento de linguagem natural”)

AND

(“avaliação” OR “pontuação” OR ”correção”OR ”classificação”)

AND

(“português”)

As pesquisas foram realizadas nas bases de dados de artigo científico ACM[1], IEEExplore[2], Engineering Village[3], Science Direct[4], SpringerLink[5], Scopus[6], Web of Science[7], SBC OpenLib[8]. No total, foram retornados 1118 artigos 1149 em inglês e apenas 39 em português.

Foram estabelecidos 9 critérios de seleção e exclusão para que os artigos seguissem como parte do estudo, dentre os critérios de inclusão estão: os artigos devem (1) ser estudos primários; (2) propor abordagens de inteligência artificial na avaliação de redações; e (3) analisar redações em português. Os critérios de exclusão seguem como sendo: (4) estudo secundários ou terciários; (5) artigos duplicados ou re-indexados; (6) artigos em diferentes idiomas; (7) estudos em literatura cinza; (8) incompletos ou ainda (9) cuja via de publicação não seja conferência e jornal. Além disso, busca dos artigos se deu no período de Janeiro/2012 até Março/2022.

A nossa pesquisa foi separada em 3 etapas principais. A primeira etapa consistiu na remoção de artigos duplicados, a segunda avaliação do título e do *abstract* (resumo) em relação aos critérios de seleção e, por fim, a terceira etapa consistiu na eliminação dos artigos de acordo com os critérios de seleção, após a leitura da introdução e das considerações finais.

Na primeira etapa, todas as referências encontradas foram organizadas na plataforma Rayyan[9], que se provou efetiva para organização e condução em trabalhos anteriores, especialmente na detecção automática de artigos duplicados (PAPADOPOULOS et al., 2020; NUNES et al., 2022; GUIMARÃES et al., 2022; MCKEOWN; MIR, 2021). Assim, foram eliminados 29 artigos duplicados de um total de 1118. Na segunda etapa, 1081 artigos foram eliminados considerando os critérios de seleção quando analisados o *abstract* ou resumo, restando 78 artigos. Por fim, mais 72 artigos foram descartados quando considerados os critérios de seleção em relação à introdução e conclusão, ficando 6 ao final.

5.2 Resultados

Em relação à primeira pergunta de pesquisa **"Quais são os principais objetivos da utilização de inteligência artificial na avaliação de redações?"**, encontrou-se que os trabalhos

[1] ⟨https://dl.acm.org/⟩
[2] ⟨https://ieeexplore.ieee.org/⟩
[3] ⟨https://www.engineeringvillage.com/⟩
[4] ⟨https://www.sciencedirect.com⟩
[5] ⟨https://link.springer.com/⟩
[6] ⟨https://www.scopus.com/⟩
[7] ⟨https://www.webofscience.com⟩
[8] ⟨https://sol.sbc.org.br/⟩
[9] https://rayyan.ai/

selecionados implementam algoritmos de IA porque eles são mais eficientes do que avaliadores humanos em determinadas atividades; não possuem vieses na avaliação; proveem um retorno importante para educadores através dos dados analíticos gerados. Os trabalhos se dividem entre dois objetivos principais: o primeiro é uma avaliação de todos os aspectos relacionados à redação, de acordo com avaliação do ENEM, por meio da utilização de classificadores automáticos (FILHO et al., 2019). O segundo consiste em avaliar um aspecto específico, como coesão e análise da capacidade dissertativa e argumentativa do aluno (MELLO et al., 2022; SOUSA et al., 2021; FILHO et al., 2018a; MELLO et al., 2021). Para analisar como estes objetivos são atingidos utilizamos a segunda pergunta de pesquisa **"Quais os principais algoritmos de inteligência artificial que são utilizados para a avaliação de redações"**. Assim, investigamos os principais algoritmos utilizados na correção de pontuação. Como resultado da pesquisa tem-se que os algoritmos de aprendizado de máquina baseados na extração de características como Máquina de Vetor de Suporte, Árvore de Decisão, XGBoost são os mais utilizados. Alguns trabalhos ainda usaram modelos baseados em redes neurais como *Bidirectional Long Short Memory* (BLSTM) e *multilingual Bidirectional Encoder Representation* (mBert) (SOUSA et al., 2021; FILHO et al., 2019; MELLO et al., 2022; FILHO et al., 2018a).

Ademais, na terceira pergunta **"Quais as métricas utilizadas?"** as principais métricas de avaliação utilizadas vão desde mais abrangentes, como Erro médio quadrático (RMSE), F1-score, Precisão e Revocação, como métricas utilizadas na área de sistemas educacionais como Quadratic Weight Kappa (QWK), *Kappa* de *Coehn* e Correlação de Person (SOUSA et al., 2021; FILHO et al., 2019; MELLO et al., 2022; FILHO et al., 2018a; FONSECA et al., 2018). Na quarta pergunta, sobre os bancos de dados utilizados, foram desde de banco de dados de redação do ENEM até textos traduzidos da língua inglesa utilizado tanto para treinamento e teste dos algoritmos e modelos (FILHO et al., 2019; SOUSA et al., 2021), como mostrado na tabela 18.

Tabela 18 – A tabela mostra o número de redações em cada uma das base de dados citadas.

ID	Fonte do Banco de Dados	Número de Redações
1	(SANTOS et al., 2018)	271
2	(FILHO et al., 2018b)	50
3	(SOUSA et al., 2021)	402
4	Redações da UOL e do Brasil Escola	1983
5	(FONSECA et al., 2018)	56.644

Quanto à questão de pesquisa **"Existe alguma evidência de que a aprendizagem de máquina auxilia na avaliação de redações?"**, nenhum dos trabalhos avaliados no mapeamento realizado incluía a avaliação dos modelos e algoritmos propostos em ambiente real, e portanto, não foi possível inferir a partir de trabalhos que propõem a correção de redação em português brasileiro que a correção automática de textos de redação auxilia efetivamente no processo de correção como um todo. Dessa forma, o trabalho foi estendido e avaliou-se trabalhos que realizam a investigação de um sistema de correção automática para língua inglesa. No trabalho de (NUNES et al., 2022), 7 dos 8 trabalhos apresentaram evidência de uma avaliação positiva por parte de professores e alunos na utilização de sistemas automáticos de correção. Para os professores, o principal ganho é a redução no trabalho de correção e foco na assistência dos alunos em si. As aplicações também aumentam o nível de confiança dos alunos durante a escrita, embora não melhore as notas em si e não seja efetivo para alunos com muita dificuldade (PALERMO; THOMSON, 2018; TANG; RICH, 2017; WARE, 2014; NUNES et al., 2022). Por fim, na última pergunta de pesquisa, **"Quais os critérios utilizados na avaliação das redações?"**, encontramos que 3 dos 6 trabalhos analisados focam na competência 3 e 4 (MELLO et al., 2022; SOUSA et al., 2021; MELLO et al., 2021). Enquanto isso, apenas um considera a competência 1 e outro a competência 2 (FILHO et al., 2019; FILHO et al., 2018a), e por fim o trabalho (FONSECA et al., 2018) considera todas as competências.

5.3 Resumo

O mapeamento buscou elucidar pontos relevantes da pesquisa relacionados à correção automática de redação em português brasileiro. Primeiramente, a maior parte dos trabalhos focam apenas em um dos aspectos relacionados à correção de pontuação, ao invés de uma análise completa, se concentrando majoritariamente em coesão e coerência (FILHO et al., 2019; FILHO et al., 2018a). Ademais, nenhum trabalho focado na língua portuguesa realizou a validação em ambiente real, o que pode limitar a capacidade de analisar o real impacto das propostas apresentadas.

6 Considerações Finais

A restauração de pontuação tem uma ampla investigação em língua inglesa, contudo, trabalhos em língua portuguesa são escassos o que foi investigado e proposto no nosso segundo capítulo da dissertação (LU; NG, 2010; TILK; ALUMÄE, 2016; COURTLAND et al., 2020; NAGY et al., 2021). Dessa forma, investigou-se diferentes abordagens para restauração de pontuação, aplicando modelos e algoritmos para predição de pontuação de maneira correta. Os melhores modelos foram avaliados tanto em textos fora do domínio em que foram treinados, com resultados inicialmente promissores para uma abordagem de avaliação fora do domínio específico, atingindo 0,735 de f1-score. Portanto os objetivos inicial do trabalho foi alcançado no que tange ao desenvolvimento de um algoritmo de restauração de pontuação em Português, contudo, é necessário o aprofundamento na pesquisa considerando a avaliação de textos de alunos de maneira automática o que foi explorado nos capítulos seguintes.

Dessa forma, até onde foi a pesquisa dessa dissertação, este é primeiro trabalho a investigar a correção automática de pontuação em textos de alunos em Português brasileiro (LIMA et al., 2023a; LIMA et al., 2022). Contudo, quando avaliou-se o modelo com textos escritos por alunos do ensino fundamental, o resultado, principalmente em relação à vírgula, tem uma maior perda, o que pode ter sido ocasionado por inconsistências na escrita do texto. Ademais, os modelos GPT-3 turbo e GPT-4 apresentaram um resultado ainda mais aquém dos demais modelos em uma abordagem *zero-shot*. Portanto, avançou-se com a pesquisa propondo uma avaliação utilizando IA explicada como forma de mitigar o problema.

Assim, este é o primeiro trabalho na área a utilizar XAI para dar transparência a modelos de restauração de pontuação. A transparência é crucial por diversos motivos principalmente na área educacional. Contudo, nem todos os modelos de inteligência artificial são transparentes e explicáveis. Dessa forma, utilizamos técnicas e ferramentas que ajudaram a melhor compreender o processo de predição de pontuação em modelos generativos como T5 o que pode permitir o entendimento desses modelos quando avaliados em textos com falhas gramaticais graves. Por fim, apesar de ser promissora, a abordagem apresenta algumas limitações que podem ser investigadas em trabalhos futuros na área.

6.1 Artigos aceitos

Os artigos resultantes desta pesquisa que foram aceitos em conferências e revistas da área foram os seguintes:

LIMA, T. B. D. et al. 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. Springer, 2022. p. 616–630.

BARBOSA DE LIMA, T. et al. Avaliação Automática de Redação: Uma revisão sistemática. Revista Brasileira de Informática na Educação, v. 31, p. 205–221, maio 2023. Disponível em: https://sol.sbc.org.br/journals/index.php/rbie/article/view/2869. DOI: 10.5753/rbie.2023.2869.

De Lima, T.; Rodrigues, L.; Macario, V.; Xavier, E.; Mello, R. F. (2023). "Automatic Punctuation Verification of School Students' Essay in Portuguese."In: ENIAC 2023, 25-29 de setembro de 2023. Disponível em: https://sol.sbc.org.br/index.php/eniac/article/view/25693. DOI: 10.5753/eniac.2023.233559

6.2 Limitações da pesquisa

Uma das maiores limitações encontradas em nossa pesquisa é o fato de os modelos não serem robustos quando avaliados com textos que contenham erros gramaticais, além dos erros de pontuação.

Ademais, quanto à utilização de modelos grandes de linguagem, apenas uma técnica de *prompt* (termo em inglês) foi avaliada e testada, o que pode ser uma limitação, dado que outras técnicas podem apresentar melhores resultados. Por fim, apenas a abordagem de *zero-shot learning*, quando o modelo não recebe nenhum exemplo, foi avaliada, em vez de técnicas como *few-shot*, o que pode ser uma fator limitador.

6.3 Trabalhos Futuros

Em trabalhos futuros, pode-se investigar melhores formas de entender o porquê de uma queda acentuada na avaliação de textos dos alunos do ensino fundamental e quais erros gramaticais se mostram mais relevantes. Além disso, pode-se investigar mais modelos, como

aqueles multilingual e de geração de texto, como mbart, mT5 e avaliar o desempenho nos textos dos alunos. Além disso, outros modelos poderiam ter sido treinados e avaliados, como modelos multilingual tais quais Roberta, mbart e mT5. È necessário, no entanto, considerar também a utilização de outras abordagens de engenharia de *prompt* (termo em inglês), além daquela utilizada neste trabalho. Utilizamos uma abordagem em que o modelo de linguagem simula o papel de uma Persona sem especificar diretamente a saída desejada, porém outras abordagens podem se mostrar mais promissoras (WHITE et al., 2023). Por último, poderíamos analisar outras abordagens como aprendizado com poucos exemplos, a fim de melhorar os resultados do modelo de maneira significativa (BROWN et al., 2020). Ademais, pode-se avaliar outras técnicas de engenharia de *prompt* (termo em inglês), além daquela apresentada nesse trabalho, bem como, a avaliação de aprendizado com poucos exemplos como forma de melhorar os resultados utilizando os modelos GPT-3.5 turbo e GPT-4.

Referências

BIEWALD, L. **Experiment Tracking with Weights and Biases**. 2020. Software available from wandb.com. Disponível em: ⟨https://www.wandb.com/⟩.

BROWN, T.; MANN, B.; RYDER, N.; SUBBIAH, M.; KAPLAN, J. D.; DHARIWAL, P.; NEELAKANTAN, A.; SHYAM, P.; SASTRY, G.; ASKELL, A. et al. Language models are few-shot learners. **Advances in neural information processing systems**, v. 33, p. 1877–1901, 2020.

CHE, X.; WANG, C.; YANG, H.; MEINEL, C. Punctuation prediction for unsegmented transcript based on word vector. In: **Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)**. [S.l.: s.n.], 2016. p. 654–658.

CHRISTENSEN, H.; GOTOH, Y.; RENALS, S. Punctuation annotation using statistical prosody models. 2001.

COURTLAND, M.; FAULKNER, A.; MCELVAIN, G. Efficient automatic punctuation restoration using bidirectional transformers with robust inference. In: **Proceedings of the 17th International Conference on Spoken Language Translation**. [S.l.: s.n.], 2020. p. 272–279.

DANILEVSKY, M.; QIAN, K.; AHARONOV, R.; KATSIS, Y.; KAWAS, B.; SEN, P. A survey of the state of explainable ai for natural language processing. **arXiv preprint arXiv:2010.00711**, 2020.

DEVLIN, J.; CHANG, M.-W.; LEE, K.; TOUTANOVA, K. Bert: Pre-training of deep bidirectional transformers for language understanding. **arXiv preprint arXiv:1810.04805**, 2018.

FEDERICO, M.; CETTOLO, M.; BENTIVOGLI, L.; MICHAEL, P.; SEBASTIAN, S. Overview of the iwslt 2012 evaluation campaign. In: **Proceedings of the international workshop on spoken language translation (IWSLT)**. [S.l.: s.n.], 2012. p. 12–33.

FILHO, A. H.; CONCATTO, F.; NAU, J.; PRADO, H. A. d.; IMHOF, D. O.; FERNEDA, E. Imbalanced Learning Techniques for Improving the Performance of Statistical Models in Automated Essay Scoring. **Procedia Computer Science**, v. 159, p. 764–773, 2019. ISSN 18770509.

FILHO, A. H.; PRADO, H. A. do; FERNEDA, E.; NAU, J. An approach to evaluate adherence to the theme and the argumentative structure of essays. **Procedia Computer Science**, v. 126, p. 788–797, 2018. ISSN 18770509.

FILHO, A. H.; PRADO, H. A. do; FERNEDA, E.; NAU, J. An approach to evaluate adherence to the theme and the argumentative structure of essays. **Procedia Computer Science**, Elsevier, v. 126, p. 788–797, 2018.

FONSECA, E.; MEDEIROS, I.; KAMIKAWACHI, D.; BOKAN, A. Automatically Grading Brazilian Student Essays. In: VILLAVICENCIO, A.; MOREIRA, V.; ABAD, A.; CASELI, H.; GAMALLO, P.; RAMISCH, C.; OLIVEIRA, H. G.; PAETZOLD, G. H. (Ed.). **Computational Processing of the Portuguese Language**. Cham: Springer International Publishing, 2018. v. 11122, p. 170–179. ISBN 978-3-319-99721-6 978-3-319-99722-3. Series Title: Lecture Notes in Computer Science.

GAZZOLA SIDNEY EVALDO LEAL, S. M. A. M. Predição da complexidade textual de recursos educacionais abertos em português. In: **Proceedings of the Brazilian Symposium in Information and Human Language Technology**. [S.l.: s.n.], 2019.

GOODING, S.; KOCHMAR, E. Complex word identification as a sequence labelling task. In: **Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics**. [S.l.: s.n.], 2019. p. 1148–1153.

GUIMARÃES, N. S.; FERREIRA, A. J.; SILVA, R. d. C. R.; PAULA, A. A. de; LISBOA, C. S.; MAGNO, L.; ICHIARA, M. Y.; BARRETO, M. L. Deduplicating records in systematic reviews: there are free, accurate automated ways to do so. **Journal of Clinical Epidemiology**, Elsevier, v. 152, p. 110–115, 2022.

KHOSRAVI, H.; SHUM, S. B.; CHEN, G.; CONATI, C.; TSAI, Y.-S.; KAY, J.; KNIGHT, S.; MARTINEZ-MALDONADO, R.; SADIQ, S.; GAŠEVIĆ, D. Explainable artificial intelligence in education. **Computers and Education: Artificial Intelligence**, Elsevier, v. 3, p. 100074, 2022.

KUMAR, V.; BOULANGER, D. Explainable automated essay scoring: Deep learning really has pedagogical value. In: FRONTIERS MEDIA SA. **Frontiers in education**. [S.l.], 2020. v. 5, p. 572367.

LENZA, P.; MARTINO, A. **Português Esquematizado**. Saraiva Educação S.A., 2021. ISBN 9786555597301. Disponível em: ⟨https://books.google.com.br/books?id=QHRIEAAAQBAJ⟩.

LIMA, T. B. D.; MIRANDA, P.; MELLO, R. F.; WENCESLAU, M.; BITTENCOURT, I. I.; CORDEIRO, T. D.; JOSÉ, J. Sequence labeling algorithms for punctuation restoration in brazilian portuguese texts. In: SPRINGER. **Intelligent Systems: 11th Brazilian Conference, BRACIS 2022, Campinas, Brazil, November 28–December 1, 2022, Proceedings, Part II**. [S.l.], 2022. p. 616–630.

LIMA, T. Barbosa de; SILVA, I. Luana Almeida da; FREITAS, E. L. S. X.; MELLO, R. F. Avaliação automática de redação: Uma revisão sistemática. **Revista Brasileira de Informática na Educação**, v. 31, p. 205–221, maio 2023. Disponível em: ⟨https://sol.sbc.org.br/journals/index.php/rbie/article/view/2869⟩.

LIMA, T. D.; RODRIGUES, L.; MACARIO, V.; XAVIER, E.; MELLO, R. F. Automatic punctuation verification of school students' essay in portuguese. In: **ENIAC 2023 ()**. [S.l.: s.n.], 2023.

LU, W.; NG, H. T. Better punctuation prediction with dynamic conditional random fields. In: **Proceedings of the 2010 conference on empirical methods in natural language processing**. [S.l.: s.n.], 2010. p. 177–186.

MAKHIJA, K.; HO, T.-N.; CHNG, E.-S. Transfer learning for punctuation prediction. In: IEEE. **2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)**. [S.l.], 2019. p. 268–273.

MARINHO, J.; ANCHIêTA, R.; MOURA, R. Essay-br: a brazilian corpus of essays. In: **Anais do III Dataset Showcase Workshop**. Online: Sociedade Brasileira de Computação, 2021. p. 53–64.

MATUSOV, E.; MAUSER, A.; NEY, H. Automatic sentence segmentation and punctuation prediction for spoken language translation. In: **Proceedings of the Third International Workshop on Spoken Language Translation: Papers**. [S.l.: s.n.], 2006.

MCKEOWN, S.; MIR, Z. M. Considerations for conducting systematic reviews: evaluating the performance of different methods for de-duplicating references. **Systematic reviews**, Springer, v. 10, p. 1–8, 2021.

MELLO, R. F.; FIORENTINO, G.; MIRANDA, P.; OLIVEIRA, H.; RAKOVIć, M.; GAšEVIć, D. Towards Automatic Content Analysis of Rhetorical Structure in Brazilian College Entrance Essays. In: ROLL, I.; MCNAMARA, D.; SOSNOVSKY, S.; LUCKIN, R.; DIMITROVA, V. (Ed.). **Artificial Intelligence in Education**. Cham: Springer International Publishing, 2021. v. 12749, p. 162–167. ISBN 978-3-030-78269-6 978-3-030-78270-2. Series Title: Lecture Notes in Computer Science.

MELLO, R. F.; FIORENTINO, G.; OLIVEIRA, H.; MIRANDA, P.; RAKOVIC, M.; GASEVIC, D. Towards automated content analysis of rhetorical structure of written essays using sequential content-independent features in Portuguese. In: **LAK22: 12th International Learning Analytics and Knowledge Conference**. Online USA: ACM, 2022. p. 404–414. ISBN 978-1-4503-9573-1.

NAGY, A.; BIAL, B.; ÁCS, J. Automatic punctuation restoration with bert models. **arXiv preprint arXiv:2101.07343**, 2021.

NUNES, A.; CORDEIRO, C.; LIMPO, T.; CASTRO, S. L. Effectiveness of automated writing evaluation systems in school settings: A systematic review of studies from 2000 to 2020. **Journal of Computer Assisted Learning**, Wiley Online Library, v. 38, n. 2, p. 599–620, 2022.

OLIVEIRA, H.; MELLO, R. F.; ROSA, B. A. B.; RAKOVIC, M.; MIRANDA, P.; CORDEIRO, T.; ISOTANI, S.; BITTENCOURT, I.; GASEVIC, D. Towards explainable prediction of essay cohesion in portuguese and english. In: **LAK23: 13th International Learning Analytics and Knowledge Conference**. [S.l.: s.n.], 2023. p. 509–519.

PĂIŞ, V.; TUFIŞ, D. Capitalization and punctuation restoration: a survey. **Artificial Intelligence Review**, Springer, p. 1–42, 2021.

PALERMO, C.; THOMSON, M. M. Teacher implementation of self-regulated strategy development with an automated writing evaluation system: Effects on the argumentative writing performance of middle school students. **Contemporary Educational Psychology**, Elsevier, v. 54, p. 255–270, 2018.

PAN, R.; GARCÍA-DÍAZ, J. A.; VALENCIA-GARCÍA, R. Evaluation of transformer-based models for punctuation and capitalization restoration in spanish and portuguese. In: SPRINGER. **International Conference on Applications of Natural Language to Information Systems**. [S.l.], 2023. p. 243–256.

PAPADOPOULOS, I.; KOULOUGLIOTI, C.; LAZZARINO, R.; ALI, S. Enablers and barriers to the implementation of socially assistive humanoid robots in health and social care: a systematic review. **BMJ open**, British Medical Journal Publishing Group, v. 10, n. 1, p. e033096, 2020.

PAPINENI, K.; ROUKOS, S.; WARD, T.; ZHU, W.-J. Bleu: a method for automatic evaluation of machine translation. In: **Proceedings of the 40th Annual Meeting of the Association for**

Computational Linguistics. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics, 2002. p. 311–318. Disponível em: ⟨https://aclanthology.org/P02-1040⟩.

SANTOS, K. S. dos; SODER, M.; MARQUES, B. S. B.; FELTRIM, V. D. Analyzing the rhetorical structure of opinion articles in the context of a brazilian college entrance examination. In: SPRINGER. **Computational Processing of the Portuguese Language: 13th International Conference, PROPOR 2018, Canela, Brazil, September 24–26, 2018, Proceedings 13**. [S.l.], 2018. p. 3–12.

SARTI, G.; FELDHUS, N.; SICKERT, L.; WAL, O. van der. Inseq: An interpretability toolkit for sequence generation models. **arXiv preprint arXiv:2302.13942**, 2023.

SOUSA, A.; LEITE, B.; ROCHA, G.; CARDOSO, H. L. Cross-Lingual Annotation Projection for Argument Mining in Portuguese. In: MARREIROS, G.; MELO, F. S.; LAU, N.; CARDOSO, H. L.; REIS, L. P. (Ed.). **Progress in Artificial Intelligence**. Cham: Springer International Publishing, 2021. v. 12981, p. 752–765. ISBN 978-3-030-86229-9 978-3-030-86230-5. Series Title: Lecture Notes in Computer Science.

SQUARISI, D. **50 Dicas para o uso da Pontuação. Disponível em: Minha Biblioteca**. [S.l.: s.n.], 2021.

SUNDARARAJAN, M.; TALY, A.; YAN, Q. Axiomatic attribution for deep networks. In: PMLR. **International conference on machine learning**. [S.l.], 2017. p. 3319–3328.

TANG, J.; RICH, C. S. Automated writing evaluation in an efl setting: Lessons from china. **JALT CALL Journal**, ERIC, v. 13, n. 2, p. 117–146, 2017.

TILK, O.; ALUMÄE, T. Bidirectional recurrent neural network with attention mechanism for punctuation restoration. In: **Interspeech**. [S.l.: s.n.], 2016. p. 3047–3051.

TJOA, E.; GUAN, C. A survey on explainable artificial intelligence (xai): Toward medical xai. **IEEE transactions on neural networks and learning systems**, IEEE, v. 32, n. 11, p. 4793–4813, 2020.

WARE, P. Feedback for adolescent writers in the english classroom. **Writing & Pedagogy**, v. 6, n. 2, 2014.

WHITE, J.; FU, Q.; HAYS, S.; SANDBORN, M.; OLEA, C.; GILBERT, H.; ELNASHAR, A.; SPENCER-SMITH, J.; SCHMIDT, D. C. A prompt pattern catalog to enhance prompt engineering with chatgpt. **arXiv preprint arXiv:2302.11382**, 2023.

