CRIAÇÃO DE BANCO DE FRASES VALIDADAS POR MACHINE LEARNING PARA PLATAFORMA DE APRENDIZAGEM EM LINGUA INGLESA

DALTON PACOLA FRANCO[1]; DANIEL BATISTÃO DE PAULA[1]; JOÃO RICARDO FAVAN[2]; ELOIZA MARTINS PRIMO CAPELOCI[2]

[1] Discentes em Big Data no Agronegócio na FATEC Pompeia "Shunji Nishimura", Pompeia-SP. dalton.franco@fatec.sp.gov.br; daniel.paula4@fatec.sp.gov.br

[2] Docentes do curso de Big Data no Agronegócio, FATEC Pompeia "Shunji Nishimura", Pompeia-SP. joao.favan@fatec.sp.gov.br; eloiza.capeloci@fatec.sp.gov.br

RESUMO

O uso do *Big Data* na Linguística tem revolucionado a forma como entendemos e estudamos as Línguas, abrangendo a análise de grandes volumes de dados linguísticos para entender padrões de uso da linguagem, aprimorando a eficácia no estudo de idiomas. Os modelos de *machine learning*, que é um subcampo da inteligência artificial, podem ser utilizados para aprendizado de idiomas, sendo possível definir o perfil de estudo de um usuário, seus temas e quais são os próximos passos com relação ao seu aprendizado. Foi desenvolvido neste trabalho um algoritmo utilizando a linguagem de programação Python com o modelo *Multilayer Perceptron* (MLP) para criar um validador de frases, onde foram utilizadas seis condições validadoras como entrada do algoritmo. As frases foram retiradas de dois sites distintos utilizando a técnica *web scraping*. As frases já validadas pelo algoritmo foram inseridas no banco de dados MySQL em quatorze assuntos distintos, facilitando a obtenção de vocabulário com um contexto já estabelecido. As frases validadas e categorizadas por assuntos foram disponibilizadas em uma API (*Application Programming Interface*) que foi desenvolvida com o *framework* Flask. O algoritmo de *machine learning* empregado para o validador foi o *Multilayer Perceptron*, que alcançou uma acurácia de 0.93 durante os testes pós-treinamento, demonstrando um desempenho esperado na tarefa de reconhecimento de frases. Para o classificador de assunto das frases foram comparados três modelos de *machine learning*: *Multilayer Perceptron*, *Decision Tree* e *Random Forest*, sendo feito uma comparação da acurácia dos modelos, o melhor resultado foi o do modelo MLP.

Palavras-chaves: *Big Data*; Linguística; Inteligência Artificial; *Multilayer Perceptron*.

INTRODUÇÃO

O aprendizado de idiomas tem se tornado imprescindível no mundo globalizado e que cada vez mais as tecnologias têm revolucionado a forma de conhecer e praticar

um novo idioma de maneira acessível, flexível e com eficácia utilizando aplicativos, plataformas *on-line* e *Big Data*.

O termo *Big Data* significa uma grande quantidade de dados que não conseguem ser processados por métodos convencionais, sendo a preocupação do *Big Data* entender esse grande volume de dados e atribuir significância, extraindo deles as melhores informações oriundas das mais diversas formas (NASCIMENTO, 2018).

O conceito de *Big Data* tem sido utilizado no campo da Linguística Aplicada para a compreensão dos fenômenos linguísticos, como reconhecimento de fala, organização de *corpus* linguístico, sistemas de tradução automática que vem melhorado com os últimos anos, incluindo a compreensão de como esses fenômenos são produzidos (VALENZUELA, 2022).

Uma ferramenta útil no contexto de *Big Data* que vem sendo utilizada para estudos linguísticos é o *web scraping,* que é uma técnica de extração de dados da internet, transformando dados não estruturados em um formato estruturado que pode ser armazenado e processado por um banco de dados por assuntos, que podem ser utilizados para leitura, aquisição de vocabulário personalizado, por exemplo. O *web scraping* engloba diversas técnicas de programação e tecnologias, incluindo análise de dados e análise de linguagem natural (MITCHELL, 2018).