7 Apêndices

Revista Brasileira de Informática na Educação – RBIE
Brazilian Journal of Computers in Education
(ISSN online: 2317-6121; print: 1414-5685)
https://sol.sbc.org.br/journals/index.php/rbie

Submission: 02/Sep/2022; 1st round notif.: 04/Nov/2022; New version: 12/Dec/2022; 2nd round notif.: 07/Mar/2023;
Camera ready: 02/Apr/2023; Edition review: 20/Apr/2023; Available online: 14/May/2023; Published: 14/May/2023;

Avaliação Automática de Redação: Uma revisão sistemática

Title: Automatic Essay Evaluation in Portuguese: A Systematic Review

Tiago Barbosa de Lima
Universidade Federal Rural de Pernambuco
ORCID: 0000-0002-0707-522X
tiago.blima@ufrpe.br

Ingrid Luana Almeida da Silva
Universidade Federal Rural de Pernambuco
ORCID: 0009-0000-9197-7535
ingrid.luana@ufrpe.br

Elyda Laisa Soares Xavier Freitas
Universidade de Pernambuco
ORCID: 0000-0001-7439-9040
elyda.freitas@upe.br

Rafael Ferreira Mello
Universidade Federal Rural de Pernambuco
ORCID: 0000-0003-3548-9670
rafael.mello@ufrpe.br

Resumo

A Avaliação Automática de Redação (do inglês, Automatic Essay Scoring - AES) tem sido tema amplamente explorado na literatura. Ela permite dispensar o esforço humano aplicado na correção de um grande número de redações em um curto espaço de tempo. A maior parte dos trabalhos disponíveis na literatura se concentra no esforço de desenvolver algoritmos que sejam capazes de corrigir automaticamente textos em inglês. No entanto, para a língua portuguesa, essa ainda é uma área que está em desenvolvimento. Neste contexto, este artigo apresenta um Mapeamento Sistemático da Literatura que busca identificar as abordagens de Inteligência Artificial que estão sendo utilizadas para oferecer suporte à avaliação de redações escritas na língua portuguesa. Os principais achados deste artigo incluem os seguintes fatos: (i) as abordagens dos trabalhos selecionados costumam focar no uso de atributos extraídos do texto em vez do uso de modelos pré-treinados baseados em Deep Learning; (ii) existe prevalência de métricas tradicionais, como Precisão, Cobertura e F-Measure na validação dos resultados; (iii) os feedbacks gerados pelas abordagens possuem um baixo detalhamento; e (iv) os artigos selecionados não analisam o impacto prático em aplicações do mundo real.

Palavras-chave: *Correção de Redação; Análise de Conteúdo; Processamento de Linguagem Natural*

Abstract

The literature has vastly explored Automatic Essay Scoring (AES) in the last few years. The critical motivation is the possibility of reducing the human effort in scoring a large number of essays in a short period. In literature, most of the work concentrates on the English language; there is still a need for progress in Brazilian Portuguese. Thus, this works provides a Systematic Mapping Study aiming to identify Artificial Intelligence methods that support Automatic Essay Correction in Brazilian Portuguese. Furthermore, the main facts this paper brings are: (i) the methods focus on feature engineering methods instead of deep learning models; (ii) there is a prevalence of traditional metrics such as precision, coverage, and f-measure to evaluate the results; (iii) feedbacks provided by the tools have low-level of details; and (iv) there is no practical evaluation of the advancement in real-world applications.

Keywords: *Essay Scoring; Content Analysis; Natural Language processing*

Cite as: Lima, T. B., Silva, I. L. A., Freitas, E. L. S. X. Mello, R. F. (2023). Avaliação Automática de Redação: Uma revisão sistemática. Revista Brasileira de Informática na Educação – RBIE), 31, 205-221. DOI: 10.5753/rbie.2023.2869.

1 Introdução

No Brasil, milhões de estudantes participam anualmente do Exame Nacional do Ensino Médio (ENEM) como parte da sua trajetória acadêmica. Esse exame mede diferentes competências dos estudantes em áreas diversas, como português, matemática e ciências. Uma dessas competências é a de escrever uma redação no formato discursivo argumentativo, seguindo um tema proposto por um texto motivador. A redação deve seguir alguns critérios[1], de acordo com as seguintes competências: (i) aderência à escrita formal do português; (ii) escrita de acordo com o estilo argumentativo discursivo; (iii) a defesa de um ponto de vista; (iv) estrutura argumentativa; e (v) a elaboração de uma proposta de intervenção no problema debatido ao longo do texto (Marinho et al., 2021). Cada competência é pontuada entre 0 e 200, onde 0 é a pior nota e 200 a melhor.

No ano de 2022, o número de inscritos no ENEM atingiu a marca de 5.3 milhões de estudantes. A vasta quantidade de redações a serem avaliadas gera uma demanda por professores habilitados, ocasionando um custo excessivo (Mello et al., 2021). Por outro lado, a investigação de métodos de correção automática proporcionou um avanço significativo na forma como são tratados os processos de correção de redação. No entanto, essa área apresenta muitos desafios, pois é preciso garantir que a correção das redações esteja sendo realizada de forma precisa. Dessa forma, diversos métodos já foram propostos - desde aqueles que utilizam *deep learning* àqueles que são baseados na extração de atributos capazes de serem usados em classificadores como Máquina de Vetor de Suporte (MVS), Árvore de Decisão e XGBoost (Chen et al., 2015; Safavian e Landgrebe, 1991; Cortes e Vapnik, 1995; Fonseca et al., 2018; Ferreira Mello et al., 2022). Além disso, também costumam ser considerados desafiadores os aspectos relacionados às avaliações dos modelos, à detecção de possíveis vieses e à apresentação de *feedbacks* assertivos e compreensíveis.

A avaliação automática de redações é uma área que apresenta bastantes estudos aplicados à língua inglesa. Na língua portuguesa, essa é uma área que vem se expandindo ao longo dos últimos anos (Costa et al., 2020). Dessa forma, investigar a literatura permite esclarecer as principais questões relacionadas, métodos utilizados, desafios e propostas apresentadas. Neste contexto, este artigo propõe um Mapeamento Sistemático da Literatura a fim de investigar propostas, desafios e limitações da avaliação automática de redações em português. O processo de pesquisa foi dividido em três etapas: (i) Busca; (ii) Seleção e (iii) Extração. Ao final da aplicação das etapas planejadas, foram selecionados 6 artigos que foram analisados levando em consideração 6 questões de pesquisa, contribuindo assim para um entendimento mais aprofundado da área em questão.

2 Trabalhos Relacionados

Para tornar as avaliações das redações cada vez mais precisas, a comunidade científica tem trabalhado os diversos aspectos que envolvem essa questão. E, bem como neste trabalho, outros artigos buscaram revisar a literatura a fim de entender o estado da arte no que se refere à avaliação automática de redações. Desse modo, esta seção apresenta um conjunto de revisões da literatura já realizadas sobre o tema e discute quais as principais diferenças para a proposta desta pesquisa.

[1]A consulta detalhada aos critérios está disponível em: http://portal.mec.gov.br/ultimas-noticias/418-enem-946573306/81381-conheca-as-cinco-competencias-cobradas-na-redacao-do-enem

Em Nau et al. (2019) é realizada uma Revisão Sistemática da Literatura que busca analisar o estado da arte na área de avaliação automática de redações. Esse trabalho realizou buscas de artigos de acordo com um conjunto de palavras-chave em três diferentes repositórios: ACL, Scopus e Science Direct. Os critérios de seleção consideram tanto título como resumo e palavras-chave, organizadas de acordo com a seguinte expressão:

essay AND (scoring OR identifying) AND ("discourse element"OR "discourse analysis"OR "discourse structure"OR "conclusion statements") AND (nlp OR "natural language processing")

No referido trabalho, artigos que propõem algum tipo de intervenção também foram considerados. Foram incluídos artigos escritos na língua inglesa ou portuguesa que foram publicados entre 01/2012 e 12/2017. Após as buscas, foram retornados 87 trabalhos, sendo selecionados 5 trabalhos para a avaliação final. Esses trabalhos foram avaliados levando em consideração as características dos *corpus* de redações utilizados que possuem variados temas, a proposta de avaliação e as técnicas utilizadas nessa proposta.

No trabalho de Costa et al. (2020) foi realizado um Mapeamento Sistemático da Literatura cujo objetivo é apresentar um panorama do estado da arte na avaliação de corretores automáticos para a língua portuguesa. Diferentemente do trabalho de Nau et al. (2019), este trabalho foca apenas em artigos que fazem a aplicação da avaliação automática de redações em textos escritos na língua portuguesa. Os critérios de inclusão se concentram em artigos que são relacionados à correção automática em língua portuguesa - seja explicitamente seja relacionado. Também são incluídos artigos de *surveys*. Os autores buscaram definir as principais estratégias utilizadas na identificação de elementos textuais, os aspectos linguísticos utilizados, as métricas e as bases de dados utilizadas nas soluções. As buscas foram realizadas nas fontes digitais Scopus e IEEE e foram retornados 787 artigos no total. Foram selecionados 6 artigos através de critérios de inclusão e exclusão, e foram inseridos 4 artigos de forma manual, totalizando 10 artigos ao final do processo - que vão desde o ano 2013 ao ano 2018. Os autores não informam quais artigos foram selecionados pelos critérios da pesquisa e quais foram inseridos manualmente.

O trabalho de Costa et al. (2020) apresenta um diferencial, pois foca em trabalhos que apresentam propostas para a avaliação de textos escritos em português. Além disso, o referido trabalho avalia aspectos textuais que incluem o domínio do uso da escrita formal, utilização de conhecimento de diversas áreas, saber selecionar e organizar as informações; construir a argumentação através de mecanismos linguísticos; e a elaboração de uma proposta de intervenção. No entanto, nenhum dos dois trabalhos avalia a existência de validação das propostas em ambiente real, além de não explorar outras fontes digitais como a *Engineering Village* e a *Web of Science* na etapa de busca dos artigos, as quais foram consideradas na presente pesquisa por meio do protocolo de mapeamento sistemático apresentado na seção 3.

3 Metodologia

Um Mapeamento Sistemático da Literatura (MSL) deve realizar uma avaliação crítica das pesquisas que abordam um determinado assunto e deve ter uma estrutura bem definida para que os resultados não sejam enviesados. Além disto, o rigor de um mapeamento da literatura precisa ser reforçado, reduzindo os efeitos aleatórios e garantindo a reprodutibilidade (Becheikh et al.,

2006). De acordo com Kitchenham e Charters (2007), os critérios de seleção podem ser aplicados liberalmente, considerando a avaliação da conclusão, o que foi realizado em uma etapa posterior. O mesmo ocorreu na revisão sistemática realizada por Nunes et al. (2022), que segue as diretrizes estabelecidas por Moher et al. (2009).

O MSL proposto neste artigo seguiu as diretrizes e o modelo de protocolo de mapeamento sistemático proposto por Kitchenham e Charters (2007), e incluiu três etapas principais:

1. **Etapa de planejamento**: Nessa etapa foram definidos os objetivos do mapeamento e o protocolo que foi seguido, bem como, as questões de pesquisas;
2. **Etapa de Execução**: os artigos foram buscados e selecionados e, por último, foi realizada a extração das informações e síntese dos resultados;
3. **Etapa de Relatório**: Refere-se à apresentação e discussão dos resultados.

A seção seguinte apresenta o protocolo definido e aplicado no MSL realizado neste trabalho.

3.1 Questões de Pesquisa

A avaliação automática de redações em língua portuguesa está relacionada a diversas questões como correção ortográfica, pontuação automática das redações em diferentes competências, entre outros aspectos. A seguinte questão de pesquisa foi, então, elaborada:

Questão de Pesquisa: Como a Inteligência Artificial (IA) tem sido utilizada para oferecer suporte à avaliação automática de redações?

Levando em consideração essa pergunta de pesquisa, o trabalho foi dividido nas seguintes subquestões:

Questão de Pesquisa 1 (Q1): *Quais são os principais objetivos da utilização de inteligência artificial na avaliação de redações?*

Questão de Pesquisa 2 (Q2): *Quais os principais algoritmos de inteligência artificial que são utilizados para a avaliação de redações?*

Questão de Pesquisa 3 (Q3): *Quais são as métricas mais utilizadas para validação?*

Questão de Pesquisa 4 (Q4): *Quais são os bancos de dados mais utilizados para validação?*

Questão de Pesquisa 5 (Q5): *Existe alguma evidência de que a inteligência artificial auxilia na avaliação de redações?*

Questão de Pesquisa 6 (Q6): *Quais os critérios utilizados na avaliação das redações?*

3.2 Estratégia de Busca

As palavras-chave deste trabalho foram definidas levando em consideração o idioma inglês e o português e quatro domínios: Educacional, Inteligência Artificial, Aplicabilidade e Idioma de Aplicação. As palavras-chave definidas e utilizadas são as seguintes:

- Educacional - redação (*essay*), tema (*prompt*), gramática (*grammar*);
- Inteligência Artificial - aprendizado de máquina (*machine learning*), *deep learning*, processamento de linguagem natural (*natural language processing*);
- Aplicabilidade - avaliação (*evaluation/assessment*), pontuação (*scoring*), correção (*correction*), classificação (*grading*);
- Idioma de Aplicação - português (*portuguese*).

A *string* de busca foi construída com o auxílio dos operadores lógicos OR e AND, sendo o operador OR utilizado entre as palavras-chave do mesmo domínio e o operador AND entre os diferentes domínios. A *string* de busca foi utilizada nos dois idiomas (Inglês e Português) - e abaixo tem-se um exemplo da *string* de busca final no idioma português.

(“redação” OR “tema” OR “grámatica”)
AND
(“aprendizado de máquina” OR “deep learning” OR “processamento de linguagem natural”)
AND
(“avaliação” OR “pontuação” OR "correção"OR "classificação")
AND
(“português”)

As *strings* de busca nos dois idiomas foram aplicadas nas seguintes bases de dados de artigos científicos: ACM[2], IEEExplore[3], Engineering Village[4], Science Direct[5], SpringerLink[6], Scopus[7], Web of Science[8], SBC OpenLib[9]. As bases de dados escolhidas são amplamente usadas no processo de revisão bibliográfica, sendo algumas delas já utilizadas em trabalhos anteriores, como Nau et al. (2019) que utilizaram a Science Direct e Scopus. Além disso, a base SBC-OpenLib provê uma ampla variedade de artigos em português, sendo ideal para buscar artigos relacionados à área de educação no idioma. Ademais, as bases ACM e Springer retornaram resultados relevantes para a nosso mapeamento, como os artigos Mello et al. (2021) e Ferreira Mello et al. (2022). Por fim, a base de artigos IEEE também foi utilizada por Costa et al. (2020).

3.3 Processo de Seleção e Extração

Foram definidos alguns critérios de inclusão e exclusão para selecionar os artigos que fariam parte do Mapeamento Sistemático da Literatura. O Quadro 1 apresenta os critérios de seleção utilizados neste trabalho.

[2] https://dl.acm.org/
[3] https://ieeexplore.ieee.org/
[4] https://www.engineeringvillage.com/
[5] https://www.sciencedirect.com
[6] https://link.springer.com/
[7] https://www.scopus.com/
[8] https://www.webofscience.com
[9] https://sol.sbc.org.br/

Nº	Tipo	Descrição
1	Inclusão	Estudos primários
2	Inclusão	Estudos que propõem abordagens de inteligência artificial na avaliação de redações
3	Inclusão	Estudos que analisam redações em português
4	Exclusão	Estudos secundários ou terciários
5	Exclusão	Estudos duplicados ou re-indexados
6	Exclusão	Artigos escritos em idioma diferente do português/inglês
7	Exclusão	Artigos publicados em literatura cinza
8	Exclusão	Estudos incompletos
9	Exclusão	Veículos de publicação diferentes de conferência ou *journals*

Quadro 1: Critérios de Seleção. Fonte: os autores (2022).

A etapa de seleção manual se iniciou com a leitura dos títulos e dos resumos de todos os artigos retornados da etapa de busca, com o objetivo de avaliar os artigos de maneira geral quanto à importância da aplicação para o mapeamento realizado. Os artigos que obedeceram os critérios de inclusão e os artigos que não apresentaram informações suficientes para exclusão passaram para a próxima etapa do processo de seleção. Nessa próxima etapa, os autores realizaram a leitura da introdução e das considerações finais dos artigos, com o objetivo de incluir ou excluir os artigos com base nos critérios de seleção.

Na etapa da extração, os autores leram os textos completos dos artigos retornados da etapa de seleção com o objetivo de extrair dados relevantes para responder às perguntas de pesquisa. O Quadro 2 apresenta todas as categorias dos dados extraídos dos artigos.

#	Tipo	Descrição
1	ID	Identificador único do artigo
2	Título	Título do artigo
3	Autores	Autores do artigo
4	Ano	Ano de publicação do artigo
5	Países	País do primeiro autor do artigo
6	Tipo de publicação	Conferência ou *Journal*
7	Tipo de estudo	Experimental, Estudo de Caso, Aplicação
8	Ferramentas	Ferramentas utilizadas no estudo
9	Validação em ambiente real	Se a abordagem foi validada em um ambiente real
10	Banco de dados	Informações sobre o banco de dados utilizado
11	Principais resultados	Quais são os principais resultados do artigo?
12	Limitações/Trabalhos Futuros	Quais são as limitações apontadas pelos autores?
13	Principais objetivos (Q1)	Qual o principal objetivo do artigo?
14	Algoritmos (Q2)	Quais os principais algoritmos utilizados?
15	Métricas de validação (Q3)	Quais as métricas de validação utilizadas?
16	Banco de dados (Q4)	Quais os bancos de dados utilizados na validação?
17	Evidência de melhora (Q5)	Há evidência de impacto positivo ou negativo na aplicação de avaliação automática de redação?
18	Critérios (Q6)	Quais os critérios utilizados na avaliação?

Quadro 2: Categorias dos dados extraídos. Fonte: os autores (2022).

4 Execução e Relatório

A primeira etapa da execução é a etapa de busca, onde a *string* de busca de cada idioma foi aplicada nas bases de dados de artigos científicos e em seguida foi realizado o *download* das referências dos artigos retornados. A busca final dos artigos foi realizada no mês de maio de 2022 e a Tabela 1 apresenta a quantidade de artigos retornados em cada uma das bases. Foram considerados em nossa busca artigos no período de Janeiro/2012 até Março/2022.

Tabela 1: Número de artigos retornados pela *string* de busca em cada base científica. Os artigos estão divididos em relação ao idioma a qual o artigo está escrito. O símbolo '-' significa que nenhum artigo foi encontrado. Fonte: os autores 2022.

Base Científica	**Número de Artigos**		
	Inglês	*Português*	Total
ACM	282	33	315
IEEExplore	6	-	6
Engineering Village	25	-	25
Science Direct	321	-	321
SpringerLink	494	-	494
Scopus	8	4	12
Web of Science	11	-	11
SBC-OpenLib	2	2	4
Total	1149	39	1188

A próxima etapa do processo de execução é a seleção. Nessa etapa foi utilizada a ferramenta Rayyan[10], dos autores Johnson e Phillips (2018), que oferece apoio ao desenvolvimento de revisões sistemáticas. A ferramenta Rayyan foi utilizada em trabalhos anteriores de revisão sistemática da literatura como o de Papadopoulos et al. (2020) e de Nunes et al. (2022). Foi também avaliada por diferentes trabalhos em relação à condução de revisões sintemáticas da literatura apresentando um desempenho similar a outras plataformas com o propósito análogo, sendo uma das mais precisas na detecção de artigos duplicados (Guimarães et al., 2022; McKeown e Mir, 2021).

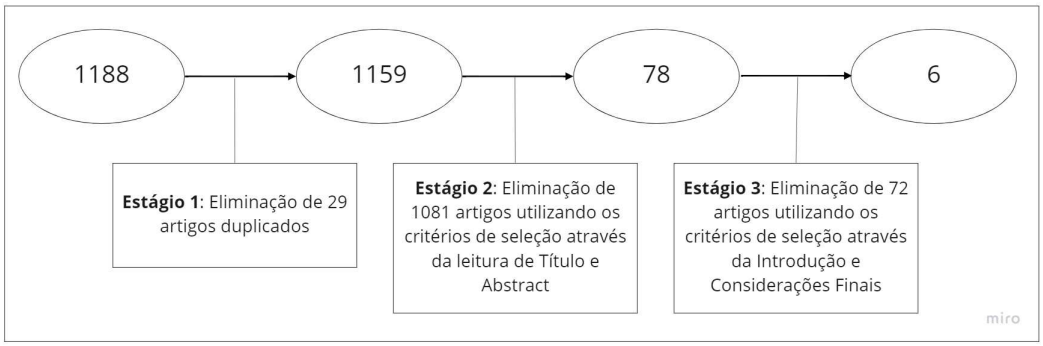

Figura 1: Fases de Seleção. Fonte: os autores (2022).

No nosso trabalho, a ferramenta auxiliou no processo de seleção em três estágios: (1) Remoção de artigos duplicados, visto que a ferramenta detecta artigos possivelmente duplicados e,

[10]https://www.rayyan.ai/

em seguida, é possível confirmar ou retificar essa informação de forma manual; (2) Identificação dos critérios de seleção através da leitura dos resumos e dos títulos; e (3) Aplicação dos critérios de seleção através da leitura da Introdução e das Considerações Finais. A Figura 4 apresenta o número de artigos selecionados em cada um desses estágios.

Dessa forma, no **Estágio 1**, 29 artigos foram detectados como duplicados pela ferramenta e conferidos e removidos da revisão pelos autores. No **Estágio 2**, os autores aplicaram os critérios de inclusão e exclusão a fim de selecionar apenas artigos de estudos primários de conferências ou revistas e que tratavam diretamente de Inteligência Artificial aplicada ao contexto de correção de redação.

Nesse estágio foram removidos cerca de 93% dos artigos e 78 artigos passaram para o próximo estágio. Os artigos que não continham informações suficientes nos resumos e nos títulos para a seleção através dos critérios de inclusão e exclusão passaram para o Estágio 3. No **Estágio 3**, os autores revisaram os artigos através da leitura da Introdução e das Considerações Finais. Nesse estágio foram selecionados 6 artigos que atenderam a todos os critérios previamente definidos para serem incluídos de forma definitiva e participarem da pesquisa.

Os trabalhos selecionados foram publicados entre os anos de 2018 e 2022. O Brasil aparece como País de publicação em 5 dos artigos selecionados, sendo que em dois deles a pesquisa foi realizada em parceria com outros países (Reino Unido, Austrália e Arábia Saudita). O país de publicação do sexto artigo é Portugal. Os 6 artigos selecionados foram publicados em conferências. O Quadro 3 apresenta uma síntese dos principais resultados encontrados relacionados às perguntas de pesquisa estudadas nesse artigo.

4.1 Principais Objetivos (Q1)

No que se refere à primeira pergunta **"Quais são os principais objetivos da utilização de inteligência artificial na avaliação de redações?"**, tais objetivos estão ligados a diferentes fatores entre eles: (i) Algoritmos de IA são mais eficientes que avaliadores humanos; (ii) não são influenciados por questões subjetivas, evitando inconsistências; e (iii) criação de *feedbacks* importantes para educadores e a geração de dados analíticos referentes ao desempenho dos alunos em sala de aula na atividade de produção textual. A partir desses objetivos, outros objetivos mais específicos são traçados, como a avaliação de diferentes abordagens, o foco em diferentes aspectos de avaliação e a melhoria nas performances.

Um dos principais objetivos identificados neste mapeamento foi a avaliação da estrutura argumentativa e a coerência de uma redação durante a defesa de um ponto de vista relacionado ao tema proposto, aparecendo como objetivo específico em 4 dos 6 trabalhos relacionados: Ferreira Mello et al. (2022), Sousa et al. (2021), Filho et al. (2018), Mello et al. (2021). Quanto aos artigos Mello et al. (2021) e Filho et al. (2019) , apesar de similares, o artigo de Ferreira Mello et al. (2022) é mais abrangente e aprofundado no assunto, explorando diversos algoritmos e *datasets* sendo, portanto, uma extensão do artigo anterior de Mello et al. (2021). Já o trabalho de Filho et al. (2019) busca resolver o problema do desbalanceamento entre classes na avaliação de uma redação quanto ao domínio da escrita formal da língua portuguesa. Por último, o trabalho de Fonseca et al. (2018) faz uma comparação entre o uso de redes neurais profundas e um sistema baseado em engenharia de atributos na avaliação de uma redação levando em consideração aspectos gerais.

Nesse sentido, pode-se concluir que os trabalhos atuais da literatura buscam por uma forma determinística e eficiente de gerar pontuações relacionadas às redações do ENEM. Sendo assim, os trabalhos propõem objetivos gerais importantes para o uso da IA no contexto de correção de redação, embora tais objetivos não sejam devidamente assegurados (como a questão de performance entre classificadores e o viés na correção de redação), tais objetivos podem servir de norte para trabalhos futuros na área.

4.2 Principais Métodos (Q2)

A segunda pergunta de pesquisa **"Quais os principais algoritmos de inteligência artificial que são utilizados para a avaliação de redações"** se concentra em desvendar os métodos utilizados na literatura. Os principais métodos de IA avaliados em português são: (i) baseados na extração de atributos, seguidos da classificação através de algoritmos de Aprendizado de Máquina (AM) supervisionado; (ii) baseados em *deep learning* em uma abordagem de classificação supervisionada.

Na primeira categoria, (i), podemos incluir os trabalhos de Ferreira Mello et al. (2022) e Mello et al. (2021), onde são explorados métodos independentes do conteúdo, que são obtidos através de ferramentas como *Coh-Metrix* e *Linguistic Inquiry Word Count (LIWC)*, e métodos dependentes do conteúdo, como Frequência do Termo e Frequência Inversa do Documento (TF-IDF, na sigla em inglês). O *Coh-Metrix* é uma ferramenta que permite a extração de diferentes atributos relacionados a diversos aspectos linguísticos como legibilidade, coesão, entre outros. Atualmente o *Coh-Metrix* possui 48 métricas e recebeu uma versão *web* desenvolvida por Camelo et al. (2020). Já a LIWC é uma ferramenta utilizada para a detecção de sentimentos no texto, que é importante nesse tipo de trabalho, pois com ela é possível detectar expressões verbais que podem conter valor semântico significativo no texto Kahn et al. (2007). Os algoritmos como *Support Vector Machine* (SVM), Árvore de Decisão e Adaboost obtiveram melhores resultados utilizando TF-IDF, ao passo que para os atributos independentes de conteúdo os resultados foram melhores usando XGBoost e *Conditional Random Fields* (CRF).