Juntamente com o *web scraping*, uma técnica utilizada para coleta e processamento de dados em Linguística que podem ser posteriormente utilizados em diversas aplicações é o *machine learning*, que pode ser empregado de várias maneiras para facilitar o aprendizado de idiomas, sendo utilizado para a criação de *chatbots* para conversação e devolutivas imediatas sobre o desempenho, seus temas preferidos e quais são os próximos passos com relação ao seu aprendizado como o trabalho de Maeda e Morais (2017) que utilizou a técnica do *machine learning* para a criação de um *chatbot* para aprendizado de Língua Portuguesa.

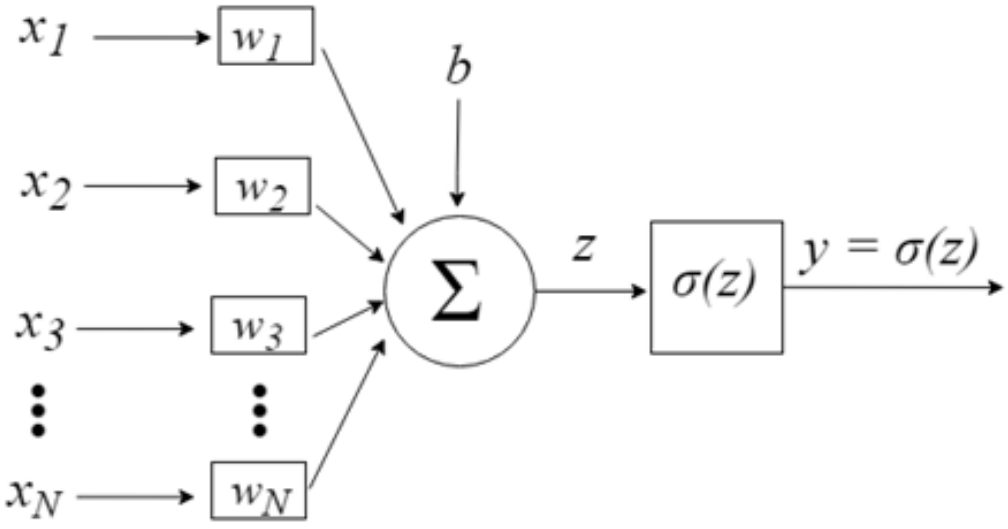
O *machine learning* é uma categoria da Inteligência Artificial que proporciona computadores a pensar e aprender por conta própria. A técnica consiste em ensinar computadores a adaptar suas operações com o intuito de aprimorá-las, visando alcançar maior precisão, cuja avaliação se baseia na frequência em que as ações desejadas são corretamente executadas (ALZUBI; NAYYAR; KUMAR, 2018).

O modelo de *machine learning* MLP (*Multilayer Perceptron*) oferece uma abordagem flexível para a análise e categorização de frases em diferentes tópicos, como demonstrado no estudo de Ali et al. (2022), em que o MLP foi treinado para diferenciar a veracidade de textos, destacando sua eficácia na detecção de notícias falsas.

Modelo Perceptron

O *Perceptron* é um neurônio artificial proposto por Frank Rosenblatt (1962), baseado nos estudos do neurofisiologista Warren McCulloch e o matemático Walter Pitts (1943) sobre um modelo matemático para o funcionamento do cérebro humano. O objeitvo do *Perceptron* é ser um modelo matemático do funcionamento de um neurônio biológico (SOUZA, 2019). Nesse neurônio é capaz de receber N entradas binárias, atribuir pesos a cada entradas e ter apenas uma resposta de saída, como descrito pela Figura 1:

Figura 1: Modelo computacional de um neurônio *Perceptron*.



Fonte: Silva; Spati; Flauzino (2010)