A abordagem se repete para o trabalho de Filho et al. (2018), que utiliza o classificador e o regressor do algortimo SVM para a avaliação de aderência ao tema em um texto dissertativo-argumentativo. O processo utilizado é semelhante ao de Ferreira Mello et al. (2022), onde atributos relacionados à contagem e repetições de palavras, além de outras métricas relacionadas ao domínio argumentativo, são utilizadas, totalizando 89 métricas. Apesar do estudo adotar um processo similar ao de Ferreira Mello et al. (2022), os resultados não foram satisfatórios em completude pelo fato do algoritmo falhar ao classificar pontuações intermediárias. Uma das possíveis razões talvez seja a tentativa falha de balancear as classes da base de dados com o método de *oversampling* Synthetic Minority Oversampling TEchnique (SMOTE). O trabalho de Filho et al. (2019) explora mais a fundo o problema do desbalanceamento de classes na área da avaliação automática de redações. As técnicas de balanceamento utilizadas são o SMOTE, o Adaptive Synthetic (ADASYN), o *Random Oversampling* e o *Random Undersampling*. Os algoritmos de regressão utilizados na avaliação são o *Least Absolute Shrinkage and Selection Operator (LASSO)* e o regressor SVM e os algoritmos de classificação utilizados foram o *Gradient Boosted Trees* e o classificador SVM. Os resultados do estudo apontam que as técnicas SMOTE e ADASYN são menos efetivas quando comparadas às outras técnicas de abordagem aleatória.

Na segunda categoria, (ii), temos o trabalho de Sousa et al. (2021) que considera o uso de modelo pré-treinado baseado em modelos recentes como *Transformer* e CRF. Inicialmente, esse trabalho propõe uma atividade de mineração de argumentos utilizando para isso o conjunto de dados de *Persuasive Essay corpus* contendo uma série de textos com diferentes tópicos. Após a realização de um processo de tradução dos textos, os dados foram treinados em uma abordagem de *token-level tagging* usando bidirectional *Long Short Term Memory* (LSTM) e *multilingual Bidirectional Encoder Representation Transformer* (mBERT). Nos experimentos de classificação realizados por Sousa et al. (2021), utilizando etiquetagem automática, o modelo mBERT obteve os melhores resultados, com 70.12 f1-macro para a versão traduzida em português em comparação com o algoritmo BLSTMCRF+Char (character), que obteve 68.59 f1-macro.

Por último, o trabalho de Fonseca et al. (2018) explora as duas categorias ao comparar uma abordagem baseada em *deep learning*, que utiliza camadas de LSTM bidirecionais e vetores de palavras treinados com o Global Vectors for Word Representation (GLOVE), e uma abordagem baseada em engenharia de atributos. Os algoritmos utilizados na abordagem baseada em engenharia de atributos foram o Gradient Boosting e a Regressão Linear. Nesse estudo, o Gradient Boosting apresentou o melhor resultado na avaliação das competências 1, 2, 3 e 4, enquanto a rede neural apresentou melhor performance para a competência 5.

Sendo assim, é possível concluir que a literatura explora algoritmos baseados na extração de características pré-definidas, sejam elas baseadas no conteúdo ou independentes do conteúdo, bem como, algoritmos baseados em *deep learning*. Além disso, algoritmos de AM são explorados de duas maneiras principais, sejam por meio de regressão ou classificação, com destaque para os algoritmos **Gradient Boost** e **XGBoost**.

4.3 Métricas de Avaliação (Q3)

Na terceira questão de pesquisa **"Quais as métricas utilizadas?"**, entre os estudos selecionados neste mapeamento sistemático da literatura grande parte empregou técnicas de validação dos ensaios combinando diferentes métricas. Estas foram: Precisão, Cobertura e *F-Measure*, que são métricas amplamente utilizadas no campo da aprendizagem de máquina e foram utilizadas nos trabalhos de Ferreira Mello et al. (2022), Sousa et al. (2021) e Mello et al. (2021) para avaliar os algoritmos em relação à verificação da estrutura argumentativa e a coerência das redações. Nos trabalhos de Ferreira Mello et al. (2022) e Mello et al. (2021), além das métricas mencionadas, também foi utilizado o *Kappa* de *Coehn* que é uma métrica bastante utilizada na área de mineração de dados educacionais.

Já os trabalhos de Filho et al. (2018) e de Filho et al. (2019) utilizaram matriz de confusão para fazer validações relacionadas à aderência ao tema e ao domínio formal da língua portuguesa, respectivamente. A matriz de confusão facilita a visualização dos erros produzidos pelos modelos em termos de verdadeiro positivo e falso negativo. Além da matriz de confusão, o estudo de Filho et al. (2019) também utilizou as métricas de Precisão e Cobertura e a Correlação de Pearson nas suas validações. Por último, o estudo de Fonseca et al. (2018) utilizou as métricas de Erro médio quadrático (RMSE) e o *Quadratic Weight Kappa* (QWK) para avaliar as redações de maneira geral. Essas são métricas que costumam ser utilizadas na literatura relacionada à avaliação automática de redações.

4.4 Banco de Dados (Q4)

Tabela 2: A tabela mostra o número de redações em cada um das base de dados citadas..

ID	Fonte do Banco de Dados	Número de Redações
1	Santos et al. (2018)	271
2	Haendchen Filho et al. (2018)	50
3	Sousa et al. (2021)	402
4	Redações da UOL e do Brasil Escola	1983
5	Fonseca et al. (2018)	56.644

A Tabela 2 mostra todos os conjuntos de bases de dados utilizados nos trabalhos presentes na revisão relacionados a questão de pesquisa **"Quais os bancos de dados mais utilizados para validação?"**. Na base de dados de Santos et al. (2018) e Nau et al. (2019) encontramos 271 e 50 redações, respectivamente, utilizadas em ambos trabalhos de Ferreira Mello et al. (2022) e Mello et al. (2021), que utilizou extração de características para análise de estrutura retórica. A segunda ainda foi utilizada no trabalho de Mello et al. (2021). A base de dados 3 foi utilizada no trabalho Sousa et al. (2021) para análise de estrutura argumentativa e persuasiva. A base de dados era originalmente escrita em inglês, mas foi traduzida para Português a fim de realizar-se as análises pretendidas em relação à estrutura argumentativa do texto. As bases 4 e 5 são redações no formato do ENEM. No caso da base 4, os dados foram extraídos de redações disponíveis na internet através de *web crawling*, enquanto a base de dados 5 foi obtida através da escrita de redações avaliadas por profissionais de educação em uma plataforma e pontuadas de acordo com as competências do ENEM (Fonseca et al., 2018).

4.5 Evidência de Melhora (Q5)

A quinta questão de pesquisa, **"Existe alguma evidência de que a aprendizagem de máquina auxilia na avaliação de redações?"**, proposta neste mapeamento, tinha como objetivo responder se há evidência de que o uso de estratégias de IA trazem melhorias ao campo da avaliação automática de redações. Apesar dos resultados promissores apresentados por todos os estudos incluídos, o que se pôde perceber foi uma lacuna no que tange à validação em ambientes reais, já que nenhum dos artigos selecionados apresentou esse tipo de validação. Inúmeros fatores podem justificar essa falta de comunicação entre a comunidade científica e o mundo real, como a falta de investimentos, falta de padronização e/ou estruturação nas bases disponíveis e até a própria magnitude de possibilidades de aplicação de estratégias de IA, que pode acabar motivando os cientistas a experimentar novas variações e/ou abordagens em contextos distintos em vez do aprofundamento de uma pesquisa que já obteve resultados preliminares.

Não obstante, o trabalho realizado por Nunes et al. (2022) investigou a existência de trabalhos que avaliam empiricamente a efetividade dos algoritmos de AES. Afinal, apenas 8 artigos foram selecionados considerando um espaço temporal de 20 anos (2000-2020), apontando apenas dois trabalhos fora dos Estados Unidos da América (EUA) e com a maioria deles avaliando estudantes que estão entre a educação primária e secundária. O período das avaliações realizadas nos trabalhos varia desde alguns dias a meses, obtendo-se avaliação positiva do uso da correção automática em 7 de 8 deles.

O trabalho Wilson e Roscoe (2020), por exemplo, avaliou alunos que tem o inglês como

língua nativa com idade média de aproximadamente 11 anos aplicando um pré e pós testes em uma análise quantitativa. Os estudantes foram separados em dois grupos, Experimental (E) e de Controle ativo (C), com 56 (64% meninas) e 58 (62% meninas) alunos cada um, respectivamente. O grupo experimental recebeu *feedback* do sistema de correção automática *Project Essay Grade* (PGE, sigla em inglês) em relação à gramática, desenvolvimento da ideia, organização, estrutura da sentença, escolha da palavra, convenções e estilo. Por fim, os alunos poderia receber *feedbacks* adicionais dos professores através de um *chat* na plataforma. Os resultados foram comparados com alunos que usaram o Google Docs e receberam o *feedback* dos professores sem utilizarem um sistema automático. Os resultados mostraram que apesar da inexistência de diferença significativa em relação à nota final de ambos os grupos no pós-teste, os alunos que usaram o sistema de correção automática se tornaram mais confiantes. Por fim, 3 educadores também tiveram uma impressão positiva do sistema em relação à usabilidade, e qualidade e quantidade dos *feedbacks* dados aos alunos pela plataforma.

Assim sendo, há uma percepção positiva tanto para alunos, através da melhora do processo de escrita ao gerar maior confiança, quanto para professores ao permitir focá-los apenas em ensinar e diminuindo o esforço em corrigir os textos. Apesar disso, os professores perceberam que estudantes com mais dificuldades em relação à escrita podem não ser capazes de utilizar as ferramentas pelo fato de produzirem textos curtos demais e com uma quantidade de erros excessiva (Palermo e Thomson, 2018; Tang e Rich, 2017; Ware, 2014; Nunes et al., 2022). Portanto, no geral, o uso de um sistema de AES apresenta benefícios significativos quando aplicados no mundo real, embora hajam poucos artigos que avaliam tal aspecto. Alguns pontos são destacados em relação a esses trabalhos, como a não consideração do processo de integração e aprendizado da plataforma como sendo um fator a influenciar, bem como, a ausência de um processo rigoroso para detectar explicações mais diversificadas.

4.6 Critérios Avaliados (Q6)

Na última questão de pesquisa, **"Quais os critérios utilizados na avaliação das redações?"**, buscou-se avaliar quais são os aspectos linguísticos que são utilizados nos trabalhos relacionados à avaliação automática de redações.

Nos artigos de Filho et al. (2018), Fonseca et al. (2018) e Filho et al. (2019) foram utilizados aspectos relacionados às competências exigidas no Exame Nacional do Ensino Médio (ENEM). No estudo de Fonseca et al. (2018) foram utilizadas as 5 competências do ENEM. No estudo de Filho et al. (2018) foi utilizada a competência 2, que investiga se a redação escrita está relacionada ao tema proposto. Já no estudo de Filho et al. (2019) foi explorada a competência 1, que está relacionada ao domínio da escrita formal da língua portuguesa.

Os estudos de Ferreira Mello et al. (2022), Sousa et al. (2021) e Mello et al. (2021) exploram aspectos linguísticos que são exigidos no ENEM, mais especificamente as competência 3 e 4. O estudo de Sousa et al. (2021) explora o aspecto relacionado à argumentação, que pode ser relacionado às competências 3 e 4 do ENEM, onde se avalia a organização de fatos e opiniões em defesa de um ponto de vista e a demonstração do conhecimento linguístico necessário para construir a argumentação. Os estudos de Ferreira Mello et al. (2022) e Mello et al. (2021) expandiram a avaliação para aspectos para além da coesão, coerência, incluindo também legibilidade e relações semânticas presentes nos textos, que são aspectos esperados em uma redação desenvolvida para

esse exame.

5 Considerações Finais

Este artigo apresenta uma visão geral dos estudos relacionados à correção automática de redação que foram coletados por meio de um Mapeamento Sistemático da Literatura. Fazendo um pequeno recorte sobre as abordagens de inteligência artificial que são utilizadas para realizar a avaliação automática de redações, conclui-se que o principal objetivo da utilização da inteligência artificial nessa área é a busca por uma forma determinística e eficiente de gerar pontuações associadas a redações escritas por alunos que buscam ingressar no ensino superior.

Um dos aspectos identificados nos trabalhos selecionados é o baixo uso de atributos extraídos de forma automatizada através de técnicas de *deep learning*. Apenas dois trabalhos exploram essa abordagem, apesar de ser uma estratégia bastante utilizada em outras áreas de Processamento de Linguagem Natural (PLN). Levando em consideração as abordagens utilizadas na validação das estratégias propostas, foi identificado que os trabalhos nem sempre utilizam análises estatísticas, como testes de hipótese, para assegurar a eficiência de um modelo quando comparado com outro. Outro fato identificado é que não há uma avaliação de diferentes abordagens em relação à performance computacional, embora esse aspecto seja citado como uma das motivações por trás do uso de algoritmos de IA.

Outra lacuna identificada entre os trabalhos selecionados é o baixo detalhamento nos *feedbacks* retornados pelos modelos de avaliação. A geração de *feedbacks* a partir das análises dos algoritmos é algo de extrema importância quando leva-se em consideração o contexto educacional e o propósito didático que algumas abordagens podem assumir em trabalhos futuros. Além disso, não foram encontradas investigações relacionadas ao viés que pode ser introduzido na correção automática de redações, ainda que esse seja um problema encontrado em outras tarefas ligadas ao PLN.

Por último, foi identificado que nenhum dos trabalhos fez a validação dos seus resultados em um ambiente real e nenhum trabalho utilizou a abordagem proposta no desenvolvimento de uma plataforma que realizasse a avaliação automática de redações. Portanto, os objetivos traçados pelos artigos selecionados acabam por não serem totalmente atingidos ou verificados em sua completude. Assim sendo, existem amplas oportunidades que podem ser exploradas em trabalhos futuros nessa área.

Referências

Becheikh, N., Landry, R., & Amara, N. (2006). Lessons from innovation empirical studies in the manufacturing sector: A systematic review of the literature from 1993–2003. *Technovation*, *26*(5-6), 644–664. https://doi.org/10.1016/j.technovation.2005.06.016. [GS Search]

Camelo, R., Justino, S., & Mello, R. (2020). Coh-Metrix PT-BR: Uma API web de análise textual para a educação. *Anais dos Workshops do IX Congresso Brasileiro de Informática na Educação*, 179–186. https://doi.org/10.5753/cbie.wcbie.2020.179. [GS Search]

Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., Chen, K., Mitchell, R., Cano, I., Zhou, T., et al. (2015). Xgboost: extreme gradient boosting. *R package version 0.4-2*, *1*(4), 1–4. https://doi.org/10.1145/2939672.2939785. [GS Search]

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297. https://doi.org/10.1007/BF00994018. [GS Search]

Costa, L., Oliveira, E. H. T. d., & Castro Júnior, A. (2020). Corretor Automático de Redações em Língua Portuguesa: um mapeamento sistemático de literatura. *Anais do XXXI Simpósio Brasileiro de Informática na Educação (SBIE 2020)*, 1403–1412. https://doi.org/10.5753/cbie.sbie.2020.1403. [GS Search]

Ferreira Mello, R., Fiorentino, G., Oliveira, H., Miranda, P., Rakovic, M., & Gasevic, D. (2022). Towards automated content analysis of rhetorical structure of written essays using sequential content-independent features in Portuguese. *LAK22: 12th International Learning Analytics and Knowledge Conference*, 404–414. https://doi.org/10.1145/3506860.3506977. [GS Search]

Filho, A. H., Concatto, F., Nau, J., Prado, H. A. d., Imhof, D. O., & Ferneda, E. (2019). Imbalanced Learning Techniques for Improving the Performance of Statistical Models in Automated Essay Scoring. *Procedia Computer Science*, *159*, 764–773. https://doi.org/10.1016/j.procs.2019.09.235. [GS Search]

Filho, A. H., do Prado, H. A., Ferneda, E., & Nau, J. (2018). An approach to evaluate adherence to the theme and the argumentative structure of essays. *Procedia Computer Science*, *126*, 788–797. https://doi.org/10.1016/j.procs.2018.08.013. [GS Search]

Fonseca, E., Medeiros, I., Kamikawachi, D., & Bokan, A. (2018). Automatically Grading Brazilian Student Essays [Series Title: Lecture Notes in Computer Science]. Em A. Villavicencio, V. Moreira, A. Abad, H. Caseli, P. Gamallo, C. Ramisch, H. Gonçalo Oliveira & G. H. Paetzold (Ed.), *Computational Processing of the Portuguese Language* (pp. 170–179). Springer International Publishing. https://doi.org/10.1007/978-3-319-99722-3_18. [GS Search]

Guimarães, N. S., Ferreira, A. J., Silva, R. d. C. R., de Paula, A. A., Lisboa, C. S., Magno, L., Ichiara, M. Y., & Barreto, M. L. (2022). Deduplicating records in systematic reviews: there are free, accurate automated ways to do so. *Journal of Clinical Epidemiology*, *152*, 110–115. https://doi.org/j.jclinepi.2022.10.009. [GS Search]

Haendchen Filho, A., do Prado, H. A., Ferneda, E., & Nau, J. (2018). An approach to evaluate adherence to the theme and the argumentative structure of essays. *Procedia Computer Science*, *126*, 788–797. https://doi.org/10.1016/j.procs.2018.08.013. [GS Search]

Johnson, N., & Phillips, M. (2018). Rayyan for systematic reviews. *Journal of Electronic Resources Librarianship*, *30*(1), 46–48. https://doi.org/10.1080/1941126X.2018.1444339. [GS Search]

Kahn, J. H., Tobin, R. M., Massey, A. E., & Anderson, J. A. (2007). Measuring emotional expression with the Linguistic Inquiry and Word Count. *The American journal of psychology*, *120*(2), 263–286. [GS Search].

Kitchenham, B. A., & Charters, S. (2007). *Guidelines for performing Systematic Literature Reviews in Software Engineering* (rel. técn. EBSE 2007-001). Keele University e Durham University Joint Report. [GS Search].

Marinho, J., Anchiêta, R., & Moura, R. (2021). Essay-BR: a Brazilian Corpus of Essays. *Anais do III Dataset Showcase Workshop*, 53–64. https://doi.org/10.5753/dsw.2021.17414. [GS Search]

McKeown, S., & Mir, Z. M. (2021). Considerations for conducting systematic reviews: evaluating the performance of different methods for de-duplicating references. *Systematic reviews*, *10*, 1–8. https://doi.org/10.1186/s13643-021-01583-y. [GS Search]

Mello, R. F., Fiorentino, G., Miranda, P., Oliveira, H., Raković, M., & Gašević, D. (2021). Towards Automatic Content Analysis of Rhetorical Structure in Brazilian College Entrance Essays [Series Title: Lecture Notes in Computer Science]. Em I. Roll, D. McNamara, S. Sosnovsky, R. Luckin & V. Dimitrova (Ed.), *Artificial Intelligence in Education* (pp. 162–167). Springer International Publishing. https://doi.org/10.1007/978-3-030-78270-2_29. [GS Search]

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & PRISMA Group*, t. (2009). Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *Annals of internal medicine*, *151*(4), 264–269. https://doi.org/10.7326/0003-4819-151-4-200908180-00135. [GS Search]

Nau, J., Haendchen Filho, A., & Dazzi, R. L. S. (2019). Identificação e Avaliação Automática da Proposta de Intervenção em Textos Dissertativos-Argumentativos: Uma Revisão Sistemática da Literatura. *Anais do Computer on the Beach*, 493–501. https://doi.org/10.4013/cld.2017.153.08. [GS Search]

Nunes, A., Cordeiro, C., Limpo, T., & Castro, S. L. (2022). Effectiveness of automated writing evaluation systems in school settings: A systematic review of studies from 2000 to 2020. *Journal of Computer Assisted Learning*, *38*(2), 599–620. https://doi.org/10.1111/jcal.12635. [GS Search]

Palermo, C., & Thomson, M. M. (2018). Teacher implementation of self-regulated strategy development with an automated writing evaluation system: Effects on the argumentative writing performance of middle school students. *Contemporary Educational Psychology*, *54*, 255–270. https://doi.org/10.1016/j.cedpsych.2018.07.002. [GS Search]

Papadopoulos, I., Koulouglioti, C., Lazzarino, R., & Ali, S. (2020). Enablers and barriers to the implementation of socially assistive humanoid robots in health and social care: a systematic review. *BMJ open*, *10*(1), e033096. https://doi.org/10.1136/bmjopen-2019-033096. [GS Search]

Safavian, S. R., & Landgrebe, D. (1991). A survey of decision tree classifier methodology. *IEEE transactions on systems, man, and cybernetics*, *21*(3), 660–674. https://doi.org/10.1109/21.97458. [GS Search]

Santos, K. S., Soder, M., Marques, B. S. B., & Feltrim, V. D. (2018). Analyzing the rhetorical structure of opinion articles in the context of a Brazilian college entrance examination. *Computational Processing of the Portuguese Language: 13th International Conference, PROPOR 2018, Canela, Brazil, September 24–26, 2018, Proceedings 13*, 3–12. https://doi.org/10.1007/978-3-319-99722-3_1. [GS Search]

Sousa, A., Leite, B., Rocha, G., & Lopes Cardoso, H. (2021). Cross-Lingual Annotation Projection for Argument Mining in Portuguese [Series Title: Lecture Notes in Computer Science]. Em G. Marreiros, F. S. Melo, N. Lau, H. Lopes Cardoso & L. P. Reis (Ed.), *Progress in Artificial Intelligence* (pp. 752–765). Springer International Publishing. https://doi.org/10.1007/978-3-030-86230-5_59. [GS Search]

Tang, J., & Rich, C. S. (2017). Automated writing evaluation in an EFL setting: Lessons from China. *JALT CALL Journal*, *13*(2), 117–146. https://doi.org/10.29140/jaltcall.v13n2.215. [GS Search]

Ware, P. (2014). Feedback for Adolescent Writers in the English Classroom. *Writing & Pedagogy*, *6*(2). https://doi.org/10.1558/wap.v6i2.223. [GS Search]

Wilson, J., & Roscoe, R. D. (2020). Automated writing evaluation and feedback: Multiple metrics of efficacy. *Journal of Educational Computing Research*, *58*(1), 87–125. https://doi.org/10.1177/0735633119830764. [GS Search]

Artigo	Objetivos(Q1)	Algoritmos(Q2)	Validação(Q3)	Banco de Dados(Q4)	Evidência (Q5)	Critérios *(Q6)
Ferreira Mello et al. (2022)	Estrutura retórica	SVM, Random Forest, Adaboost, XGBoost e CRF	Kappa, Cobertura, Precisão e F-Measure	Banco de dados de Santos et al. (2018) e Haendchen Filho et al. (2018)	Não	Competências 3 e 4
Sousa et al. (2021)	Análise de argumentação	BLSTM e CNN	Cobertura, Precisão e F-Measure	Banco próprio	Não	Competência 3 e 4
Filho et al. (2018)	Fuga ao tema	R-SVM e C-SVM	Precisão e correlação	Redações da UOL e do Brasil Escola	Não	Competência 2
Fonseca et al. (2018)	Avaliação das competências do ENEM	BLSTM e CNN	QWK e RMSE	Banco próprio	Não	Todas as competências
Filho et al. (2019)	Domínio formal do português	SVM e GBT	Relação de verdadeiro positivo	Banco da UOL	Não	Competência 1
Mello et al. (2021)	Estrutura retórica	SVM, Random Forest, Adaboost, XGBoost e CRF	Kappa, Cobertura, Precisão e F-Measure	Santos et al. (2018).	Não	Competências 3 e 4

Quadro 3: Resumo das cinco perguntas de pesquisa.* As competências citas são do ENEM. Fonte: os autores (2022).

Sequence Labeling Algorithms for Punctuation Restoration in Brazilian Portuguese Texts

Tiago B De Lima[1][0000−0002−0707−522X], Pericles Miranda[1], Rafael Ferreira Mello[1], Moesio Wenceslau[1], Ig Ibert Bittencourt[2], Thiago Damasceno Cordeiro[2], and Jário José[2]

[1] Universidade Federal Rural de Pernambuco, Rua Dom Manuel de Medeiros, Recife, Pernambuco, 52171-900, Brazil
{tiago.blima,rafael.mello,moesio.wenceslau, pericles.miranda}@ufrpe.br
[2] Av. Lourival Melo Mota, S/N - Cidade Universitária, Maceió - AL, 57072-970
ig.ibert@gmail.com, {thiago,jjsj@ic}@ic.ufal.br

Abstract. Punctuation Restoration is an essential post-processing task of text generation methods, such as Speech-to-Text (STT) and Machine Translation (MT). Usually, the generation models employed in those tasks produce unpunctuated text, which is difficult for human readers and might degrade the performance of many downstream text processing tasks. Thus, many techniques exist to restore the text's punctuation. For instance, approaches based on Conditional Random Fields (CRF) and pre-trained models, such as the Bidirectional Encoder Representations from Transformers (BERT), have been widely applied. In the last few years, however, one approach has gained significant attention: casting the Punctuation Restoration problem into a sequence labeling task. In Sequence Labeling, each punctuation symbol becomes a label (e.g., COMMA, QUESTION, and PERIOD) that sequence tagging models can predict. This approach has achieved competitive results against state-of-the-art punctuation restoration algorithms. However, most research focuses on English, lacking discussion in other languages, such as Brazilian Portuguese. Therefore, this paper conducts an experimental analysis comparing the Bi-Long Short-Term Memory (BI-LSTM) + CRF model and BERT to predict punctuation in Brazilian Portuguese. We evaluate those approaches in the IWSLT 2012-03 and OBRAS dataset in terms of precision, recall, and F_1-score. The results showed that BERT achieved competitive results in terms of punctuation prediction, but it requires much more GPU resources for training than the BI-LSTM + CRF algorithm.

1 Introduction

Punctuation Restoration, also called Punctuation Prediction [14], is the task to insert missing punctuation marks in a sequence of unpunctuated text [22,19]. It is an important step in the post-processing of Speech-To-Text (STT), which usually does not take punctuation into account [2], and other text-generation techniques, such as Machine Translation (MT) [28]. Thus, various methods have been

considered in the last years for solving this problem, including traditional Machine Learning (ML) algorithms, Deep Learning techniques [22], and pre-trained methods, such as the Bidirectional Encoder Representations from Transformers (BERT) [7].

The recent literature characterizes the punctuation restoration problem as a Sequence Labeling task [29,19,22]. In this approach, every token is assigned to one of the possible labels [29], which usually represents the existence of a punctuation mark (e.g., COMMA, PERIOD, QUESTION) or no punctuation at all (e.g., EMPTY). This approach has achieved state-of-the-art results when combined with Neural Networks [22]. However, it is also possible to apply this approach with other techniques, such as Bidirectional encoders [26], Conditional Random Fields (CRF) [16], and pre-trained models (e.g., BERT) [22].

Although sequence labeling has obtained good results in punctuation restoration, the analysis of text produced in languages other than English is not well explored in the literature. Specifically, sequence labeling approaches for punctuation restoration in Brazilian Portuguese are still a gap in the community.

Thus, this paper explores different sequence labeling methods for punctuation restoration in Brazilian Portuguese texts. More specifically, we compare the performance of the BI-LSTM with CRF model [23] against the BERT, pre-trained in Portuguese, model [24] in two different datasets: the International Workshop on Spoken Language Translation (IWSLT) 2012, which consists of TED talks transcripts, and OBRAS, a corpus of Brazilian Portuguese literature texts. We compare the models' performance in terms of precision, recall, and F_1-score.

The results show that sequence labeling with the pre-trained BERT model is a promising approach for punctuation restoration in Brazilian Portuguese texts, surpassing the BI-LSTM+CRF model in most cases of the IWSLT 2012 dataset. When tested in the out-of-domain dataset OBRAS, the BERT model reached competitive results compared to other approaches in the literature.

The rest of the paper is structured as follows: Section 2 provides an overview of deep learning techniques applied to punctuation restoration; Section 3 discusses related works in the literature; Section 4 describes the datasets, algorithms, and experimental settings; Section 5 details the experimental results; and, lastly, Section 6 presents the conclusions and future research directions.

2 Preliminaries

There are a variety of approaches for solving the punctuation restoration task, ranging from rule-based methods to deep learning techniques [22]. This section covers some learning algorithms applied to punctuation restoration over the years.

Deep learning (DL) has become increasingly popular over the last years, gaining significant attention as part of many Natural Language Processing (NLP) tasks and frameworks [8]. In this context, Sequence to Sequence Neural Networks [25] are a promising approach for NLP tasks due to the range of their applications, such as Machine Translation (MT) [30], Text Simplification (TS)

[4], and Question Answer (QA) [5]. Furthermore, they can be adapted to other domains and are the building blocks of Deep Encoder representations [27].

For punctuation restoration, the Deep Learning approach has played an important role recently, gaining significant attention and leading with state-of-art results [22]. Specifically, BI-LSTM [21] and, more recently, BERT [2] are two viable and promising approaches. Therefore, we will briefly explore those algorithms in the following subsections.

2.1 BI-LSTM

The Bidirectional LSTM (BI-LSTM) is a Recurrent Neural Network (RNN) that can learn both forward and backward temporal information from an input sequence [10]. This characteristic improved learning in many tasks, such as Machine Translation [13] and Sequence Labeling [10].

However, BI-LSTMs suffer from learning in big datasets due to high memory requirements [27]. Thus, Attention Mechanisms are an alternative to mitigate those issues by allowing the network to focus on capturing only essential information for the task, improving learning and stability [27].

Furthermore, it is also possible to combine BI-LSTMs with CRF networks, resulting in a BI-LSTM-CRF network, which boosts accuracy in some sequence labeling tasks [10].

2.2 BERT and Self-Attention

The Bidirectional Encoder Representation from Transformers (BERT) is a pre-trained language model [7] that has achieved state-of-art results in many NLP tasks by simply appending extra output layers to the original model, and fine-tuning to the task [7,21,13]. BERT is built on top of an attention mechanism known as Self-Attention [7] and can extract patterns and features by conditionally masking and predicting tokens in a large set of corpus [7]. In the following paragraphs, we explain BERT's input/output representation and its attention mechanisms.

In BERT, the input consists of a series of tokens, and the word encoding process is as follows [7]: First, three distinct embedding strategies, namely token encoding, sentence encoding, and positional encoding, process the input. Then, the output is the sum of those encodings for each token. Figure 1 depicts this process.

As aforementioned, BERT is built on top of an attention mechanism called Self-Attention, which relates different positions of a sequence to represent the sequence better [27]. In general, attention is achieved through attention functions mapping a query (Q, a vector) and a set of key-value pairs (K and V, respectively, both vectors) to an output [27]. A simple attention function is the Scaled Dot-Product Attention [27] shown in Equation 1, where d_k is the number of dimensions of the key. Specifically, in self-attention layers, Q, K, and V are from the previous layer of the network, which enables the network to account for all sequence positions processed by previous layers [27].

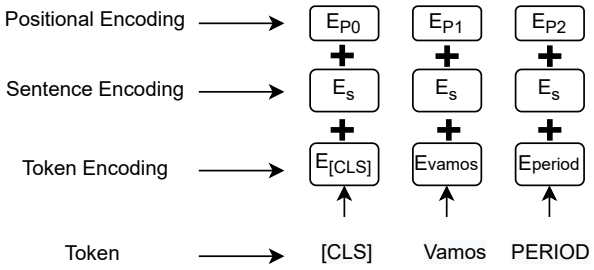

Fig. 1. A simple BERT encoder layer.

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{1}$$

3 Related Works

Speech-to-text and other text-generation tasks procedures often fail to produce grammatically correct and readable text outputs from a punctuation point of view [12,26], motivating the development of algorithms to perform the restoration of the missing punctuation [22].

One of the most successful strategies for handling punctuation restoration is to consider the problem as a sequence labeling task [19,22], whose objective is to classify tokens in the sequence to pre-defined labels or classes [6]. For punctuation restoration, the classes indicate if a given token should have a punctuation mark inserted after it [22].

Table 1 shows an example of punctuation restoration through label prediction in a text sequence where the classes are COMMA, PERIOD, QUESTION, and O (i.e., no punctuation, also referred to as the EMPTY class).

Table 1. Example of punctuation restoration by sequence labeling.

Input	Carlos	could	you	please	come	here
Prediction	COMMA	O	COMMA	COMMA	O	QUESTION
Output	Carlos,	could	you,	please,	come	here?

However, there are other possible approaches for punctuation restoration that do not necessarily rely on sequence labeling. For instance, there are deterministic rule-based alternatives [22], N-gram language models [22], and Machine Translation (MT) approaches [26].

Furthermore, even in the sequence labeling approach, there are slight variations in the applied methods. For instance, [16] considered Conditional Random Fields (CRF), achieving good results in both English and Chinese languages. However, CRF relies on minimal linguistics assumptions, which might not be satisfied in all applications. On the other hand, ML and Deep Learning classifiers have been extensively considered [22] and have almost no linguistic assumptions. For ML methods, both single model approaches, such as BI-LSTM [26], and multi-model approaches, such as BI-LSTM with CRF layers [10], are common in the literature.

However, deep learning-based algorithms are far the preferred approach for punctuation restoration in the last years, leading with state-of-the-art results by applying sequence labeling to punctuation restoration [22]. Table 2 summarizes popular ML, and Deep Learning approaches for Punctuation Restoration present in the literature. In the following paragraphs, we briefly overview some of these methods.

Table 2. Popular ML and Deep Learning approaches for Punctuation Restoration.

Reference	Method	Dataset	Language	Metrics
[26]	BI-LSTM	IWSLT2011	English, Estonian	Precision, Recall, F_1-score, slot error rate (SER)
[11]	LSTM	MGB Challenge dataset (MGB-1)	English	Precision, Recall, F_1-score, Perplexity
[16]	Factorial CRF	IWSLT09	English, Chinese	Precision, Recall, F_1-score
[18]	BERT + BI-LSTM + CRF	IWSLT2012	English	Precision, Recall, F_1-score
[14]	Multilingual LSTM	SpeechRecognition TV show	43 Languages	Precision, Recall, F_1-score
[19]	BERT	IWSLT2012	English, Hugarian	Precision, Recall, F_1-score
Our	BERT BI-LSTM+CRF	IWSLT2012 OBRAS	Brazilian Portuguese	Precision, Recall, F_1-score

Tilk and Alumäe [26] considered a BI-LSTM with attention for punctuation restoration both in English and Estonian. The results showed that Global Vectors for Word Representation (GloVe) [20], an unsupervised algorithm for obtaining vector representations of worlds, significantly improved the performance of the model.

Ondřej et al. [11] present a different approach by casting the punctuation restoration problem as a Machine Translation task of translating non-punctuated text to punctuated text.

Makhija et al. [18] used minimal annotated data with a BI-LSTM + CRF model combined with BERT embedding mechanisms, providing contextual information to the sequence tagging task.

Li and Lin [14] present a new approach: multilingual punctuation restoration. The main idea was to use byte-pair encoding, which allows sharing of information across different related languages. They considered a set of 43 languages, including English, Spanish, France, Italian, and Portuguese, and achieved good results. Specifically, the model achieved more than 80% F_1-score in Portuguese.

Nagy et al. [19] considered BERT, trained with the IWSLT 2012-03 dataset, for the English and treebank for the Hungarian language. It achieved competitive results with the state-of-the-art models, showing the potential of BERT, a pre-trained model, to address punctuation restoration for different languages.

However, although there are works considering languages other than English, few addressed the punctuation prediction for Brazilian Portuguese, and Portuguese [14]. Thus, to fill these gaps in the community, we address punctuation restoration for Brazilian Portuguese texts. Specifically, we investigate learning algorithms (BERT and BI-LSTM + CRF) for punctuation restoration, as a sequence labeling task, in Brazilian Portuguese. Table 2 summarizes all related work and contrasts them with the current research (last line).

4 Materials and Methods

This work assesses the BI-LSTM + CRF and BERT algorithms for punctuation restoration in Brazilian Portuguese texts. The algorithms were chosen due to their promising results in other languages [26,18,19], and the lack of works applying them to Brazilian Portuguese.

In order to evaluate those algorithms, we first define an annotated corpus for training and evaluation. Details of such corpus are explained in subsection 4.1. Afterward, we configure and train the algorithms for the task. Hyperparameters and details are explained in subsection 4.2. Lastly, we explain the evaluation metrics and critical difference analysis in subsection 4.3.

4.1 Datasets

In the experiments, we considered two datasets: the IWSLT 2012-03 dataset[3], and the OBRAS corpus[4]. Following a sequence labeling design [19], each token (word) of the datasets are annotated according to the following classes:

- COMMA, for commas, and dash marks;
- PERIOD, for periods, semicolons, and exclamation marks;
- QUESTION, for question marks;
- O, for no punctuation (i.e., the EMPTY class).

The IWSLT 2012-03 dataset comprises speech-to-text transcripts of TED talk presentations, including a Brazilian Portuguese version. We used the NLTK

[3] Available at: `https://wit3.fbk.eu/2012-03`
[4] Available at: `https://www.linguateca.pt/OBRAS/OBRAS.html`

tokenize package to obtain words after applying the following punctuation conversions: (i) semicolons to periods; (ii) exclamation marks to periods; and (iii) dash marks to commas. Additionally, all words were converted to lower case to avoid bias in the prediction [19], and each document was considered unique to build the TRAIN, DEV, and TEST sets. Table 3 shows the number of sentences, words, and labels for the IWSLT dataset for each set (TRAIN, DEV, or TEST).

Table 3. Number of sentences, words, and labels for the IWSLT dataset.

Labels	TRAIN	DEV	TEST
O	1,929,873	14,069	22,208
COMMA	169,384	1,169	2,270
PERIOD	147,379	935	1,721
QUESTION	11,595	87	152
Sentences	139,653	1,570	887
Words	2,258,231	16,260	26,351

The OBRAS dataset contains a range of different Brazilian Portuguese literature texts available in the open domain. The tokenization process was analogous to that applied in the IWSLT data, also using the NLTK tokenize package. Besides, we also converted all words to lower case. It is worth mentioning that this dataset was unavailable during the training phase of the algorithms. That is the reason we considered it an out-of-domain dataset. Table 4 shows the number of sentences, words, and labels of the OBRAS dataset.

Table 4. Number of sentences, words, and labels for the OBRAS dataset.

Labels	Count
O	2,298,811
COMMA	303,424
PERIOD	202,573
QUESTION	15,380
Sentences	193,236
Words	2,820,188

4.2 Compared Algorithms

The BI-LSTM + CRF with pre-trained embedding and the BERT model are two famous and promising approaches for Punctuation Restoration in the literature [22]. Nonetheless, the performance of these algorithms has yet to be investigated

for punctuation restoration in Brazilian Portuguese texts. In this work, we evaluate those two algorithms using the CRF model as a baseline. In the following paragraphs, we explain their configurations.

CRF: We consider the CRF as a baseline algorithm for sequence labeling and punctuation restoration, as done in [16,17]. We used the same features as in [17], all hyperparameters were set empirically and are shown in Table 5.

Table 5. Hyperparameters for CRF.

Hyperparameter	**Value**
Algorithm	L-BFGS
c_1	0.1
c_2	0.1
Max Iterations	100
All Possible Transitions	TRUE

BI-LSTM: The BI-LSTM model was trained with Word2Vec skip-gram 300 (BI-LSTM+Skip$_s$300) [9] and early stopping. The hyperparameters were set empirically, and are shown in Table 6. We used the implementation present in the FLAIR Framework [1].

Table 6. Hyperparameters and training configuration for the BI-LSTM.

Hyperparameter	**Value**
Learning Rate	0.1
Max Epochs	100
Mini-batch size	32
Patience	3
Annealing factor	0.5
Shuffle	TRUE
Training with DEV	FALSE
Batch Growth Annealing	FALSE

BERT: The experiments used the pre-trained BERT model present in the Simple Transformers Framework[5]. For the fine-tuning phase of BERT, the hyperparameters were set empirically and are shown in Table 7.

4.3 Evaluation

We evaluated the performance of the algorithms in terms of Precision (P), Recall (R), and F_1-score. Those evaluation metrics are defined as follows:

[5] https://simpletransformers.ai/

Table 7. Hyperparameters and model setup for BERT.

Parameter	Value
Early stopping	TRUE
Num. train epochs	12
Train batch size	16
Eval batch size	8

$$\text{Precision} = \frac{TP}{(TP + FP)}, \tag{2}$$

$$\text{Recall} = \frac{TN}{(TN + FN)}, \tag{3}$$

$$F_1 = \frac{2 \times \text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}} \tag{4}$$

where TP are true positives, TN are true negatives, FP are false positives and FN are false negatives.

We decided to use those metrics to evaluate the capacity of each model to not only predict correct punctuation (precision) but also to recall missing predictions in the original text. Except when stated otherwise, the results refer to a single run of the model in the respective dataset. All experiments were run in the Google Colaboratory.

5 Results and Discussion

In this section, we present and discuss the results. From here on, except when stated otherwise, we will refer to the models by their algorithms' name, that is, we refer to BERT-BASE as BERT, and BI-LSTM+Skip$_s$300 by BI-LSTM.

In subsection 5.1, we present and discuss the performance of the algorithms in the IWSLT 2012-03 dataset. Additionally, in subsection 5.2, we evaluate the performance of the algorithms in the out-of-domain dataset OBRAS, and provide a cost analysis of each algorithm in subsection 5.3.

5.1 Evaluation on IWSLT 2012-03 dataset

Table 8 presents the obtained results in each evaluation metric for every class in the task. As it can be seen, the BERT algorithm outperformed all other competitors in recall (R) and F_1-score. Besides, BERT also achieved superior precision for COMMA and QUESTION.

Table 9 presents the overall results, called *micro average metrics*, reached by the models. The BERT model achieved better average performance for all metrics, with 0.833 precision, 0.789 recall, and 0.810 F_1-score.

Table 8. Precision, Recall and F_1-score, in a single run, for all classes representing punctuation marks.

	COMMA			PERIOD			QUESTION		
	P	R	F_1	P	R	F_1	P	R	F_1
CRF	0.556	0.306	0.395	0.869	0.836	0.852	0.318	0.046	0.080
BI-LSTM+Skip$_s$300	0.670	0.530	0.592	**0.924**	0.842	0.881	0.750	0.572	0.649
BERT-BASE	**0.770**	**0.719**	**0.744**	0.911	**0.887**	**0.899**	**0.844**	**0.711**	**0.771**

Table 9. Micro average metrics, in a single run, over the IWSLT 2012-03 dataset.

Model	Precision	Recall	F_1-score
CRF	0.738	0.525	0.614
BI-LSTM+Skip$_s$300	0.792	0.667	0.724
BERT-BASE	**0.833**	**0.789**	**0.810**

On the other hand, the BI-LSTM + CRF outperformed BERT in precision for PERIOD, while achieving similar precision to BERT in COMMA. Thus, the BI-LSTM + CRF may be a viable alternative when question marks are not part of the language. In any case, the results suggest that the BI-LSTM + CRF struggles to handle unbalanced datasets.

It is worth mentioning that the baseline, the CRF algorithm, was surpassed by both models when comparing the results for each class. However, it achieved similar results to both BI-LSTM and BERT for COMMA. In opposition, it achieved poor results to QUESTION. Since it uses hand-crafted features, we conjecture that CRF depends on cased information to predict the punctuation, as suggested by [17], and deals poorly with unbalanced datasets.

In summary, the BERT model achieved an overall better performance than the BI-LSTM + CRF model and CRF algorithm, regarding all considered classes. We believe that BERT's bidirectional encoder procedure is able to capture meaningful information better than other contextual embedding strategies.

5.2 Out-of-domain Evaluation

Cross-domain evaluation is important when one wants to guarantee the method can be applied/tested in different domains, especially when data can hardly be available for some domains [15]. Hence, testing models in a different domain they were once trained is extremely valuable.

The results presented in Section 5.1 showed that BERT, pre-trained in the IWSLT 2012-03 dataset, achieved better overall results when compared to the other models. Thus, we evaluated the BERT model in the out-of-domain OBRAS dataset, and the results are shown in Table 10.

The results clearly show that the BERT model reached a good precision in all classes and on the micro averages. However, the recall metric dropped significantly for QUESTION labels, which could impact Automatic Speech Recognition

Table 10. Evaluation results (precision, recall, and F_1-score) of the BERT-BASE model in the OBRAS dataset.

Label	P	R	F_1
COMMA	0.697	0.608	0.649
PERIOD	0.877	0.865	0.871
QUESTION	0.626	0.427	0.508
Micro Averages	0.771	0.703	0.735

(ASR) by missing punctuation marks that should be predicted. In any case, the overall results indicate that the model's predictions are mostly correct.

5.3 Cost Analysis

Experimental Setup: The BERT and BI-LSTM algorithms were trained in a Tesla P100-PCIE 16GB memory. For the BERT algorithm, the ADAM with Weight Decay optimiser was used, while for the BI-LSTM algorithm, the SGD with Weight Decay was used.

We conducted computational cost analysis in terms of GPU resources to understand how much resources must be allocated to train a BERT model and a BI-LSTM + CRF model in the IWSLT 2012-03 dataset. We used the Weights & Biases (W&B) framework [3] to collect the data. Figure 2 shows the GPU memory access, GPU utilization, and GPU memory allocation requests.

We found that the BI-LSTM + CRF algorithm had lower GPU time and memory consumption than the BERT algorithm, being easier for training when GPU resources are scarce. Furthermore, the pre-trained BERT model has 110 million parameters, surpassing the BI-LSTM + CRF model, requiring more resources during evaluation.

Thus, although the BERT model achieves better results in punctuation restoration for Brazilian Portuguese, it has high resource requirements for training and prediction. It might make it difficult for deployment in small devices that need to perform speech-to-text in real-time without relying on an internet connection. For such scenarios, the BI-LSTM + CRF model appears to be a better option, which balances both processing performance and prediction precision.

6 Conclusion

Punctuation Restoration is a relevant topic that has been widely investigated in the last few years since punctuation information is important for many NLP tasks and facilitates human understanding. However, there is still a gap in the literature investigating punctuation restoration in languages other than English, for instance, for Brazilian Portuguese language. Thus, to help mitigate this gap, in this work, we explored state-of-art ML algorithms applied for punctuation restoration of Brazilian Portuguese texts.

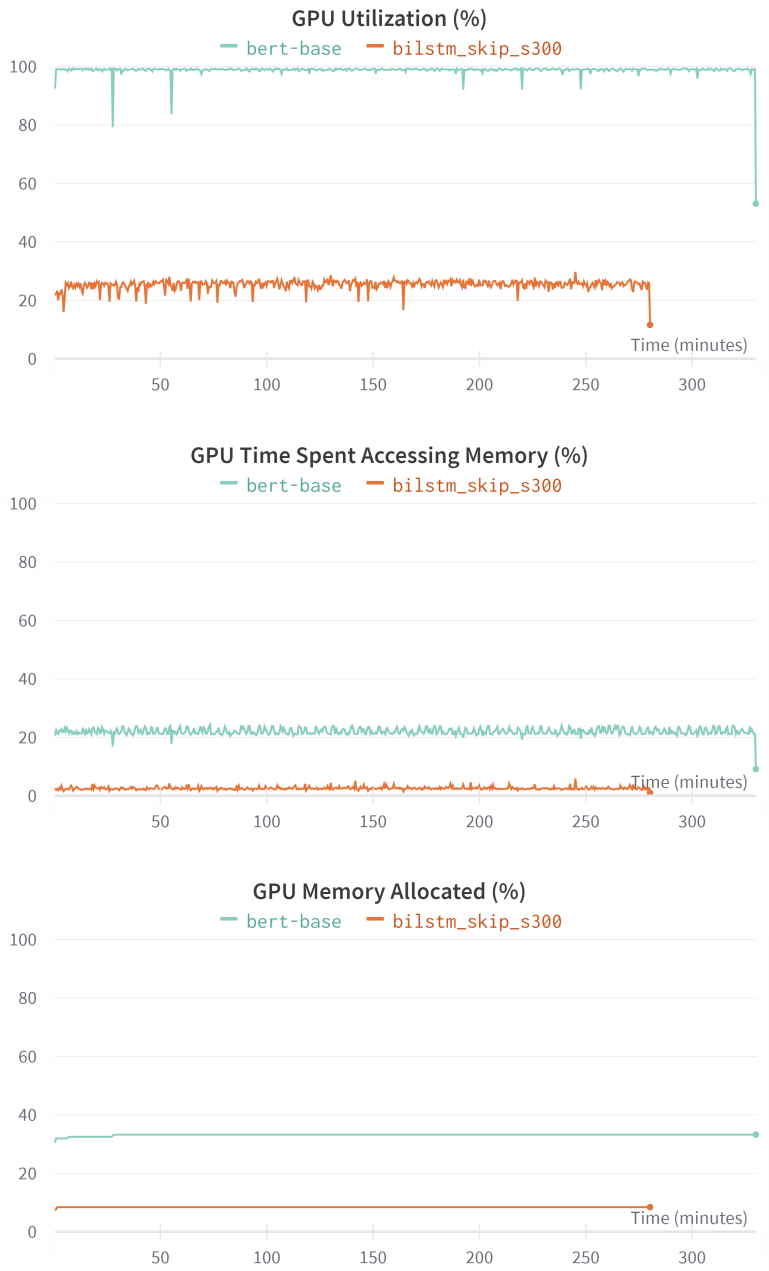

Fig. 2. Consumption of GPU resources for training both BI-LSTM +CRF and BERT.

As with most of the recent literature, we also treat the punctuation restoration problem as a sequence labeling task. We performed an experimental analysis of the BERT and BI-LSTM + CRF algorithms in two Brazilian Portuguese datasets: the IWSLT 2012-03 (Brazilian translated version) dataset and the OBRAS corpus.

The results showed that BERT is a promising approach, surpassing the BI-LSTM + CRF algorithm in most cases of the IWSLT 2012-03 dataset. It also achieved good performance in the out-of-domain OBRAS dataset, suggesting robustness for cross-domain applications. However, through a computation cost analysis, we found that the BERT algorithm requires much more GPU resources for training than the BI-LSTM + CRF algorithm.

In any case, we hope our findings contribute to pursuing the state-of-the-art and enriching the literature for punctuation restoration in Brazilian Portuguese. In future works, we intend to investigate multilingual models to address the problem of punctuation restoration in different languages and consider more datasets and models.

References

1. Akbik, A., Bergmann, T., Blythe, D., Rasul, K., Schweter, S., Vollgraf, R.: FLAIR: An easy-to-use framework for state-of-the-art NLP. In: NAACL 2019, 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations). pp. 54–59 (2019)
2. Alam, T., Khan, A., Alam, F.: Punctuation restoration using transformer models for high-and low-resource languages. In: Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020). pp. 132–142 (2020)
3. Biewald, L.: Experiment tracking with weights and biases (2020)
4. Botarleanu, R.M., Dascalu, M., Crossley, S.A., McNamara, D.S.: Sequence-to-sequence models for automated text simplification. In: International Conference on Artificial Intelligence in Education. pp. 31–36. Springer (2020)
5. Chandra, Y.W., Suyanto, S.: Indonesian chatbot of university admission using a question answering system based on sequence-to-sequence model. Procedia Computer Science **157**, 367–374 (2019)
6. Chiticariu, L., Krishnamurthy, R., Li, Y., Reiss, F., Vaithyanathan, S.: Domain adaptation of rule-based annotators for named-entity recognition tasks. In: Proceedings of the 2010 conference on empirical methods in natural language processing. pp. 1002–1012 (2010)
7. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
8. Guo, J., He, H., He, T., Lausen, L., Li, M., Lin, H., Shi, X., Wang, C., Xie, J., Zha, S., et al.: Gluoncv and gluonnlp: Deep learning in computer vision and natural language processing. J. Mach. Learn. Res. **21**(23), 1–7 (2020)
9. Hartmann, N., Fonseca, E., Shulby, C., Treviso, M., Rodrigues, J., Aluisio, S.: Portuguese word embeddings: Evaluating on word analogies and natural language tasks. arXiv preprint arXiv:1708.06025 (2017)
10. Huang, Z., Xu, W., Yu, K.: Bidirectional LSTM-CRF models for sequence tagging. CoRR **abs/1508.01991** (2015)

11. Klejch, O., Bell, P., Renals, S.: Punctuated transcription of multi-genre broadcasts using acoustic and lexical approaches. In: 2016 IEEE Spoken Language Technology Workshop (SLT). pp. 433–440 (2016)
12. Kolár, J., Lamel, L.: Development and evaluation of automatic punctuation for french and english speech-to-text. In: Interspeech. pp. 1376–1379 (2012)
13. Li, W., Gao, S., Zhou, H., Huang, Z., Zhang, K., Li, W.: The automatic text classification method based on bert and feature union. In: 2019 IEEE 25th International Conference on Parallel and Distributed Systems (ICPADS). pp. 774–777. IEEE (2019)
14. Li, X., Lin, E.: A 43 language multilingual punctuation prediction neural network model. vol. 2020-October, pp. 1067–1071. International Speech Communication Association (2020)
15. Liu, Z., Xu, Y., Yu, T., Dai, W., Ji, Z., Cahyawijaya, S., Madotto, A., Fung, P.: Crossner: Evaluating cross-domain named entity recognition. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 35, pp. 13452–13460 (2021)
16. Lu, W., Ng, H.T.: Better punctuation prediction with dynamic conditional random fields. In: Proceedings of the 2010 conference on empirical methods in natural language processing. pp. 177–186 (2010)
17. Lui, M., Wang, L.: Recovering casing and punctuation using conditional random fields. In: Proceedings of the Australasian Language Technology Association Workshop 2013 (ALTA 2013). pp. 137–141 (2013)
18. Makhija, K., Ho, T.N., Chng, E.S.: Transfer learning for punctuation prediction. In: 2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC). pp. 268–273 (2019)
19. Nagy, A., Bial, B., Ács, J.: Automatic punctuation restoration with bert models (1 2021)
20. Pennington, J., Socher, R., Manning, C.D.: Glove: Global vectors for word representation. In: Empirical Methods in Natural Language Processing (EMNLP). pp. 1532–1543 (2014)
21. Pires, T., Schlinger, E., Garrette, D.: How multilingual is multilingual bert? arXiv preprint arXiv:1906.01502 (2019)
22. Păiş, V., Tufiş, D.: Capitalization and punctuation restoration: A survey. Artif. Intell. Rev. **55**(3), 1681–1722 (mar 2022)
23. Souza, F., Nogueira, R., Lotufo, R.: Portuguese named entity recognition using bert-crf. arXiv preprint arXiv:1909.10649 (2019)
24. Souza, F., Nogueira, R., Lotufo, R.: BERTimbau: pretrained BERT models for Brazilian Portuguese. In: 9th Brazilian Conference on Intelligent Systems, BRACIS, Rio Grande do Sul, Brazil, October 20-23 (to appear) (2020)
25. Sutskever, I., Vinyals, O., Le, Q.V.: Sequence to sequence learning with neural networks. Advances in neural information processing systems **27** (2014)
26. Tilk, O., Alumäe, T.: Bidirectional recurrent neural network with attention mechanism for punctuation restoration. In: Interspeech. pp. 3047–3051 (2016)
27. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
28. Wang, F., Chen, W., Yang, Z., Xu, B.: Self-attention based network for punctuation restoration. In: 2018 24th International Conference on Pattern Recognition (ICPR). pp. 2803–2808 (2018)
29. Yi, J., Tao, J., Bai, Y., Tian, Z., Fan, C.: Adversarial transfer learning for punctuation restoration. arXiv preprint arXiv:2004.00248 (2020)

30. Zheng, Z., Zhou, H., Huang, S., Chen, J., Xu, J., Li, L.: Duplex sequence-to-sequence learning for reversible machine translation. In: Advances in Neural Information Processing Systems. vol. 34, pp. 21070–21084. Curran Associates, Inc. (2021)

Automatic Punctuation Verification of School Students' Essay in Portuguese

Tiago Barbosa de Lima[1], **Luiz Rodrigues**[3], **Valmir Macario** [1], **Elyda Freitas** [2,4], **Rafael Ferreira Mello** [1,2]

[1]Departamento de Computação – Universidade Federal Rural de Pernambuco (UFRPE)
Rua Dom Manuel de Medeiros, s/n, Dois Irmãos - CEP: 52171-900 - Recife – PE – Brazil

[2]Centro de Estudos e Sistemas Avançados do Recife (CESAR)
Rua Cais do Apolo, 220, Recife, - CEP: 50030-390 Recife – PE – Brazil

[3]NEES - UFAL – Av. Lourival Melo Mota, S/N - Cidade Universitária,
Maceió - AL, 57072-970.

[4]Departamento de Sistemas de Informação
Universidade de Pernambuco (UPE) – Caruaru, PE – Brazil

{tiago.blima,valmir.macario,rafael.mello}@ufrpe.br

luiz.rodrigues@nees.ufal.br, elyda.freitas@upe.br

Abstract. *Textual production is a key activity at different levels of education. The analysis of essays encompasses several criteria, such as lexical and syntactic errors, cohesion, and coherence. Within these criteria, how the students include punctuation (i.e., final mark and comma) could influence the quality of the final production. Thus, the literature has proposed several approaches to verifying punctuation correction in students' essays for English. However, despite the advancements in natural language processing models for other languages, there is a significant gap concerning punctuation verification. Therefore, this paper proposed a new approach based on state-of-the-art language models to develop a punctuation prediction method for Portuguese. The proposed model was applied to evaluate the textual productions of students in Brazilian public schools. Finally, the results of this study and its practical implications for educational settings are further discussed.*

1. Introduction

Punctuation is a relevant aspect of learning a new language [Suliman et al. 2019]. The incorrect use of punctuation might lead to a diverse range of miss interpretations [Suliman et al. 2019]. Despite its importance, the literature shows that correct punctuation is a significant problem for pupils and second language learners [Awad 2012]. Therefore, several studies have proposed algorithms and models to analyze the main errors made by pupils and second language learners, aiming to build automatic assistance software to help them [Kurup et al. 2016a]. In this context, Natural Language Processing (NLP) has significantly developed punctuation verification models [Sahami et al. 2011].

Moreover, many grammatical checkers that provide feedback about punctuation errors have been released recently (e.g., Grammarly). However, the problem of automatic punctuation checking is not widely addressed in non-English languages, such as

Portuguese. Despite there being some tools, such as cogroo[1] and language tool[2], none provide enough resources for punctuation verification. Cogroo and Language-tool are able to check minor errors such as sentences punctuated at the beginning or repetitive use of periods or commas in sequence. Therefore, other punctuation misuses as commas separating the subject and verb or the lack of sentences to separate the appositive still lack automatic correction tools. Therefore, exploring punctuation correction for Brazilian Portuguese text is still an open problem in the literature.

The primary approach to analyzing students' essays to detect punctuation errors is to create models that predict the correct punctuation and then compare the outcomes with the students' texts [Suliman et al. 2019, Nagy et al. 2021]. More specifically, the task consists of predicting after which words the punctuation is necessary [Vāravs and Salimbajevs 2018]. Previous papers proposed approaches for punctuation prediction in English [Nagy et al. 2021] and Portuguese [Lima et al. 2022], where the authors evaluated different algorithms (e.g., LSTM, BERT, and Conditional Random Fields (CRF)). Particularly, BERT reached the best results in both cases. Although the promising results of the related paper, to the best of our knowledge, no previous work has applied these models specifically for educational settings.

Furthermore, recently the T5 model [Raffel et al. 2020] has reached more robust results in several NLP tasks, compared to BERT, which achieved state-of-the-art in related work. However, to our best knowledge, T5 has not been used for punctuation prediction yet. Therefore, this paper assesses the performance of BERT and T5 models for punctuation prediction in Brazilian Portuguese within educational texts for elementary school students[Gazzola et al. 2019]. The original dataset was conceived to address automatic readability classification but we adapted the text to punctuation restoration text. We also evaluated the models in a new dataset created containing essays from students in Brazilian public schools and the results reveal that BERT reached better results and generalization for this task. There are promising results from both models, we trained and evaluated the models as punctuation restoration tasks and then we used the model to predict the correct punctuation of the student's essays. The models achieve competitive performance in well-structured sentences, despite a poor outcome in incorrectly written sentences. Finally, the practical implications for education are further discussed such as the main causes for poor performance in non well-structed sentences and possible improvements.

2. Background

This section presents background information on language models and the bottleneck of punctuation analysis in NLP.

2.1. Language Models

Language Models (LM) learn the semantic structure of a specific language from unlabeled text corpora, allowing the automatic creation of relationships between words and sentences. These models might boost several NLP tasks, such as Named Entity Recognition, Text Classification, and Question-Answering (QA) systems [Devlin et al. 2019]. LM gained significant prominence with the development of the BERT model and, later, with T5 [Raffel et al. 2020].

[1]http://www.cogroo.org/

[2]https://languagetool.org/pt-BR

BERT is a language model capable of performing different tasks when fine-tuned. It was first released by [Devlin et al. 2019] in two versions: base and large. Similarly to the English version, [Souza et al. 2020] trained a base from the multi-language checkpoint version of BERT (mBERT) with 110M parameters and large BERT version from the original English version of BERT large with 340M parameters respectively [Devlin et al. 2019, Souza et al. 2020, Devlin 2018].

The literature shows that the pre-trained BERT is suitable for different tasks because they learn deep textual representation [Devlin et al. 2019]. Thus, the concept enables the creation of a diverse range of applications through fine-tuning and minimal architectural changes.

T5 is a text-to-text LM with multi-task capability. It aims to predict a sequence of text [Raffel et al. 2020], differently from BERT, which predicts a single word in a given context. This characteristic allows T5 to perform multiple tasks through text generation, such as text summarization, QA, and translation [Raffel et al. 2020]. Additionally, since the T5 model supports multiple tasks in the same model, one must add a specific tag for each task.

Overall, those robust LMs can carry textual representation to different levels, allowing their application in a diverse range of textual contexts, such as education, health, and justice [Kundu 2021, Perrotta and Selwyn 2020].

2.2. Punctuation analysis

Investigating the use of punctuation is essential to build up strategies for improving students' written communication. There is a significant effort to evaluate grammatical correction systems and punctuation verification systems in the educational context [Kinoshita et al. 2006, Adriaens 1994]. It pushed the development of tutoring systems, data analyses, and other methods to evaluate students' performance. For instance, the Cogroo project can recognize simple punctuation errors [Kinoshita et al. 2006]. On the other hand, there are meaningful advancements on this topic in English with the evaluation of the second language learners and college students [Awad 2012].

In this context, [He 2009] evaluated the performance of a tutoring system for automatic punctuation. The purpose was to provide flag feedback anytime a student does not include mandatory punctuation and suggest improvements on this concern by giving step-by-step instructions to fix multiple errors. The study assessed the performance of the proposed method with ten students, who showed significant improvements in the post-test after using the software considering eight English punctuation rules [He 2009]. Furthermore, the system provides insights into how automatic tutoring software can help students across punctuation challenges [He 2009].

Another work [Nagata and Nakatani 2010] goes even further in analyzing the learning effect of automatic educational software in English. The paper evaluates the impact of precision-oriented and recall-oriented software on learning by comparing the results with a real human tutor. The study analyzed 22 different valid essays of 10 sentences or more made by Japanese college students [Nagata and Nakatani 2010]. The first group wrote without any intervention, the second with human tutoring (4 students), and the third and fourth with precision-orient (6 students) and recall-oriented tutoring system (7 students). The researchers evaluated a grammar corrector system that is closer to a hu-

man tutor when based on precision feedback, where critical errors are detected, but other errors the students must find alone.

In addition to the presented papers, Grammarly is an online tool - also available for mobile, Windows, and MAC - for text corrections powered by an AI engine[3]. The main objective of the Grammarly app is to provide online feedback while the user is typing the tasks, not only for grammatical mistakes but also for punctuation and other tools such as plagiarism. Several studies propose to measure not only the real improvement of punctuation promoted by Grammarly but also the overall student view of the platform through survey questions [Im 2021, ONeill and Russell 2019, Cavaleri and Dianati 2016].

Punctuation plays a crucial role in enhancing text comprehension, necessitating the awareness of NLP models towards punctuation in text prediction tasks [Tilk and Alumäe 2016]. However, certain tasks like Automatic Speech Recognition (ASR) and some models do not predict punctuation correctly in the text [Tilk and Alumäe 2016]. Therefore, the punctuation restoration task aims to utilize machine learning techniques, such as sequence labeling, to automatically predict the missing punctuation [Klejch et al. 2016, Makhija et al. 2019, Nagy et al. 2021].

In general, deep learning models provided the most significant results in the last years when combining pre-training embeddings or using pre-trained models such as BERT for punctuation restoration. For example, the strategy proposed by [Nagy et al. 2021] consists of treating the punctuation restoration problem as a sequence labeling task in which each token receives one of the labels according to the Inside–outside–beginning (BIOS) tagging annotation [Ramshaw and Marcus 1995] where O (no-punctuation) and labels I-COMMA (,), I-PERIOD (.), or I-QUESTION (?), which precedes words with the punctuation. The best performing algorithm of this work obtained an 80.6 F1 score for all labels with the BERT-base developed by [Courtland et al. 2020].

Other works [Nagy et al. 2021, Lima et al. 2022, Tilk and Alumäe 2016, Makhija et al. 2019] used the IWSLT 2012-03 dataset to address punctuation restoration tasks both in English and Portuguese. The IWSLT 2012-03 proposed by [Federico et al. 2012] consists of tedtalks transcriptions in different languages, including Portuguese and English, originally proposed by [Federico et al. 2012] to address Spoken Language Translation (SLT), Speech Recognition and Machine Translation (MT). In turn, the work [Hentschel et al. 2021] adopted another strategy to not only make the punctuation restoration faster but also multitask. They used the ELECTRA model to inject errors in the transcription of the ASR model to make the model more robust. The authors obtained a significant improvement of 11% using a model smaller than BERT. Therefore, punctuation restoration is a widely used strategy to recover punctuation from ASR output, showing significant results in the literature not only in English but also in Portuguese.

Those works provide meaningful insights into the main problems and possible solutions in future works. Since the use of punctuation is a critical evaluation factor for pupils and second language learns, different works evaluate students' punctuation automatically or manually in English [Kurup et al. 2016b, ONeill and Russell 2019,

[3]https://app.grammarly.com/

Im 2021]. However, as far as went our research none of them evaluate the use of punctuation by students in Brazilian Portuguese automatically or manually. Besides, there are only limited tools to address punctuation verification in Brazilian Portuguese text [Kinoshita et al. 2006]. Moreover, the state-of-the-art LM for punctuation prediction tasks (BERT and T5) has not been applied to educational settings. As such, this study proposes the following research questions:

RESEARCH QUESTION 1 (RQ1):
To what extent can BERT and T5 predict the correct punctuation for Portuguese texts?

RESEARCH QUESTION 2 (RQ2):
To what extent can BERT and T5 accurately estimate punctuation errors in students' textual productions?

3. Method

This section presents the datasets, as well as procedures for model selection, assessment, and development adopted in this study.

3.1. Data Description

This study adopted two datasets to train the LMs and evaluate students' punctuation performances when writing essays. The first dataset, named NILC dataset, encompasses a series of school books from different educational levels [Gazzola et al. 2019]. The primary objective of the corpus was to evaluate text complexity. The original NILC dataset includes textbooks focused on elementary, middle, high school, and under-graduated levels. Overall, the dataset consists of 1695 texts and 13016 total sentences. Table 1 describes the number of instances used for training, validation, and testing procedure. The dataset was split using a stratified strategy to maintain the same proportion of both educational levels at training and test.

It is important to highlight that we considered all exclamation marks, semi-colons, and question marks to be periods, similar to both previous works [Nagy et al. 2021, Lima et al. 2022]. Moreover, to the best of our knowledge, we were the first to use this dataset to address punctuation restoration.

Table 1. The final number of texts, sentences, and labels after pre-processing of NILC and MEC datasets.

split	Number of Texts	Number of Sentences	Sentences Elementary I	Sentences Elementary II	I-PERIOD	I -COMMA
train	613	9371	4898	4473	11961	9424
test	597	2604	1361	1243	2621	3335
validation	485	1041	544	497	1424	1044
Total	1695	13016	6803	6213	16006	13803
MEC	256	2004	-	-	2004	1082

The second dataset presented in this paper, called MEC dataset, comprises 265 essays (2004 sentences) by students in middle-school public schools in Brazil. Two expert coders annotated the dataset using three categories: insertion (the student included the

punctuation in the wrong place), missing (the student did not include the required punctuation), and exchange (the student included the wrong punctuation). The coders reached an average agreement of 0.569, according to Cohen's Kappa, which represents a moderate agreement [Landis and Koch 1977]. The dataset encompasses 2004 and 1082 instances of period and comma errors, respectively. As this is a small dataset, it was used only for testing purposes, not for training or validation.

3.2. Model Selection

As detailed in section 2.1, the punctuation restoration is a sequence labeling task. Thus, the language models can infer the results without an additional classification algorithm. In this context, we assessed the performance of BERT and T5 for the problem of punctuation prediction and verification. For the BERT model, we used the Portuguese version released by [Souza et al. 2020] in two different architectures: base (with 110M parameters) and large (with 330M of parameters). It is important to mention that BERT is an encoder-only model that predicts words[Devlin et al. 2019]. Thus, we use it to predict the punctuation directly.

On the other side, the T5 model comprises both encoding and decoding strategies. It means that the T5 architecture allows the use of the same model for different tasks by changing the input tag of input texts [Raffel et al. 2020]. Precisely for this study, we predict the entire sentence, with the corresponding punctuation, aiming to evaluate its correctness. The Portuguese T5 was first released by [Carmo et al. 2020] with four pre-trained models. As the authors recommended, we used the most recent models (i.e., ptt5-base-portuguese-vocab and ptt5-large-portuguese-vocab) with 220M and 760M of parameters, respectively [Carmo et al. 2020].

3.3. Model evaluation

To address RQ1, we assessed the selected models with the NILC dataset using the train, validation, and test split described in table 1. We adopted the evaluation process recommended in the literature [Akbik et al. 2018] to compare the results of the sequential-based models (BERT-based and T5-based).

To evaluate BERT, which does single-word prediction, we applied the traditional NLP evaluation measures used by previous works [Nagy et al. 2021, Lima et al. 2022]: precision, recall, and f-score. In short, precision assesses how accurate the model is in predicting a specific category, while recall measures the number of correctly retrieved instances in the dataset. F1-score is the harmonic mean of both measures, which provides a general performance indicator.

For the T5, which does full sentence prediction, the adequate measure to evaluate is the Bilingual Evaluation Understand (BLEU score) [Papineni et al. 2002]. BLEU captures and evaluates the overlap between the predicted and the reference sentences [Garg and Agarwal 2018]. It has been widely used in the Machine Translation domain for years and was adapted to other tasks, such as QA and Text simplification. After the validation step of the T5 model, we also used precision, recall, and f-score to analyze the results in the test set.

To address RQ2, we selected the best model identified in RQ1 to assess their capability to detect errors automatically in the student-written texts, MEC dataset. In this

case, we decided not to fine-tune the models using the MEC dataset due to the limited number of instances available. Therefore, we measure the performance of the models with precision, recall, and f-score.

3.4. Experimental Setup

We used a google cloud T4 Tesla GPU of 16GB architecture to execute the experimentation. For each model, we evaluated five epochs using the hyper-parameter specified in Table 2, as suggested by [Akbik et al. 2018].

Table 2. Model hyper-parameters for BERT and T5 models.

Parameter	BERT	T5
Learning rate	5.00e-5	5.00e-5
Train batch size	8	2
Eval batch size	8	2
Seed	42	42
Optimizer	Adam with betas=(0.9,0.999) epsilon=1e-08	Adam with betas=(0.9,0.999) epsilon=1e-08
LR scheduler type	linear	linear
Number of epochs	5	5

4. Results

This section presents our results for RQ1 and RQ2.

4.1. RQ1: to what extent can BERT and T5 predict the correct punctuation for Portuguese texts?

The first research question aimed to compare the results of BERT and T5 algorithms with the NILC dataset. Initially, we focused on the analysis of the training and validation process. Figures 1 and 2 present the results of the execution from epochs 1 to 5 in the validation dataset of BERT and T5, respectively. Overall, the best results were reached with four epochs for the base models and five for the large ones. Thus, these were the models selected for the rest of the experimentation.

As can be observed in Figure 1, the detection of the score, comparing the variations of the BERT model, portrays the convergence of the predictive capacity of the model over time. The BERT Base model shows more smoothness in detecting the score and stabilizing itself in constant accuracy with the course of training and validation. The BERT Large model, in contrast, by better capturing phrase-level information in the lower and hierarchical information in the intermediate layers of the language [Jawahar et al. 2019], reaches the highest levels of score prediction.

Unlike the previous scenario, Figure 2 shows the evolution in terms of the BLEU measure of the T5 model. This measure was evaluated to observe the agreement of the model's predicted output with the expected one. It is essential to highlight that the T5 Base model presented the best BLEU measurement in epoch 3. Another essential characteristic is that both T5 Large and Base models presented similar training/validation curves.

4.2. RQ2: to what extent can BERT and T5 accurately estimate punctuation errors in students' textual productions?

Tables 3 and 5 present the comparative results to answer RQ2. That is, under the comparative aspect between the models, to observe Precision, Recall, and F1-Score measures in

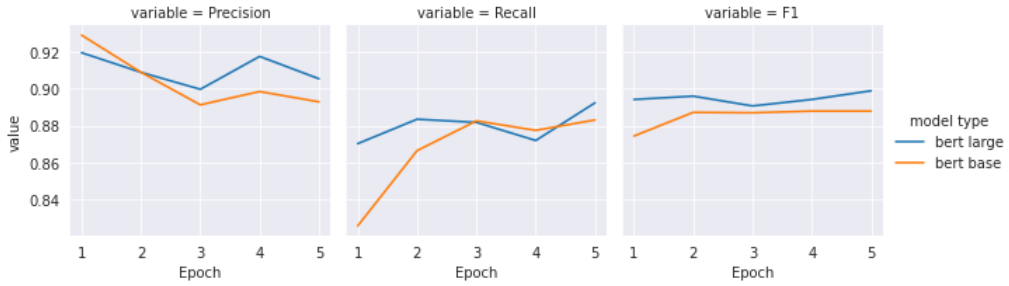

Figure 1. BERT training Evolution with the validation set.

Figure 2. T5 model training performance on the validation set.

both models BERT and T5, considering their Base and Large variants. Hence, the tables enable observing the predictive capacity of the models concerning the evaluation of the punctuation of the texts in the Portuguese Language.

Considering a more controlled dataset, with texts produced and extracted from educational books, Table 3 presents the results obtained by the models and their respective variations. The predictive capacity, in general, was between 0.74 and 0.84 in all measures. Evaluating only *comma* and considering the entire period, the accuracy value rises to 0.98 and 0.99. In terms of the mean, the values obtained were between 0.85 and 0.91. Thus, the best accuracy for evaluating the score, in terms of average, is in the T5 model and its variants BASE and Large, with measures of 0.90 and 0.91, respectively.

Table 3. Table shows the result for all models and measures evaluated at the NILC test dataset with measures Precision (P), Recall (R) and F1-score (F).

	BERT BASE			BERT LARGE		
	P	**R**	**F**	**P**	**R**	**F**
COMMA	0.802	0.772	0.787	0.81	0.784	0.797
PERIOD	**0.997**	0.993	**0.995**	0.996	0.993	0.994
AVG	0.891	0.873	0.882	0.895	0.88	0.887
	T5 BASE			**T5 LARGE**		
	P	**R**	**F**	**P**	**R**	**F**
COMMA	0.831	0.747	0.787	0.842	0.762	0.8
PERIOD	0.995	**0.989**	0.992	**0.998**	**0.994**	**0.996**
AVG	0.906	0.858	0.88	0.914	0.868	0.89

Table 4. Comparison with previous works related to punctuation restoration with measures Precision (P), Recall (R) and F1-score (F).

Paper	Model	Language	P	R	F
[Makhija et al. 2019]	BERT-Punct LARGE	English	79.5	83.7	81.4
[Courtland et al. 2020]	Roberta-base	English	84	**83.9**	**83.9**
[Nagy et al. 2021]	BERT base uncased	English	75.8	85.1	79.8
[Lima et al. 2022]	BERT base cased	Portuguese	83.3	78.9	81

The table shows the previous results of punctuation restoration works on the widely used IWLST2012 public dataset. Differently from our work, the previous paper considered commas, periods and question marks instead of treating question marks as periods.

Table 5 presents the comparative results with the models considering the MEC Dataset. Unlike the previous scenario, which considered the NILC Dataset, the results obtained in this comparison showed a better performance in evaluating the punctuation by the BERT model.

Table 5. Table shows the result for all models and measures evaluated at the MEC dataset with measures Precision (P), Recall (R) and F1-score (F).

	BERT BASE			BERT LARGE		
	P	**R**	**F**	**P**	**R**	**F**
COMMA	0.12	0.368	0.181	0.123	0.381	0.186
PERIOD	**0.984**	**0.999**	0.991	**0.97**	0.996	**0.983**
AVG	0.707	0.797	**0.732**	0.698	0.799	0.727
	T5 BASE			**T5 LARGE**		
	P	**R**	**F**	**P**	**R**	**F**
COMMA	0.049	0.126	0.07	0.047	0.139	0.07
PERIOD	0.8	0.009	0.018	0.697	0.011	0.021
AVG	0.603	0.04	0.032	0.527	0.044	0.034

Finally, Table 6 presents descriptive statistics of the proportion of errors that were returned. Those consider Different Numbers of Labels (Test case 1), partial evaluation, which corresponds to an Equal Number of Labels but Wrong Placement (Test case 2), and, Full match (Test case 3). That is an approximate assessment of where the probable punctuation error might be. At this stage, some linguistic mechanisms, such as ambiguities, were concentrated, which could result in two possible ways to evaluate the score in the dataset.

Table 6. Number of examples in each case evaluated.

Test Case	Number of Punctuation	Proportion
1	237	54.11%
2	15	3.42%
3	186	42.47%
Total	438	100%

5. Discussion

Punctuation plays a vital role in enhancing the clarity and readability of communication. By providing precise markers, it facilitates effective communication. The results obtained

from our evaluation indicate strong promise when utilizing more recent Natural Language Processing algorithms. Our best result, achieved using the T5-Large algorithm, achieves an impressive average F1-score of 0.89, surpassing the performance of previous work by [Courtland et al. 2020]. Several factors contribute to this positive outcome. Firstly, in our analysis, we considered labels for both periods and commas since the students' datasets treated question marks as periods without analyzing them separately.

As a result, the total number of periods in the training set increases. However, we intentionally refrain from using an excessively large dataset. Doing so may cause the algorithm to overly generalize within a specific context, which differs from the TEDTALK IWLST2012 dataset described in the papers referenced in Table 4. Furthermore, the dataset consists of texts specifically tailored for children, which contributes to achieving a higher level of accuracy. However, it is worth noting that these texts are comparatively simpler compared to more complex and mature content intended for a different audience.

We also address RQ2, which explores the extent to which BERT and T5 models can accurately detect punctuation errors in students' written work. While both models demonstrate above-average performance, particularly BERT, they encounter difficulties likely stemming from the dataset size. The limited number of samples may not provide enough information for the models to effectively evaluate aspects such as pauses, rhythm, and intonation within the text. These aspects are crucial for various text genres like narratives and dissertations. Furthermore, students' essays often contain significant grammatical errors that can introduce punctuation inconsistencies. Language models like BERT, which capture intricate linguistic features at the phrasal level, may face challenges in correctly labeling punctuation, especially when confronted with grammatical mistakes in a sentence [Jawahar et al. 2019]. This can significantly impact the accurate labeling of punctuation, particularly for commas.

Finally, the discussed models demonstrate the ability to verify writing style and provide corrective feedback, showcasing minimum threshold scores that should be present in students' texts. However, a notable limitation of these models is their assessment of commas in student sentences, despite performing well on well-structured sentences. Given the multitude of grammatical errors and text inconsistencies that can lead to erroneous predictions, it would be beneficial to evaluate the maximum number of grammatical errors before assessing punctuation. This approach would help mitigate the problem effectively. Additionally, delving into how the model arrived at a specific label can enhance robustness. Therefore, incorporating explainable AI (XAI) in future research could further improve the results. By building more robust models, we can assist middle-school teachers in essay assessments while boosting students' confidence and enhancing their writing skills [Wilson and Roscoe 2020, He 2009]. Moreover, as one of the pioneering studies addressing automatic punctuation in students' essays, this work opens the door for further research in this area by identifying key limitations and suggesting new directions for investigation. The results highlight that punctuation serves as a valuable tool for evaluating textual continuity, representing intonation and conveying emotions in narrative texts. It underscores the importance of punctuation in writing, enabling more precise, accurate, and effective communication.

6. Final Remarks

Punctuation verification has been addressed in different formats over the year. However, the topic is not fully discussed in Brazilian Portuguese. Thus, this paper presents a benchmark evaluation of BERT and T5 language models to address the punctuation restoration task in Brazilian Portuguese text for children. Also, as far as went our research, no paper to previous data has yet addressed the punctuation verification of students' essays before, then we present a novel dataset for punctuation verification of Brazilian students that can help research in the field in the close future.

The results show that models can be applied with success in well-structured sentences, however, improvements are necessary for unstructured texts. Moreover, punctuation verification with ML has promising results, for future works comparison with ruled-based approaches and LLM prompting engineering would be of good importance.

The results present the evaluation from the local perspective of error correction and its overall relationship shows a strong deficiency in predicting punctuation in a not well-structured text. However, some mechanisms, such as some datasets to emphasize important words and phrases and their due grammatical classes, could be used to enrich the datasets further and, consequently, make the models reach higher levels of score evaluation.

References

Adriaens, G. (1994). Simplified English grammar and style correction in an MT framework: The LRE SECC project. In *Proceedings of Translating and the Computer 16*, London, UK. Aslib.

Akbik, A., Blythe, D., and Vollgraf, R. (2018). Contextual string embeddings for sequence labeling. In *Proceedings of the 27th international conference on computational linguistics*, pages 1638–1649.

Awad, A. (2012). The most common punctuation errors made by the english and the tefl majors at an-najah national university. . *Vol.*, 26:23.

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

Cavaleri, M. R. and Dianati, S. (2016). You want me to check your grammar again? the usefulness of an online grammar checker as perceived by students. *Journal of Academic Language and Learning*, 10(1):A223–A236.

Courtland, M., Faulkner, A., and McElvain, G. (2020). Efficient automatic punctuation restoration using bidirectional transformers with robust inference. In *Proceedings of the 17th International Conference on Spoken Language Translation*, pages 272–279.

Devlin, J. (2018). Multilingual bert readme document.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

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

Garg, A. and Agarwal, M. (2018). Machine translation: a literature review. *arXiv preprint arXiv:1901.01122*.

Gazzola, M. G., Leal, S. E., and Aluísio, S. M. (2019). Predição da complexidade textual de recursos educacionais abertos em português. In *Symposium in Information and Human Language Technology - STIL*. SBC.

He, X. (2009). A web-based intelligent tutoring system for english dictation. In *2009 International Conference on Artificial Intelligence and Computational Intelligence*, volume 4, pages 583–586.

Hentschel, M., Tsunoo, E., and Okuda, T. (2021). Making Punctuation Restoration Robust and Fast with Multi-Task Learning and Knowledge Distillation. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7773–7777. ISSN: 2379-190X.

Im, H.-J. (2021). The use of an online grammar checker in english writing learning. *Journal of Digital Convergence*, 19(1):51–58.

Jawahar, G., Sagot, B., and Seddah, D. (2019). What does BERT learn about the structure of language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657, Florence, Italy. Association for Computational Linguistics.

Kinoshita, J., Salvador, L. d. N., and de Menezes, C. E. D. (2006). CoGrOO: a Brazilian-Portuguese grammar checker based on the CETENFOLHA corpus. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy. European Language Resources Association (ELRA).

Klejch, O., Bell, P., and Renals, S. (2016). Punctuated transcription of multi-genre broadcasts using acoustic and lexical approaches. In *2016 IEEE Spoken Language Technology Workshop (SLT)*, pages 433–440. IEEE.

Kundu, S. (2021). AI in medicine must be explainable. *Nature Medicine*, 27(8):1328–1328.

Kurup, L., Joshi, A., and Shekhokar, N. (2016a). Intelligent Tutoring System for learning English punctuation. In *2016 International Conference on Computing Communication Control and automation (ICCUBEA)*, pages 1–6, Pune, India. IEEE.

Kurup, L., Joshi, A., and Shekhokar, N. (2016b). Intelligent tutoring system for learning english punctuation. In *2016 International Conference on Computing Communication Control and automation (ICCUBEA)*, pages 1–6. IEEE.

Landis, J. R. and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

Lima, T. B. D., Miranda, P., Mello, R. F., Wenceslau, M., Bittencourt, I. I., Cordeiro, T. D., and José, J. (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, pages 616–630. Springer.

Makhija, K., Ho, T.-N., and Chng, E.-S. (2019). Transfer learning for punctuation prediction. In *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 268–273. IEEE.

Nagata, R. and Nakatani, K. (2010). Evaluating performance of grammatical error detection to maximize learning effect. In *Coling 2010: Posters*, pages 894–900, Beijing, China. Coling 2010 Organizing Committee.

Nagy, A., Bial, B., and Ács, J. (2021). Automatic punctuation restoration with BERT models.

ONeill, R. and Russell, A. (2019). Stop! grammar time: University students' perceptions of the automated feedback program grammarly. *Australasian Journal of Educational Technology*, 35(1).

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Perrotta, C. and Selwyn, N. (2020). Deep learning goes to school: Toward a relational understanding of ai in education. *Learning, Media and Technology*, 45(3):251–269.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P. J., et al. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140):1–67.

Ramshaw, L. and Marcus, M. (1995). Text chunking using transformation-based learning. In *Third Workshop on Very Large Corpora*.

Sahami, M., desJardins, M., Dodds, Z., and Neller, T. (2011). Educational advances in artificial intelligence. In *Proceedings of the 42nd ACM technical symposium on Computer science education*, SIGCSE '11, pages 81–82, New York, NY, USA. Association for Computing Machinery.

Souza, F., Nogueira, R., and Lotufo, R. (2020). Bertimbau: pretrained bert models for brazilian portuguese. In *Brazilian conference on intelligent systems*, pages 403–417. Springer.

Suliman, F., Ben-Ahmeida, M., and Mahalla, S. (2019). Importance of Punctuation Marks for Writing and Reading Comprehension Skills. *(Faculty of Arts Journal)* - , (13):29–53.

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

Vāravs, A. and Salimbajevs, A. (2018). Restoring Punctuation and Capitalization Using Transformer Models. In Dutoit, T., Martín-Vide, C., and Pironkov, G., editors, *Statistical Language and Speech Processing*, volume 11171, pages 91–102. Springer International Publishing, Cham. Series Title: Lecture Notes in Computer Science.

Wilson, J. and Roscoe, R. D. (2020). Automated writing evaluation and feedback: Multiple metrics of efficacy. *Journal of Educational Computing Research*, 58(1):87–125.

Explorando a Correção Automática de Pontuação de em Textos de Alunos

1

Abstract. *O aprendizado da correta pontuação em um texto permite uma comunicação clara e objetiva sem ambiguidades. Neste contexto, os educadores possuem um papel essencial conduto a carga de trabalho excessiva pode ser um impeditivo para a efetiva aplicação de métodos educacionais que vem de fato a ajudar os alunos. Dessa forma, nos últimos anos diversos algoritmos de Inteligência Artificial (IA) têm sido proposto em diversas aplicações educacionais, no entanto, a falta de transparência no processo de predição a é um impeditivo. Avaliamos, portanto, 5 modelos diferentes para restauração de pontuação e correção das pontuações de alunos do ensino fundamental utiulizando também o GPT-3 e GPT-4 sem passar qualquer exemplo de treinamento (zero shot learning) para corrigir os textos dos alunos em relação a pontuação. Os nossos resultados que quando treinado com os conjunto de dados Tedtalk2012 IWLST2012 e NILC o modelo BERT-LARGE obtém os melhores resultados com F1-score de 0.822% e 0.888% conduto a abordagem de zero shot com GPT-3 e GPT-4 se mostrou limitada quanto a capacidade generalização em relação aos modelos convencionais o que pode ser adicionando alguns exemplos ou através de técnicas de prompt engenering. Por fim, os modelos apresentam limitações em sentenças com erros gramaticais e na aplicação completa das regras de pontuação. Melhoramos também o processo de transparência do modelo BERT-LARGE através da ferramenta de Explainable AI, Captum no modelo BERT-LARGE. Dessa maneira, os resultados mostram a possibilidade da utilização de sistema de IA na área educacional através de um processo transparente o que pode ajudar educacadores de maneira efetiva na correçãod e textos de alunos.*

1. Introdução

A pontuação é um aspecto de grande relevância no texto escrito, visto que por meio dela o autor consegue expressar a real intenção em relação ao que pretende comunicar [Lenza and Martino 2021]. Além disso, a pontuação sinaliza pausas, inflexões da voz, separa expressões, além de esclarecer o sentido da frase, entre outras aplicações na língua escrita [Lenza and Martino 2021]. Contudo, o mau o uso da pontuação pode levar a uma compreensão errada do texto e mostrar um baixo domínio do idioma escrito [Suliman 2019]. Dessa forma, a capacidade de utilizar a pontuação de maneira adequada no processo da escrita formal faz parte dos critérios de avaliação e correção da escrita dos estudantes [Suliman 2019, Kurup et al. 2016]. Neste contexto, a correção da pontuação do aluno pode oferecer o direcionamento para melhoria da escrita, bem como, auxiliar na avaliação das principais dificuldades dos alunos em um determinado idioma [Suliman 2019].

Assim, diferentes trabalhos propõem a avaliação dos aspectos de pontuação nos escritos de alunos através do desenvolvimento de plataformas didáticas específicas, como

as desenvolvidas no trabalho de [Kurup et al. 2016], que fornecem aos alunos a possibilidade de terem as pontuações corrigidas automaticamente enquanto aprendem a correta utilização destas. Além disso, o uso de ferramentas como *Grammarly* durante a escrita proporciona um maior nível de confiança durante a escrita, principalmente entre estudantes nativos do idioma [ONeill and Russell 2019, Gaikwad et al. 2010, Cho et al. 2015]. Ademais, o desenvolvimento dessas plataformas pode ser aprimorado através de modelos de aprendizagem de máquina e técnicas de Processamento de Linguagem Natural (PLN), que permitem a correção automática de textos, como investigado em [Barbosa de Lima et al. 2023].

No contexto de PLN, a restauração de pontuação prevê a pontuação de uma sentença considerando elementos que podem ser apenas o texto de uma sentença. Existem diferentes técnicas para realização da predição usando somente o texto de um sentença, desde aquelas baseadas rotulagem de sequências como reconhecimento de entidade nomeada até aquelas baseadas em *sequence-to-sequence* [Klejch et al. 2017, Tilk and Alumäe 2016]. Em geral, modelos de aprendizagem profunda, como os algoritmos como *Bidirectional Encoder Representation* (BERT) na versão base e large, e T5, com múltiplas camadas de aprendizado, apresentam resultados melhores em relação a modelos tradicionais de aprendizado de máquina, como Conditional Random Fields (CRF) entre outros [Nielsen 2015, Lima et al. 2022, Tilk and Alumäe 2016, Nagy et al. 2021].

A predição de pontuação, ou restauração de pontuação, também pode ser usada como parte do pós-processamento a fim de melhorar a legibilidade de textos gerados por modelos de Reconhecimento Automático de Fala [Păiş and Tufiş 2021].

Assim como utilizado no contexto de Reconhecimento Automático de Fala, poderíamos utilizar a predição de pontuação no processo de correção de textos educacionais semelhantemente ao realizado em [Kurup et al. 2016]. Isso permitiria, por exemplo, a otimização do processo de correção, permitindo que educadores foquem em outras atividades como tutoria [Kurup et al. 2016]. Além disso, como investigado no mapeamento realizado por [Barbosa de Lima et al. 2023], um dos principais objetivos do desenvolvimento de algoritmos de correção automática de redação é prover *feedbacks* para alunos e educadores, já que algoritmos são mais eficientes e não são influenciados por questões subjetivas. Portanto, a aplicação da predição de pontuação em um contexto educacional traria um maior direcionamento para trabalhos futuros na área. Portanto, a aplicação da predição de pontuação em um contexto educacional traria um maior direcionamento para trabalhos futuros na área.

No entanto, aumentar a transparência e a confiabilidade em modelos de Inteligência Artificial (IA), a qual faz parte de requisitos regulatórios de alguns países em áreas como educação, tem sido um obstáculo para modelos de aprendizagem profunda [Tjoa and Guan 2020, Danilevsky et al. 2020, Khosravi et al. 2022]. Neste contexto, a área de eXplanaible AI (XAi) tem por objetivo principal trazer transparência aos algoritmos de IA. Dentre as técnicas desenvolvidas tem-se a *local explanability*, onde as características da entrada do modelo que mais contribuem para uma determinada predição são destacadas, e a *global explanability* onde se busca o entendimento do modelo como um todo [Danilevsky et al. 2020].

Até onde foi o trabalho realizado por [Barbosa de Lima et al. 2023] relacionado à aplicação de IA na área de educação em português avaliou a capacidade dos modelos de IA de aderir a regras de pontuação em língua portuguesa de maneira satisfatória em textos educacionais. Ademais, embora haja vasta literatura sobre o assunto em inglês, até onde foi possível investigar, nenhum trabalho explorou quais características do texto levam a uma predição do modelo, que é um requisito para assegurarmos a transparência e a confiabilidade de tais algoritmos [Tilk and Alumäe 2016, Nagy et al. 2021, Courtland et al. 2020].

Sendo assim, realizar uma análise automática das predições de pontuações geradas a fim de fornecer direcionamentos para trabalhos futuros na área de correção automática de textos em língua portuguesa, este trabalho analisa, através de um *benchmark*, os algoritmos de restauração de pontuação, quais sejam: BERT, T5, *Bidirectional Long Short Term Memory* (BLSTM) e Contional Random Fields (CRF), utilizados em diversos trabalhos da área de restauração de pontuação [Lima et al. 2022, Tilk and Alumäe 2016, Makhija et al. 2019, Courtland et al. 2020]. Na avaliação foram utilizados os conjuntos de dados IWLST TEDTALK2012, que contêm palestras realizadas no Tedtalk2012 e textos educacionais do Núcleo de Linguística e Computação da Universidade de São Paulo (NILC), para níveis educacionais do ensino Fundamental I e II. Nós obtivemos 82.2 F1-score utilizando BERT-LARGE testando no tradicional conjunto de dados IWLST TEDTALK2012 em português. Os resultados apresentados mostram que a técnica de predição de pontuação, aplicada em textos educacionais, tem potencial para apoiar a correção de textos de alunos no futuro.

2. Referencial Teórico

O objetivo da predição de pontuação consiste em restaurar a pontuação de texto gerado de maneira automática por algoritmos de Reconhecimento de Fala [Tilk and Alumäe 2016]. Diversas abordagens foram desenvolvidas até então, tendo mais sucesso aquelas que utilizam algoritmos de aprendizagem profunda, que assim como em atividades de Reconhecimento de Entidade Nomeadas (REN) prevem um determinado rótulo para cada *token* em um sentença. [Nagy et al. 2021, Courtland et al. 2020]. Por outro lado, as técnicas de explainable AI (XAi) têm tido o objetivo de elucidar mais claramente o processo de predição de um algoritmo em relação a uma predição. Nesse sentido, técnicas baseadas na avaliação de características de entrada do modelo têm sido aplicadas em diversos contextos, como o educacional [Oliveira et al. 2023]. Nesta seção, apresentaremos os principais algoritmos utilizados na restauração de pontuação e técnicas capazes de gerar uma explicação para as predições dessa pontuação.

2.1. Restauração de Pontuação

A rotulação de sequência em uma técnica de PLN utilizada na classificação de *tokens* em diversas atividades como reconhecimento de entidades nomeadas, identificação de palavras complexas entre outras [Gooding and Kochmar 2019, Devlin et al. 2018]. Na restauração de pontuação, podemos utilizar a rotulação de sequência para indicar a pontuação após cada palavra, anotação realizada em diversos trabalhos da área [Tilk and Alumäe 2016, Nagy et al. 2021]. Palavras sem pontuação recebem o rótulo vazio (O); seguindo o estilo de anotação, as que precisam receber vírgula recebem o

Table 1. A tabela mostra a anotação utilizada na detecção automática de pontuação.

você	acertou	não	foi
O	I-COMMA	O	I-PERIOD

rótulo I-COMMA; e as que precisam de ponto, I-PERIOD, como no exemplo disposto no Quadro 1.

Um dos modelos disseminados na área de PLN, mais especificamente de reconhecimento de entidades nomeadas, o *Conditional Random Fields* (CRF), pode ser utilizado tanto em combinação com uma rede neural, ficando na última camada, como sozinho, fazendo predições a partir da extração de características [Lu and Ng 2010, Lima et al. 2022, Lafferty et al. 2001, Manning et al. 2014]. Porém, no trabalho de [Lu and Ng 2010], uma variação do CRF, o Factorial-CRF, obtém melhores resultados na restauração de pontuação, mas os resultados não se alteram significativamente quando comparado com diferentes configurações de extração de características usando n-grams [Gravano et al. 2009]. Existem ainda outras técnicas para melhorar os resultados, como retirar palavras desnecessárias do vocabulário, como proposto por [Baldwin et al. 2013]. Assim sendo, utilizando CRF pode-se obter resultados significativos a partir de técnicas simples, por meio de um *framework* estatístico robusto.

Além das técnicas mencionadas anteriormente, criar representações de espaço vetorial de palavras permite capturar regularidades semânticas e sintáticas refinadas usando aritmética vetorial [Pennington et al. 2014]. O *Global Vectors for Word Representation* (glove), por exemplo, permitiu uma melhora significativa nos resultados de atividades como Reconhecimento de Entidades Nomeadas, que dependem de uma representação capaz de indicar o significado da palavra em um determinado contexto [Pennington et al. 2014]. Dessa forma, a proposta de [Tilk and Alumäe 2016] utiliza *Bidirectional Long Short Memory* (BLSTM, sigla em inglês) com *glove embedding* de 300 dimensões e uma camada final de CRF para predição da etiquetas nos textos. Ainda considerando o uso de apenas de texto para predição de pontuação, o trabalho realizado por [Makhija et al. 2019] utiliza BERT large, um modelo de compreensão textual pre-treinado capaz de ser aplicado a diferentes atividades como resposta a perguntas, reconhecimento de entidades nomeadas, entre outras. Os *embeddings* do BERT-large foram usados de melhorar os resultados de uma rede neural BLSTM para rotulação de sequência, obtendo 0.814 F1-score [Lima et al. 2022].

2.2. Explainable AI (XAi)

Nos últimos anos os modelos de inteligência artificial têm sido tornado cada vez mais complexo em termos de explicabilidade [Danilevsky et al. 2020]. Algoritmos tradicionais e aprendizado de máquina como regressão linear, árvore de decisão e outros provem formas simples de interpretação através das entradas e saídas dos modelos [Danilevsky et al. 2020]. Contudo, algoritmos de aprendizagem profunda não são facilmente entendíveis devido a complexidade com que são construídos através da combinação de várias camadas de aprendizado [Danilevsky et al. 2020]. Dessa forma o aprimoramento do retorno de diversas plataforma de IA, XAI tem sido explorada em duas abordagens principais, chamadas explicabilidade local e global [Yang et al. 2018,

Valenzuela-Escárcega et al. 2018, Ribeiro et al. 2016]. Isso não só fornece uma maior transparência para os usuários de sistemas de IA, como também ajuda os desenvioladores a melhorem o sistema a partir das análise geradas através da XAI [Danilevsky et al. 2020].

Um das formas de aplicação da XAI possibilita a visualização destacada de atributos de entrada evidenciando as principais características que levaram à predição (explicabilidade local) ou a procura por padrões dentro do modelo (explicabilidade global) fazem parte do processo de explicabilidade da IA conhecido como XAi [Doshi-Velez and Kim 2017]. Ferramentas como *SHapley Additive exPlanations (SHAP)* e *Captum*, atribuem a cada elemento de entrada o nível de influência que este tem na predição do rótulo pretendido [Ribeiro et al. 2016, Khosravi et al. 2022]. A ferramenta Captum foi utilizada no trabalho de [Kumar and Boulanger 2020] e avalia a aplicação da XAI em algoritmos de correção automática de redações através da extração de 1592 características do texto relacionadas a sofisticação, coesão e complexidade do texto. Enquanto isso, a ferramenta *Captum* foi adotada no trabalho de [Oliveira et al. 2023] para a verificação das características que mais contribuem para a coesão em redações em português e inglês.

A SHAP unifica diferentes outras ferramentas de XAi, como LIME e DeepLift, sendo agnóstica em relação ao modelo [Ribeiro et al. 2016, Štrumbelj and Kononenko 2014, Shrikumar et al. 2017, Datta et al. 2016, Bach et al. 2015]. Em modelos mais simples, a explicação é o próprio modelo, já em modelos mais complexos, como modelos de aprendizagem profunda, as saídas do modelo são estimadas a partir do uso de um modelo mais simples que possa manter três propriedades: (i) a acurácia, a saída do modelo mais simples é a mesma do original; (ii) ausência de uma características dada como entrada não impacta no resultado; e (iii) consistência, as propriedades anteriores devem ser mantidas e há apenas um atributo aditivo para o modelo [Ribeiro et al. 2016]. Assim, dada uma entrada, é possível estimar a influência de cada uma das características passadas para o modelo a partir de um valor-base estimado pela ferramenta [Ribeiro et al. 2016].

Enquanto isso, a *Captum* é baseada na arquitetura *transformers*, proposta por [Vaswani et al. 2017]. Ela fornece explicação local para uma variedade de atividades de IA, entre elas Reconhecimento de Entidade Nomeada, Perguntas e Resposta , Classificação de Texto, entre outras [Khosravi et al. 2022]. Além disso, permite visualizarmos o *token* que mais contribui para o rótulo predito e o que menos contribui. Por ser baseada em *deep learning*, *Captum* provê uma melhor escalabilidade enquanto dedica a cada entrada uma importância em relação à predição, medindo a fidelidade, sensitividade máxima em relação à entrada [Khosravi et al. 2022]. Além disso, também é uma ferramenta agnóstica ao modelo e pode ser melhor integrada com o *framework* pyTorch. Por fim, assim como SHAP, *Captum* também provê uma explicação baseada na influência de cada atributo de entrada do modelo destacando as partes mais relevantes da entrada [Khosravi et al. 2022].

2.3. ChatGPT e Few Shot Learning

A tecnologia de modelos de linguagem tem evoluto rapidamente nos últimos anos desde de modelos conhecidos como *encoder-only* como BERT, DEBERTa, RoBERTA e outras, modelos *decoder-only* como GPT-2, BART e outros, bem como, aqueles

baseados em *sequence-to-sequence* como T5 [Devlin et al. 2018, Radford et al. 2019, Lewis et al. 2020, He et al. 2020, Raffel et al. 2020]. Esses modelos pre-treinados permitiram que fossem atingidos o estado da arte em diversas atividades de PLN como pergunta e resposta, similaridade textual, NER, tradução entre outras [Devlin et al. 2018, Radford et al. 2019, Lewis et al. 2020, He et al. 2020, Raffel et al. 2020]. Com essa abordagem os modelos eram apenas capazes de performa uma única atividade quando realizado o ajusto fino com milhares de exemplos perdendo assim a capacidade de generalização de uma modelo de linguagem [Brown et al. 2020]. [Raffel et al. 2020] mostrou, no entanto, que modelos de linguagem são capazes de realizada diferente tarefas se perderem a capacidade de generalização podem ser adaptados para novas atividades sem ser necessário um novo modelo transferindo conhecimento entre elas [Brown et al. 2020, Alzubaidi et al. 2023]. Apesar disso, além de computacionalmente custoso para treinar cada nova atividade, em muitos casos, há uma ecasses de conjunto de dados devidamente rotulado para atividade que se está propondo o que pode ser um fator limitador em diversos cenários.

A fim de endereçar esse problema, no artigo [Brown et al. 2020], os autores mostraram que os modelos de linguagem são capazes de aprender novas atividades com nenhum ou poucos exemplos da nova atividade assim como os humanos. Os autores avaliaram que o novo modelo linguagem, GPT-3, é capaz de gerar um resultado significativo em mais de 20 atividades de diferentes e ainda outras preparados para o cenário de zero shot, quando nenhum exemplo é passado, *few shot* quando apenas alguns poucos exemplos são passados e *in-context learning*, quando o contexto da atividade é passado até o limite de entrada do modelo. O modelo GPT-3 é um modelo auto-regressivo treinado para predizer a próxima sentença em uma abordagem auto-regressiva [Brown et al. 2020]. O conjunto de dados utilizados consistiu de 5 conjuntos de textos com bilhões de *tokens* cada um que foram balanceados não proporcionalmente, bem como, de-duplicados e filtrados antes do treinamento com alguns conjuntos de dados vistos até 3.4 vezes [Brown et al. 2020]. Dessa forma, em alguma atividades o modelo com de pergunta e resposta utilizando o conjunto de dados TriviaQA, o modelo GPT-3 consegue superar o modelo T5 11B, que até então detinha o estado da arte [Brown et al. 2020].

Mais recentemente, a openai lançou o modelo de linguagem GPT-4 capaz de obter resultados ainda melhores do que o GPT-3 em diversas atividades de NLP [OpenAI 2023]. Enquanto o GPT-3 obtinha resultados entre os 10% piores resultados o GPT-4 figura entre os 10% melhores inclusive ao responder perguntas de exames desenhado para humanos que vão desde o teste para exercer advocacia nos Estatods Unidos (Uniforma Bar Exam) a exames relacionados a física e psicologia [OpenAI 2023]. Ademais, o GPT-4 também possui a capacidade de predizer texto a partir de imagem se convertendo assim em um modelo multimodel. Modelos multi-modais permitem receber como entrada ou saída mais de uma modalidade de dado, seja imagem, texto, ou áudio o que permite a aplicação de uma gama ainda maior de atividades [OpenAI 2023, Xu et al. 2023]. Por fim, o modelo GPT-4 mantém algumas das limitações do modelo GPT-3 como alucinações, limitações da janela de contexto, bem como não aprender a partir do histórico [OpenAI 2023].

Dessa forma, modelos de linguagem tem evoluído constantemente melhorando a capacidade de generalização e adaptação a novas atividades. Apesar dos avanços na

área, além de computacionalmente custoso, em muitos casos, há uma ecasses de conjunto de dados devidamente rotulado para atividade que se está propondo o que pode ser um fator limitador em diversos cenários [Alzubaidi et al. 2023]. Assim, modelos como GPT-3 e GPT-4 são facilitadores o treinamento de novas atividades de PLN por permitirem resultados significativos com poucos exemplos.

3. Trabalhos Relacionados

A restauração de pontuação tem sido o foco de diversos trabalhos da área, contudo, o conjunto de dados utilizado pela maior parte dos trabalhos se encontra em inglês. Nesta seção, avaliamos as principais metodologias aplicadas à restauração de pontuação utilizando rotulação de sequência. Em diversos trabalhos, o conjunto de dados utilizado para treinar e avaliar o modelo, em geral, é o conjunto de dados proposto por [Federico et al. 2012] que consiste em horas de áudios de palestra do TEDTALK contendo descrições em diferentes versões referenciadas por anos (e.g IWLST 2009) de cada uma delas em diversos idiomas como inglês e português.

No trabalho [Lu and Ng 2010], os autores propõe a utilização de factorial-CRF para inserção de símbolos de pontuação em textos de transcrição de Inglês e Chinês sem depender de características do áudio. O trabalho utilizou o conjunto de dados BTEC (Basic Travel Expression Corpus) que consiste em diálogos de turismo para avaliar o modelo em inglês e o conjunto de dados e CT (Challenge Task) consiste em diálogos entre idiomas relacionados a viagem que fazem parte do *corpus* IWLST2009 [Paul et al. 2010]. Os resultados mostram uma melhoria estatisticamente significativo dos resultados quando utilizado o factorial-CRF ao invés do modelo tradicional linear-CRF principalmente para o inglês chegando a mais de 88% quando avaliado no conjunto de dados BTEC. Embora apresente a avanços significativos nenhum modelo baseado em redes neurais foi avaliado no artigo, bem como, técnicas para realizar a remoção de palavras desnecessárias do texto como realizado no trabalho [Baldwin et al. 2013].

O artigo [Courtland et al. 2020] propõe uma nova forma mais eficiente de realizar o treinamento e predição de modelos de restauração de sentença prevendo a pontuação para sequência de *tokens* de uma única vez. O trabalho também propõe a utilização a agregação das predições em diferentes janelas de contexto o que melhorar ainda mais os resultados. Assim, [Courtland et al. 2020] utiliza o conjunto de dados IWSLT 2012 TED Talks em inglês com 2.1M, 296k, 12.6k de palavras para treinamento, validação e teste respectivamente. Os modelos pre-treinado avaliados foram XLNet-base,T5-base,BERT-base,ALBERT-base, DistilRoBERTa e RoBERTa-large. O modelo que mais se destaca é o RoBERTa-large que obtém uma melhoria de 48.7% em ganhos relativos e 15.3% em ganhos absolutos em relação ao estado da arte não havendo uma busca sistemática de hiper-parâmetros que pudessem melhorar ainda mais os resultados apresentados.

O trabalho de [Tilk and Alumäe 2016] utilizou as transcrições do conjunto de dados IWLST2012 em inglês para treinamento e validação, com 2.1M e 296K palavras, respectivamente, enquanto a versão IWLST2011 é usada apenas para teste. O trabalho propõe a utilização de word embeddings pre-treinado como forma de repassar para modelo alguma informação semântica das palavras e melhorar os resultados. o modelo de linguagem BERT apresenta melhores resultados para capturar a função sintática e semântica das palavras dentro do contexto utilizado sendo utilizado no tra-

balho [Makhija et al. 2019] onde os autores substituem a camada de *embeddings* pre-treinado por uma camada com BERT BASE e LARGE. Isso resulta em uma melhora de absoluta de 17 pontos de f1 score geral em relação ao trabalho [Tilk and Alumäe 2016]. Dessa forma, a utilização de *embeddings* pre-treinados mostra prover um ganho significativo principalmente quando informações de contexto são repassadas alinhadas a informações semânticas e sintáticas.

No artigo [Nagy et al. 2021] utiliza o modelo BERT na versão base e large e versão *multilanguage* Hubert e mBERT para classificar corretamente a pontuação de texto e obteve um F1 macro médio de 0.798 para o idioma inglês e 0.822 para o Húngaro. Os autores substituíram exclamação e ponto e vírgula por ponto final; e dois pontos e aspas por vírgula, removendo os hífens entre palavras (quando não separados por espaço. Quando havia o espaço eles substituíam por vírgulas). Eles também consideram todo o texto como minúsculo a fim de evitar vieses relacionados à predição do ponto final. Embora tenha apresentado resultados significativos, o trabalho não avaliou nenhum um outro algoritmo além do BERT, ademais, apenas foram avaliados textos em inglês e Húngaro [Nagy et al. 2021]. O modelo pré-treinado BERT base gerou o melhor resultado e obteve média de 0.810 F1 micro score quando comparado aos modelos CRF e BLSTM+skip-gram *embedding* de 300 dimensões. Em avaliação fora do domínio, considerando textos de obras literárias antigas em português brasileiro, o modelo BERT base e obteve média de 0.735 F1 micro score. Uma abordagem semelhante também foi seguida por [Lima et al. 2022] na avaliação das transcrições do conjunto de dados IWSLT tedtalk2012-03 em Português do Brasil proveniente de palestras do TEDTALK e proposto por [Federico et al. 2012], sendo utilizadas 139.653, 1.570 e 887 sentenças para treino, teste e validação, respectivamente.

Nesse artigo, seguiremos uma abordagem semelhante à adotada por [Lima et al. 2022], avaliando os principais modelos de restauração de pontuação, acrescentando à abordagem a análise do modelo pré-treinado T5 e BERT large avaliados no trabalho de [Courtland et al. 2020].

4. Perguntas de Pesquisa

Análise da pontuação gerada por modelos de IA a fim de verificar se ela é consistente. Destarte, as seguintes Perguntas de Pesquisa (PP) serão endereçadas ao longo desse trabalho:

PP1) Com que precisão os modelos de restauração de pontuação conseguem predizer corretamente a pontuação de uma sentença de acordo com as regras da norma padrão do idioma português? O objetivo da PP1 é identificar qual o melhor modelo e conjunto de dados a ser utilizado na restauração de pontuação e se ele segue as regras da norma culta do português.

PP2) Os modelos XAI são capazes de fornecer feedbacks adequados sobre a predição de pontuação? O objetivo da pergunta PP2 é responder se é possível obter um *feedback* que mostre a razão pela qual aquele rótulo foi predito. Isso seria útil particularmente em um cenário educacional onde a correção da pontuação precisa ser justificada para o

estudante. Para responder à PP2 utilizamos explamaible AI avaliando o modelo BERT-large.

5. Materiais e Métodos

Nesse artigo, propõe-se prever automaticamente a pontuação utilizando uma das técnicas de restauração baseada apenas no texto, a qual foi utilizada em trabalhos que analisam textos em português anteriormente. Para avaliação dos conjuntos de dados selecionados, utilizamos 4 diferentes abordagens: A primeira abordagem constitui-se da utilização do algoritmo CRF para a rotulação de sequência semelhantemente ao realizado no trabalho [Lu and Ng 2010]. Em segundo, treinaremos os modelos BLSTM+CRF com a utilização de *embeddings* pré-treinados skipgrams com 300 dimensões. Em seguida avaliaremos as abordagens utilizando modelos pre-treinados para português brasileiro, BERT-BASE, BERT-LARGE [Souza et al. 2020] e T5-BASE [Carmo et al. 2020]. Por fim, avaliamos uma abordagem zero-shot com as versões dos modelos de linguagem GPT-3.5-turbo e GPT-4 lançados pela OpenAI [Brown et al. 2020, OpenAI 2023]. Isso constitui-se uma ampliação do escopo do artigo de [Lima et al. 2022], de forma a gerar um *feedback* automático em relação a qualquer texto escrito no que se refere à pontuação usada.

5.1. Conjunto de Dados e Pré-processamento

Para o treinamento utilizamos dois conjuntos de dados, um para treino, validação e teste dos modelos (conjunto de dados TEDTALK2012 e NILC) e outro apenas para teste (MEC) devido à baixa quantidade de exemplos. Essas fontes de dados estão detalhadas a seguir.

O conjunto de dados TEDTALK2012 Utilizamos nesse artigo a versão IWLST TEDTALK2012 em português, que contém 385 textos no total, com 155.787, 1.048 e 1.886 sentenças para treino, validação e teste. O banco de dados é composto por áudios e transcrições de palestras do TEDTALK2012 utilizados tanto para Reconhecimento Automático de Fala quanto para Tradução Automática de Máquina além de restauração de pontuação [Federico et al. 2012, Tilk and Alumäe 2016]. Por fim, o conjunto de dados é utilizado para métricas de comparativos entre diversos trabalhos do estado da arte em relação a restauração de pontuação [Courtland et al. 2020, Makhija et al. 2019, Tilk and Alumäe 2016]. Neste trabalho, decidiu-se por permanecer com a divisão original do conjunto de dados IWLST TEDTALK2012, assim como em trabalhos anteriores da literatura como [Tilk and Alumäe 2016].

O conjunto de dados NILC é composto de textos educacionais agrupados de fontes diversas e classificados em quatro níveis educacionais: ensino fundamental I e II, ensino médio e superior. Os textos possuem de 300 a 596 palavras. Eles foram utilizados a fim de classificar automaticamente os textos de acordo com um dos níveis educacionais, utilizando 53 métricas de legibilidade agrupadas entre palavras, sentenças e relação entre sentenças [Murilo Gazzola 2019]. Escolhemos textos referentes ao ensino fundamental I e II devido à faixa etária dos alunos que produziram as redações que serão avaliadas. No nível de ensino fundamental I foram utilizados textos de notícias adaptados para crianças

entre 8 e 11 anos, que também foram utilizadas no trabalho de [Scarton and Aluísio 2010] para classificação automática de texto legíveis para crianças e adultos. Para ensino fundamental II, os textos utilizados correspondem a livros-texto, textos provenientes de exames educacionais do Sistema de Avaliação da Educação Básica (SAEB), E-Books, Campanha Nacional de Escolas da Comunidade (CNEC) Educacão, totalizando 1697 textos e 13016 sentenças, referenciados doravante como NILC. Por fim, o conjunto de dados foi utilizado para treinamento, teste e validação dos modelos.

O conjunto de dados MEC foi obtido por meio da análise de redações de estudantes do ensino fundamental para este artigo e possui 385 redações e 2168 sentenças. Será referenciado doravante como MEC. 5 de suas sentenças são apresentadas na Tabela 2. A quantidade de sentenças e de *tokens* para todos os conjuntos de dados podem ser vistos na Tabela 3.

Table 2. Exemplos de sentenças geradas pelos alunos do ensino fundamental, disponíveis no conjunto de dados MEC.

ID	Sentença
1	nesse dia.
2	eu e pedro estava em casa depois.
3	nos foi para escola ai.
4	depois nós foi para casa mais estava chovendo ai.
5	mesmo chovendo nós foi.

Ambos conjuntos de dados, NILC e TEDTALK2012, foram utilizados para treino, validação e teste, enquanto o conjunto de dados de redações de alunos foi utilizado só para testes e validação dos modelos, devido à baixa qualidade e quantidade de exemplos. Por fim, como o conjunto de dados do MEC dispõe apenas de correções de ponto final e vírgula, seguimos uma abordagem similar a [Nagy et al. 2021, Lima et al. 2022] e substituímos o símbolo de interrogação (?), exclamação (!) e ponto e vírgula (;), os quais foram tratados como ponto final, e as palavras foram colocadas em minúsculas.

Table 3. A tabela mostra a quantidade de rótulos e o total de tokens para no conjunto de dados TEDTALK2012 e NILC.

		I-COMMA	I-PERIOD	Total Tokens
TEDTALK2012	validation	1068	981	2049
	test	2155	1797	3952
	train	157486	151496	308982
	TOTAL	160709	154274	314983
NILC	validation	1424	1044	18596
	test	3335	2621	44161
	train	11961	9424	44161
	TOTAL	16720	13089	106918
MEC	**TOTAL**	2215	1622	52694

5.2. Extração de Características

Para nossa primeira abordagem de treinar o algoritmo CRF, escolhemos executar uma extração de características que vão desde da extração de simples n-grams como no tra-

balho realizado por [Lu and Ng 2010]. Uma outra característica relevante é inclusão de *part of speech (PoS)* que gera ganhos significativos relacionados a restauração de pontuação em [Shi et al. 2021]. As demais características estão relacionadas a palavra em si, como o case (maiuscula ou minuscula), se é um número ou sufixo. As característica são extraídas através da própria linhagem de programação (i.e *python*) ou através da utilização da biblioteca *spacy* [Honnibal et al. 15] como mostrado na tabela 4.

A segunda abordagem consiste na utilizado do modelo BLSTM com *word2vc embeddings* de 300 dimensões desenvolvido por [Hartmann et al. 2017] que consiste em uma representação semântica. Os modelos *word2vec skip-gram* são treinados de maneira que o modelo recebe uma palavra inicial e tem que prever a próxima logo em seguida. O *word2vc embeddings* já foi utilizado para extração de características em outras abordagens relacionadas a predição de pontuação como em [Che et al. 2016]. Não julgamos necessário a utilização do BERT *embedding* junto ao modelo BLSTM devido ao fato dos modelos BERT base e bert large sem BLSTM terem resultados semelhantes a abordagem proposta em [Makhija et al. 2019]. Isso fica evidente nos trabalhos de [Courtland et al. 2020] e [Nagy et al. 2021] onde o modelo BERT base atingi 80.6 f1 score e no caso de [Nagy et al. 2021] BERT BASE uncased atinge 79.8 F1 quando comparado ao resultado de 79.4 f1-score para o mesmo conjunto de dados IWLST2012 do artigo [Makhija et al. 2019]. O que seria também custoso computacionalmente dado que os modelos apresentam resultados similares. Dessa forma, nós utilizamos a mesma abordagem realizada em [Lima et al. 2022] onde o *word2vc embeddings* foi incorporado junto ao modelo BLSTM.

Table 4. Características extraídas e usadas como entrada do algoritmo CRF para a palavra casa. Utilizamos a biblioteca spacy com o modelo para aquisição do postag de cada palavra.

Nome	Valor
bias	1.0
word.lower()	'casa'
word[-3]	a
word[-2]	s
word.isupper()	FALSO
word.istitle()	FALSO
word.isdigit()	Falso
postag	NOUN
postag[2]	U
word.islower()	VERDADEIRO
word[0].isupper()	FALSO
word[0].islower()	FALSO
not word[0].isalnum()	FALSO
not word.isalnum()	FALSO
word.isalpha()	VERDADEIRO

5.3. Modelos

O algoritmo CRF foi selecionado devido à sua utilização anterior em diferentes trabalhos da literatura, inclusive em português [Lima et al. 2022], obtendo resultados significativos em alguns deles, como o de [Lu and Ng 2010]. Este último utiliza F-CRF para predição de pontuação, obtendo até 93.19 F1-score, avaliado no conjunto de dados IWLST08. Por ser um modelo mais simples e estatístico, consideramos o CRF como *baseline*, adotando uma extração de diferentes características como Part of Speech

(Pos), obtidas com a biblioteca spacy [Honnibal et al. 15]. Estas foram utilizadas como entrada do modelo usado no trabalho [Lima et al. 2022] e podem ser verificadas na Tabela 4. Nós utilizamos a biblioteca crf-suite em conjunto com a biblioteca scikit-learn [Pedregosa et al. 2011, Wijffels and Okazaki 2018]. Para obter os resultados executamos 100 interações usando a biblioteca sklearn crf suite assim como realizado no trabalho [Lima et al. 2022][1].

O BLSTM, por sua vez, já foi utilizado tanto para predição de sentenças em inglês como em português [Tilk and Alumäe 2016, Lima et al. 2022]. Assim como nos trabalhos anteriores, optou-se por usar o modelo BLSTM combinado com CRF na última camada. Além disso, adicionar *embeddings* pré-treinados se mostra capaz de melhorar significativamente os resultados do modelo [Makhija et al. 2019]. Por isso, a abordagem apresentada no presente artigo adiciona ao modelo BLSTM o *embedding* pré-treinado skip-gram com 300 dimensões pré-treinado apresentado por [Hartmann et al. 2017]. Escolhemos o skip-gram *embeddings* a fim de reproduzir a mesma configuração do modelo utilizada no artigo [Lima et al. 2022]. Não combinamos o BERT *embeddings* nessa abordagem por já comparar com o modelo BERT diretamente, bem como, devido ao custo computacional envolvido. Também executando por 100 épocas, a fim de seguir a mesma abordagem do trabalho anterior de [Lima et al. 2022](ver Tabela 5).

Os outros modelos escolhidos, BERT base e BERT large, também já apresentaram resultados significativos em trabalhos anteriores em inglês e português [Nagy et al. 2021, Lima et al. 2022]. Bem como, o modelo T5 BASE utilizado no trabalho [Courtland et al. 2020] para restauração de pontuação em inglês usando o conjunto de dados IWLST2012 Tedtalk. Por fim, optou-se executar 5 épocas com o conjunto NILC como realizado no trabalho [Makhija et al. 2019] e apenas 1 com o conjunto de dados TEDTALK2012 (vê tabela 5) dado que o conjunto de dados IWLST2012 Tedtalk possui um quantidade maior de *tokens*, e rótulos para o treinamento (ver Tabela 3). Um resumo dos parâmetros utilizados em cada modelo podem ser encontrados na tabela 5. Treinamos os modelos BERT BASE e LARGE com um tamanho de *batch size* maior (8) e menor nos demais devido a custo computacional (4 para BLSTM e 2 para T5). Os modelos BERT-BASE, BERT-LARGE e T5-BASE foram treinados utilizando o *framework* transformers [Wolf et al. 2020].

Table 5. Os parâmetros de treinamento de cada modelo.

	NILC		**IWST2012 TEDTALK**	
	batch size	Num. Epochs	batch size	Num. Epochs
BERT BASE	8	5	8	1
BERT LARGE	8	5	8	1
BLSTM+skipgram	4	100	2	100
T5 BASE	2	1	1	1

5.4. Zero Shot

Considerando os recentes avanços na construção de Modelos Grandes de Linguagem (LLM, sigla em inglês), avaliamos a utilização de Em uma abordagem *zero-shot* , nenhum exemplo é repassado ao modelo, mas ao invés disso uma instrução chamada de *prompt*.

[1]https://sklearn-crfsuite.readthedocs.io/en/latest/

Dessa forma, poderemos avaliar se um modelo de linguagem suficientemente grande é capaz de executar a mesma atividade de predição de pontuação sem que nenhum exemplo seja efetivamente fornecido. A fim de avaliar essa abordagem elaboramos um *prompt* utilizados em ambos os modelos que os instrui a agir como correto gramatical em Português brasileiro e a colocar ponto final e vírgula onde necessário da seguinte maneira:

prompt: *f"Act like punctuation corrector in brazilian portuguese: Place the 'period' and 'comma' punctuation marks in the following sentence without any other corrections: 'sentence' "*

De acordo com o trabalho de [Giray 2023] poderíamos separar nosso *prompt* nas seguintes etapas: 1) instrução: "Act like punctuation corrector"; 2) delimitação do contexto "in brazilian portuguese", 3) indicação do formato da saída " Place the 'period' and 'comma' punctuation marks in the following sentence without any other corrections" e 4) a entrada em si com a sentença original do aluno. Dessa forma, utilizamos a API da openAPI para acessar os modelos GPT-3.5-turbo e o modelo GPT-4. [2].

5.5. Métricas de Avaliação

As métricas de avaliação consistem na precisão (P), revocação (R) e F1-score (F1), amplamente usadas na avaliação de modelos de rotulação de sequência em diversos trabalhos da área desde aqueles que utilizam CRF e factorial-CRF [Lu and Ng 2010] até aqueles que utilizam modelos pre-treinados [Courtland et al. 2020, Lima et al. 2022].

Enquanto a precisão mede a quantidade de verdadeiros positivos em relação ao total (1), a revocação (2) mede o quanto foi retido pelo modelo de predição em relação àquilo que deveria ser predito e não foi. Por fim, o F1-score é o balanceamento de ambas as métricas (3). Para avaliação geral escolhemos a média micro da predição de cada modelo, assim como no trabalho de [Lima et al. 2022].

$$P = \frac{TP}{(TP + FP)} \tag{1}$$

$$R = \frac{FP}{(TP + FP)} \tag{2}$$

$$F1 = \frac{FP}{(TP + FP)} \tag{3}$$

6. Resultados

A seção apresenta os principais resultados obtidos através do treinamento e testes dos algoritmos e modelos selecionados neste artigo. As avaliação vão desde formas mais quantitativas através das métricas precisão, revocação e F1-score até a utilização de *explainable* AI.

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

6.1. PP1) Com que precisão os modelos de restauração de pontuação conseguem predizer corretamente a pontuação de uma sentença de acordo com as regras da norma padrão da língua portuguesa?

A nossa primeira perguntas de pesquisa é respondida através dos resultados apresentados: (1) tabela 6 que apresentam os resultados gerais de cada modelo treinado e testado considerando o conjunto de dados TEDTALK2012; (2) tabela 7 que apresenta os resultados dos modelos treinado e testado considerando o conjunto de dados NILC; por fim (3) a tabela 8 ajuda a responder à nossa primeira pergunta de pesquisa com os modelos apenas avaliados usando o conjunto de dados MEC.

1. Modelos testados e treinados com conjunto de dados TEDTALK2012 Na Tabela 6 vemos os resultados dos modelos treinados com o conjunto de dados TEDTALK2012. O modelo BERT-LARGE obteve melhor resultado de maneira geral com 0.822 de bleu score enquanto o modelo T5 BASE obteve o melhor resultado em termos de precisão com 0.915. Além disso, podemos observar que o modelo BERT-LARGE obtém o melhor resultado quando a predição da vírgula com 0.797 f1-score, a métrica aumenta a medida que avaliamos o ponto final, ainda assim o modelo BERT-BASE tem um resultado ligeiramente melhor de 0.995 f1-score. Por fim, o BERT LARGE obteve os melhores resultados gerais para revocação e F1-score, considerando o conjunto de dados TEDTALK2012.

Table 6. Resultados dos experimentos do conjunto de teste TEDTALK2012, para Precisão (P), Revocação (R) e F1-score (F).

	IWLST2012 TEDTALK								
	COMMA			PERIOD			Geral		
	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+Skipgrams	**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

2. Modelos treinados e testados com conjunto de dados NILC Os resultados apresentados na Tabela 7, mostram que o modelo T5 BASE obteve o melhor resultado em termos de precisão geral com 0.909, enquanto modelo BERT-LARGE obteve melhor revocação e f1-score com 0.880 e 0.888 respectivamente. Além disso, o modelo T5-BASE obtém o melhor resultado de precisão em relação a predição de vírgula enquanto o modelo BERT LARGE obtém melhores resultados para revocação e f1-score. Por fim, os modelos obtiveram resultados similares entre si quanto a predição do ponto final, com o algoritmo CRF obtendo os melhores resultados para precisão e BERT-BASE para revocação e f1-score com 0.993 e 0.995 respectivamente.

3. Modelos testados com conjunto de dados MEC A Tabela 8 mostra os modelos treinados com ambos os conjuntos de dados TEDTALK2012 e NILC e avaliados com as três métricas de avaliação precisão, revocação e f1-score. Na avaliação do modelo treinado com o primeiro conjunto de dados, TEDTALK2012, temos que o modelo que melhor se sobre sai em termos gerais é o T5-BASE com 0.607 F1-score, tendo o modelo

Table 7. Resultados dos experimentos do conjunto de teste NILC, para Precisão (P), Revocação (R) e F1-score (F).

	NILC								
	COMMA			PERIOD			Geral		
	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+Skipgrams	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

BERT-LARGE alcança melhor desempenho quando se trata de revocação. Além disso, o modelo T5-BASE obtém o melhor resultado de precisão e f1-score 0.154 e 0.205 respectivamente para predição da vírgula enquanto o algoritmo CRF obtém o melhor resultado para revocação (0.507). Algo semelhante acontece com relação a classe de ponto final, o BERT-BASE, com o modelo obtendo o melhor resultado de precisão e f1-score 1 e 0.989 respectivamente e o algoritmo CRF obtendo o melhor resultado para revocação com 0.994.

Table 8. Resultado dos experimentos para todos os modelos treinados com ambos conjuntos de dados e testado no conjunto de dados do MEC, considerando as métricas de Precisão (P), Revocação (R) e F1-score (F). (Os resultados foram convertidos para percentagem para fins comparativos).

MEC - TRAINED WITH TEDTALK2012									
	COMMA			PERIOD			GENERAL		
	P	R	F	P	R	F	P	R	F
CRF	0.086	0.179	0.116	0.912	0.974	0.942	0.470	0.679	0.556
BLSTM+Skipgrams	0.110	0.384	0.171	0.878	0.969	0.921	0.378	0.752	0.503
BERT BASE	0.104	0.494	0.172	0.876	**0.955**	0.914	0.289	0.761	0.418
BERT LARGE	0.104	**0.516**	0.173	0.878	**0.955**	0.915	0.284	**0.770**	0.415
T5 BASE	**0.154**	0.307	**0.205**	**0.947**	0.967	**0.957**	**0.523**	0.722	**0.607**
MEC - TRAINED WITH NILC									
	COMMA			PERIOD			GENERAL		
	P	R	F	P	R	F	P	R	F
CRF	0.083	0.164	0.111	**0.960**	0.974	**0.967**	0.494	0.674	**0.570**
BLSTM+Skipgrams	0.108	0.329	0.162	0.847	**0.975**	0.907	0.396	0.736	0.515
BERT BASE	**0.124**	**0.382**	0.187	0.806	0.961	0.877	0.360	0.717	0.479
BERT LARGE	0.117	0.365	0.178	0.759	0.959	0.847	0.347	0.709	0.466
T5 BASE	0.131	0.385	**0.195**	0.940	0.968	0.954	**0.429**	**0.749**	0.546
	Zero Shot								
	COMMA			PERIOD			GENERAL		
	P	R	F	P	R	F	P	R	F
GPT-3.5-turbo	0.068	0.316	0.112	0.235	0.626	0.341	0.152	0.514	0.234
GPT-4	0.072	0.406	0.123	0.479	0.902	0.625	0.240	0.742	0.363

A tabela 8 também mostra que o modelo CRF obteve uma média balanceada melhor para todas as métricas com exceção da revocação quando treinado com o conjunto de dados NILC repetindo o que acontece quando os modelos são treinados com o conjunto de dados TEDTALK2012. Os modelos obtiveram, no geral, um desempenho melhor quando treinados com o conjunto de dados TEDTALK2012, em termos de F1-score, do que os modelos treinados com o conjunto de dados NILC, com exceção do algoritmo CRF. A Tabela 8 mostra ainda como os modelos obtêm um resultado melhor predizendo ponto

final ao invés de vírgula quando treinados com o conjunto de dados TEDTALK2012. Por fim, apesar da abordagem utilizando GPT-4 se mostrar promissora, os resultados com zero shot learning usando o modelo GPT-3 e GPT-4 obtiveram um resultado aquém relação aos demais.

Por meio dos resultados apresentados na Tabela 6 e 7, temos que o modelo BERT-LARGE obteve o melhor resultado considerando o conjunto de testes do NILC. Portanto, escolhemo-lo para realizar a análise também para responder a pergunta de pesquisa 2.

6.2. PP2) Os modelos XAI são capazes de fornecer feedbacks adequados sobre a predição de pontuação?

Para responder a essa pergunta, utilizamos os dois exemplos nos quais o nosso modelo BERT-LARGE prevê corretamente os rótulos para regras gramaticais, aplicando-os agora à ferramenta *Captum* para uma análise mais aprofundada.

A ferramenta *Captum* permite que os *tokens* que os mais contribuem para uma predição do modelo sejam destacados em verde e aqueles que contribuem negativamente em vermelho. No nosso contexto, caso o modelo venha a predizer uma vírgula, os *tokens* do texto que contribuíram para essa saída serão destacados em verde como na figuras 1 e 2 e aqueles que contribuem para ponto final (mesmo a predição do modelo sendo vírgula) serão destacados em vermelho.

Figure 1. Exemplo de como a regra R6 é entendida pelo modelo de restauração de pontuação usando o modelo do NILC conforme o exemplo da Tabela 9.

Legend: Negative Neutral Positive

True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
I-COMMA	I-COMMA (1.00)	form	2.31	[CLS] form ##oso 21 de novembro de 2004 [SEP]
I-COMMA	I-COMMA (0.99)	##oso	2.18	[CLS] form ##oso 21 de novembro de 2004 [SEP]

Figure 2. Exemplo de como a regra R8 é entendida pelo modelo de restauração de pontuação usando o modelo do NILC conforme o exemplo da Tabela 9.

Legend: Negative Neutral Positive

True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
I-COMMA	I-COMMA (0.99)	se	1.12	[CLS] er ##ra se por exemplo ao colocar se vir ##gu ##la entre sujeito e verbo [SEP]
I-COMMA	I-COMMA (1.00)	exemplo	0.67	[CLS] er ##ra se por exemplo ao colocar se vir ##gu ##la entre sujeito e verbo [SEP]

Na Figura 1 é possível observar que o modelo destaca corretamente elementos da linguagem que fazem com que a pontuação seja necessária logo após a palavra ‘Formoso’ (nome de uma cidade). Por outro lado, na Figura 2, observa-se que o termo ‘por exemplo’ é destacado fortemente (pontuação/score 1.12), ainda que a expressão ‘ao colocar’ tenha uma influência negativa sobre o rótulo a ser predito, mas que é insuficiente para alterar o resultado final.

6.3. Análise de Erros dos Modelos de Acordo com a Regra gramatical

Para avaliarmos a capacidade do modelo prever corretamente as pontuações e respondermos a pergunta de pesquisa PP2, além das métricas tradicionalmente utilizadas na avaliação de modelos de restauração de pontuação (F1-score, Precisão e Revocação) faremos também uma avaliação em relação às regras gramaticais da língua portuguesa. Os exemplos apresentados foram adaptados do livro [Lenza and Martino 2021].

Table 9. Exemplo de predição de pontuação usando o modelo BERT-large treinado com o conjunto de dados NILC para os exemplos relacionados a regras de pontuação.

Regra		Exemplos
R1	predição	carlos, meu vizinho bateu com o carro.
	referência	Carlos, meu vizinho, bateu com o carro.
R2	predição	pai eu estou com fome.
	referência	Pai, eu estou com fome.
R3	predição	comprei arroz leite, carne e chuchu.
	referência	Comprei arroz, leite, carne e chuchu.
R4	predição	uma flor essa menina.
	referência	Uma flor, essa menina.
R5	predição	o decreto regulamenta os casos gerais, a portaria, os particulares.
	referência	O decreto regulamenta os casos gerais; a portaria, os particulares.
R6	predição	formoso, 21 de novembro de 2004.
	referência	Formoso, 21 de novembro de 2004.
R7	predição	a rosa entreguei a para a menina.
	referência	A rosa, entreguei-a para a menina.
R8	predição	erra se, por exemplo, ao colocar se virgula entre sujeito e verbo.
	referência	Erra-se, por exemplo, ao colocar-se vírgula entre sujeito e verbo.

As orações ou termos explicativos se separam em dois, os restritivos e não restritivos. No primeiro caso, a vírgula não pode ser utilizada porque limitaria o sentido do precedente, no segundo caso ela deve ser usada para fins de explicativos do termo anterior, cobrimos, portanto, essa definição na regra R1 [Squarisi 2021]. A regra R2 isola o vocativo do restante da frase em todos os casos [Squarisi 2021]. Os termos e orações coordenadas são aqueles que possuem um significado conjunto quando colocados lado a lado, contudo são independentes e podem ser substituídas por ponto final [Squarisi 2021]. As orações coordenas podem ser assindéticas, que necessitam obrigatoriamente apenas da vírgula, e sindéticas que necessitam também da preposição. No nosso trabalho apenas verificaremos a separação de termos coordenados com mesma função por meio da regra R3 [Squarisi 2021].

Também podemos ocultar um verbo (R4) por meio do uso da vírgula, bem como, até outro termo dependendo do contexto em que se aplica (R5) [Squarisi 2021]. Na regra R6 separamos locais e datas entre si na frase [Squarisi 2021]. Uma oração na língua portuguesa tem uma sequência de elementos bem definida como sendo sujeito (1), verbo (2), complemento verbal (3), adjunto adverbial (4) [Squarisi 2021]. Todos as vezes que algum dos elementos 2 ou 3 são deslocados para o início é necessário isolá-los com vírgula, o que está contemplado na R7 [Squarisi 2021]. Por fim, a regra R8 isola algumas expressões tais como: por exemplo, ou seja, digo, minto, então. Exemplos para cada uma das regras podem ser vistos na Tabela 9 [Squarisi 2021]. Estas regras foram selecionadas por estarem bem documentadas e discutidas através da elaboração de exemplos nos livros [Lenza and Martino 2021, Squarisi 2021]. É possível identificar pela Tabela 9 que o modelo acerta o ponto final em todos os casos, com exceção do exemplo

para regra R5. Porém, acerta a colocação da vírgula de maneira correta em apenas duas regras gramaticais (R6 e R8), falhando nas demais.

7. Discussão

De acordo com os principais resultados obtidos relacionados à PP1 apresentados nas tabelas 6 e 7 mostram que os modelos pré-treinados BERT LARGE cased e o T5 BASE obtêm melhores resultados para o conjunto de dados usado do que as abordagens mais tradicionais como o modelo BLSTM+Skipgram e o algoritmo CRF mesmo com poucas épocas de treinamento. Desde que os modelos pré-treinados sejam treinados com uma quantidade de *tokens* como a disposta no conjunto de dados TEDTALK2012 (maior do que 300 mil) ou por 5 épocas. Quando os modelos são avaliados com o conjunto de dados MEC há uma queda considerável no desempenho, como vemos na Tabela 8, com o modelo CRF obtendo o melhor resultado quando comparado aos modelos pré-treinados. Apesar dos avanços nos usos dos modelos de linguagem como GPT-3 e GPT-4 a abordagem com zero shot learning (sem apresentar nenhum exemplo) ficou abaixo do desempenho dos outros modelos.

Uma hipótese para que o modelo CRF tenha um resultado superior a modelos pre-treinados quando considerando os dados do MEC é que em algumas das sentenças avaliadas exitem erros gramaticais, como pode ser visto nos exemplos da Tabela 2. Isso mostra que os modelos ainda não são robustos o suficiente para lidarem com esse tipo de ruído. Isso é reforçado pelo fato dos resultados da Tabela 6 e 7 mostrarem que os modelos são capazes de identificar o ponto final com até 0.95 de F1-score.

Apesar de não podermos comparar diretamente os resultados com trabalhos anteriores da literatura devido à diferença no idioma e na metodologia aplicada, é possível obter resultados competitivos com outros artigos do estado da arte. De acordo com os resultados da Tabelas 6, 7 e treinar com o modelo BLSTM com embeddings não produz necessariamente um resultado significativamente melhor do que utilizando modelos pré-treinados como BERT e T5 BASE, com os resultados obtidos sendo abaixo dos modelos pré-treinados na maioria dos casos.

Semelhantemente ao que ocorre em trabalhos anteriores [Nagy et al. 2021] e [Lima et al. 2022], os modelos pré-treinados usados foram especializados sem modificações na estrutura do modelo ou o congelamento de algumas camadas durante o processo de treinamento como realizado no trabalho [Courtland et al. 2020]. Dessa forma, observamos um ganho significativo de performance (82.2 F1-score ante a 79.8 [Nagy et al. 2021] e 81 [Lima et al. 2022]). Portanto, dois fatores principais podem ter influenciado os resultados: (i) maior quantidade de exemplos de treinamento em relação ao trabalho de [Lima et al. 2022]; e (ii) o mapeamento do rótulo de QUESTION (interrogação) para PERIOD (ponto final), o que difere de ambos os trabalhos [Nagy et al. 2021, Lima et al. 2022]. Por fim, segundo [White et al. 2023] a qualidade da saída obtida está ligada diretamente a qualidade do *prompt*, portanto, na avaliação quantitativa da abordagem zero-shot usando GPT-3.5 turbo e GPT-4 a exploração de outras alternativas de *prompts* pode ter um impacto significativo no resultado final. A área de *prompt engeenering* ainda é relativamente nova [Giray 2023] havendo diversas categorias de *prompt* a serem exploradas em diferentes contextos e, portanto, como apenas uma versão delas foi explorada isso constitui-se uma limitação [White et al. 2023]. Além

disso, uma abordagem baseada em *few-shot learning* poderia vir a melhorar significativamente os resultados futuramente também [Brown et al. 2020].

Para avaliarmos a capacidade do modelo prever corretamente as pontuações e respondermos a pergunta de pesquisa PP2, além das métricas tradicionalmente utilizadas na avaliação de modelos de restauração de pontuação (F1-score, Precisão e Revocação) faremos também uma avaliação em relação às regras gramaticais da língua portuguesa. Em relação a PP2 e às regras gramaticais, o modelo não é capaz de aderir à maior parte delas, como podemos ver pela predição dos exemplos na Tabela 9. Principalmente em relação às regras como a substituição do verbo, no exemplo: ”Uma flor, essa menina”. Contudo, podemos dizer que regras mais simples podem ser aprendidas, como o uso dos termos 'por exemplo' entre vírgulas e a separação de local e data. Um dos motivos a serem levados em consideração é a grande variabilidade do uso da vírgula, que dependendo do contexto podem apresentar texto gramaticalmente corretos, embora tenham significados distintos como observado por [Lenza and Martino 2021, Squarisi 2021]. Isso pode fazer com que o algoritmo faça uma predição correta, mas inadequada para o contexto e da intenção do autor ao escrever o texto [Lenza and Martino 2021, Squarisi 2021]. Um das formas de mitigar esse problema é através do uso de aprendizado por reforço a fim de forçar padrões corretos de pontuação de acordo com o contexto utilizado, bem como, fornecer maior contexto para a predição do modelo [Zhong et al. 2017].

Portanto, restauração de pontuação não é inteiramente capaz de prever corretamente a maior parte das regras gramaticais relacionadas ao uso da vírgula na língua portuguesa, especialmente se forem regras mais específicas. Nas regras as quais o modelo foi capaz de prever corretamente (regras R6 e R8), os *tokens* relevantes para predição, de acordo com a ferramenta *Captum*, são aqueles que obrigariam o uso da vírgula no contexto usado na sentença, tal como o uso da vírgula para separar local e data (R6) e para isolar expressões como 'por exemplo' (R8), o que pode ser visto nas Figuras 1 e 2. Em cenários do mundo real tal atribuição implica que uma pontuação incorreta seja corrigida e motivo do erro adequadamente explicado. Sendo assim, a aplicação de Xai se mostra capaz de fornecer o *feedback* necessário para a correção da pontuação quando a predição é realizada de maneira correta.

Assim sendo, as restauração de pontuação se mostra uma abordagem que pode ser utilizada para corrigir possíveis erros dos alunos com precisão, prevendo algumas das pontuações corretas da sentença quando ela está bem escrita e livre de maiores erros gramaticais. No entanto, uma maior incidência de erros linguísticos não relacionados a pontuação pode ser um impeditivo para o uso efetivo de uma ferramenta que utilize a predição de pontuação como abordagem de correção de pontuação. Dessa maneira, uma forma de minimizar o problema é utilizar a correção de pontuação como último item a ser avaliado em uma aplicação de correção textual, bem como, treinar modelos com conjuntos de reais de redações para que os modelos possam ser robustos a esse tipo de ruído. Além disso, a maioria das regras não puderam ser capturadas pelo classificador, principalmente aquelas menos usadas; uma mitigação para o problema seria treinar um classificador mais simples, binário (um que indique se a pontuação está correta ou não) com um conjunto de dados maior para cada um das regras devidamente curado por um especialista. Por fim, nos exemplos de regras gramaticais em que o modelo acerta, a ferramenta *Captum* torna o *feedback* mais informativo, destacando corretamente os *tokens* mais relevantes

para aquela predição.

8. Conclusão

O presente trabalho mostra como a restauração de pontuação pode ajudar a melhorar o processo de escrita dos alunos. É possível obter comparados a trabalhos anteriores da literatura em relação à restauração de pontuação com os diferentes modelos, principalmente os que utilizam modelos pré-treinados e são treinados com um conjunto de dados maior. Contudo, os algoritmos se mostram incapazes de fornecer uma predição correta para a maioria dos exemplos exibidos na Tabela 9, que se referem às regras de aplicação da vírgula na língua portuguesa. Isso pode se dar pelo fato das diversas nuances existentes no uso da vírgula, que dependem fortemente do contexto ao redor da sentença e da real intenção do autor desta [Lenza and Martino 2021, Squarisi 2021].

Dessa maneira, os modelos apresentados nesse trabalho atingem o objetivo de termos modelos de restauração de pontuação comparado com outros resultados da literatura, conduto uma investigação mais aprofundada é necessária quando a aplicação dos modelos em relação a regras de norma culta padrão da língua portuguesa. Além disso, desde que os modelos estejam devidamente ajustados as regras é possível obter resultados de explicativos utilizando ferramentas como *Captum*.

Por fim, a simplificação do problema entre pontuação correta ou incorreta pode ser usado para tentar expandir o conceito para os tipos de avaliação, bem como, através da XAI obter um *feedback* relevante no âmbito educacional, explicando como o modelo chegou aos resultados apresentados. Ademais, uma forma de mitigar o baixo desempenho em relação a regras gramáticas pode ser o desenvolvimento de modelos que tenham como entrada não apenas a sentença, mas um contexto associado.

References

Alzubaidi, L., Bai, J., Al-Sabaawi, A., Santamaría, J., Albahri, A., Al-dabbagh, B. S. N., Fadhel, M. A., Manoufali, M., Zhang, J., Al-Timemy, A. H., et al. (2023). A survey on deep learning tools dealing with data scarcity: definitions, challenges, solutions, tips, and applications. *Journal of Big Data*, 10(1):46.

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., and Samek, W. (2015). On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140.

Baldwin, T., Cook, P., Lui, M., MacKinlay, A., and 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*, pages 356–364.

Barbosa de Lima, T., Luana Almeida da Silva, I., Laisa Soares Xavier Freitas, E., and Ferreira Mello, R. (2023). Avaliação automática de redação: Uma revisão sistemática. *Revista Brasileira de Informática na Educação*, 31:205–221.

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

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

Che, X., Wang, C., Yang, H., and Meinel, C. (2016). Punctuation prediction for unsegmented transcript based on word vector. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 654–658.

Cho, E., Niehues, J., Kilgour, K., and Waibel, A. (2015). Punctuation insertion for real-time spoken language translation. In *Proceedings of the 12th International Workshop on Spoken Language Translation: Papers*, pages 173–179.

Courtland, M., Faulkner, A., and McElvain, G. (2020). Efficient automatic punctuation restoration using bidirectional transformers with robust inference. In *Proceedings of the 17th International Conference on Spoken Language Translation*, pages 272–279.

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

Datta, A., Sen, S., and Zick, Y. (2016). Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems. In *2016 IEEE symposium on security and privacy (SP)*, pages 598–617. IEEE.

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

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

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

Gaikwad, S. K., Gawali, B. W., and Yannawar, P. (2010). A review on speech recognition technique. *International Journal of Computer Applications*, 10(3):16–24.

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

Gooding, S. and Kochmar, E. (2019). Complex word identification as a sequence labelling task. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1148–1153.

Gravano, A., Jansche, M., and Bacchiani, M. (2009). Restoring punctuation and capitalization in transcribed speech. In *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4741–4744. IEEE.

Hartmann, N., Fonseca, E., Shulby, C., Treviso, M., Rodrigues, J., and Aluisio, S. (2017). Portuguese word embeddings: Evaluating on word analogies and natural language tasks. *arXiv preprint arXiv:1708.06025*.

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

Honnibal, M., Montani, I., and AI, E. (2015–). spacy.

Khosravi, H., Shum, S. B., Chen, G., Conati, C., Tsai, Y.-S., Kay, J., Knight, S., Martinez-Maldonado, R., Sadiq, S., and Gašević, D. (2022). Explainable artificial intelligence in education. *Computers and Education: Artificial Intelligence*, 3:100074.

Klejch, O., Bell, P., and 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 (ICASSP)*, pages 5700–5704. IEEE.

Kumar, V. and Boulanger, D. (2020). Explainable automated essay scoring: Deep learning really has pedagogical value. In *Frontiers in education*, volume 5, page 572367. Frontiers Media SA.

Kurup, L., Joshi, A., and Shekhokar, N. (2016). Intelligent tutoring system for learning english punctuation. In *2016 International Conference on Computing Communication Control and automation (ICCUBEA)*, pages 1–6. IEEE.

Lafferty, J. D., McCallum, A., and 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*, ICML '01, page 282–289, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Lenza, P. and Martino, A. (2021). *Português Esquematizado*. Saraiva Educação S.A.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., and Zettlemoyer, L. (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*, pages 7871–7880, Online. Association for Computational Linguistics.

Lima, T. B. D., Miranda, P., Mello, R. F., Wenceslau, M., Bittencourt, I. I., Cordeiro, T. D., and José, J. (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*, pages 616–630. Springer.

Lu, W. and 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*, pages 177–186.

Makhija, K., Ho, T.-N., and Chng, E.-S. (2019). Transfer learning for punctuation prediction. In *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 268–273. IEEE.

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

Murilo Gazzola, Sidney Evaldo Leal, S. M. A. (2019). Predição da complexidade textual de recursos educacionais abertos em português. In *Proceedings of the Brazilian Symposium in Information and Human Language Technology*.

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

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

Oliveira, H., Ferreira Mello, R., Barreiros Rosa, B. A., Rakovic, M., Miranda, P., Cordeiro, T., Isotani, S., Bittencourt, I., and Gasevic, D. (2023). Towards explainable prediction of essay cohesion in portuguese and english. In *LAK23: 13th International Learning Analytics and Knowledge Conference*, pages 509–519.

ONeill, R. and Russell, A. (2019). Stop! grammar time: University students' perceptions of the automated feedback program grammarly. *Australasian Journal of Educational Technology*, 35(1).

OpenAI (2023). Gpt-4 technical report.

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

Paul, M., Federico, M., and Stüker, S. (2010). Overview of the iwslt 2010 evaluation campaign. In *Proceedings of the 7th International Workshop on Spoken Language Translation: Evaluation Campaign*, pages 3–27.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

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

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Ribeiro, M. T., Singh, S., and 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*, pages 1135–1144.

Scarton, C. E. and 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(1):45–61.

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

Shrikumar, A., Greenside, P., and Kundaje, A. (2017). Learning important features through propagating activation differences. In *International conference on machine learning*, pages 3145–3153. PMLR.

Souza, F., Nogueira, R., and Lotufo, R. (2020). BERTimbau: pretrained BERT models for Brazilian Portuguese. In *9th Brazilian Conference on Intelligent Systems, BRACIS, Rio Grande do Sul, Brazil, October 20-23 (to appear)*.

Squarisi, D. (2021). *50 Dicas para o uso da Pontuação. Disponível em: Minha Biblioteca*.

Štrumbelj, E. and Kononenko, I. (2014). Explaining prediction models and individual predictions with feature contributions. *Knowledge and information systems*, 41:647–665.

Suliman, F. (2019). Importance of punctuation marks for writing and reading comprehension skills. *(Faculty of Arts Journal)*, pages 29–53.

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

Tjoa, E. and Guan, C. (2020). A survey on explainable artificial intelligence (xai): Toward medical xai. *IEEE transactions on neural networks and learning systems*, 32(11):4793–4813.

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

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., Gilbert, H., Elnashar, A., Spencer-Smith, J., and Schmidt, D. C. (2023). A prompt pattern catalog to enhance prompt engineering with chatgpt. *arXiv preprint arXiv:2302.11382*.

Wijffels, J. and Okazaki, N. (2007-2018). crfsuite: Conditional random fields for labelling sequential data in natural language processing based on crfsuite: a fast implementation of conditional random fields (crfs). R package version 0.1.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T. L., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. M. (2020). Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Xu, P., Zhu, X., and Clifton, D. A. (2023). Multimodal learning with transformers: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Yang, C., Rangarajan, A., and 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 (HPCC/SmartCity/DSS)*, pages 1563–1570. IEEE.

Zhong, V., Xiong, C., and Socher, R. (2017). Seq2sql: Generating structured queries from natural language using reinforcement learning. *arXiv preprint arXiv:1709.00103*.