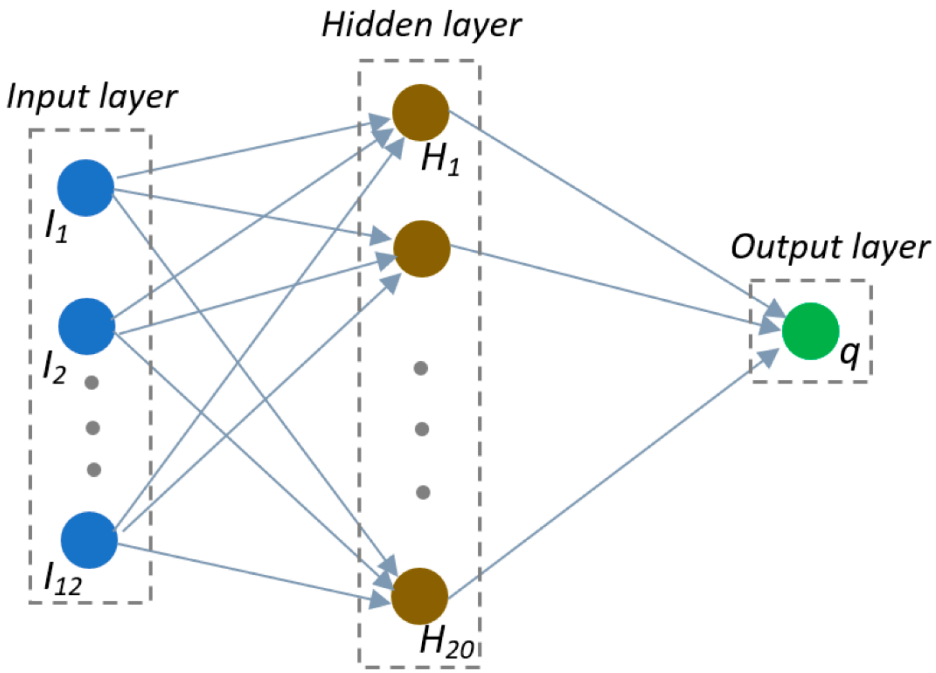
No exemplo mostrado na Figura 1, o *Perceptron* possui N entradas: x1, x2, x3, …, xn. A regra para calcular a saída proposta por Rosenblatt (1962) consiste em introduzir pesos para cada entrada w1, w2, w3, …, wn e multiplicar o valor de cada entrada com seu respectivo peso. O somatório acumula as multiplicações entre pesos e entradas, mais o potencial de repouso bias, gerando na saída, o peso ponderado do

neurônio, que irá passar por uma função de ativação responsável por produzir a saída do neurônio (SOUZA, 2019).

O MLP (*Multilayer Perceptron*) é uma técnica amplamente utilizada de aprendizado supervisionado em Redes Neurais Artificiais (RNA) cuja arquitetura tem sido empregada com sucesso em diversos problemas de previsão documentados na literatura. A arquitetura de uma rede MLP é composta essencialmente por três camadas distintas: a camada de entrada, uma ou mais camadas ocultas e a camada de saída. Cada camada oculta pode conter uma ou mais funções de ativação, que desempenham um papel crucial no processo de aprendizado da rede neural (OLATUNJI et al, 2019).

Um MLP representa uma evolução do conceito original proposto por Rosenblatt em 1959. Nessa versão são introduzidas uma ou mais camadas intermediárias entre as camadas de entrada e saída como apresentado na Figura 2:

Figura 2: Diagrama esquemático de um MLP



Fonte: ALSHAHRI; ELBISY (2023)

Os neurônios são organizados em camadas e as conexões seguem um padrão unidirecional, sempre indo das camadas inferiores para as superiores. Para as redes do tipo feedforward, feedbak e recorrentes os neurônios na mesma camada não se conectam uns aos outros, criando um arranjo hierárquico. O número de neurônios na camada de entrada é igual ao número de medidas para o problema em questão,

enquanto o número de neurônios na camada de saída corresponde ao número de classes a serem classificadas (RAMCHOUN et al, 2016).

A complexidade do MLP e suas aplicações no processamento de linguagem natural são exploradas em trabalhos de classificação, como exemplo o trabalho de Junior; Silva e Lopes (2022) que classifica frases retiradas da rede social Twitter que mencionam autorrelato de sintomas de COVID-19 em usuários infectados com a doença ou usuário não infectado pela doença, utilizando o modelo de *machine learning* MLP, revelando o potencial revolucionário das Redes Neurais Artificiais na compreensão e geração de texto, demonstrado assim que, por meio do modelo MLP, resultados altamente satisfatórios foram obtidos na classificação de texto. A linguagem natural, com sua complexidade inerente, pode ser tratada e compreendida, evidenciando a capacidade do MLP em lidar com problemas de processamento de linguagem natural.

Neste contexto, o objetivo deste trabalho foi criar um validador de frases em Língua Inglesa com base em técnicas de *machine learning*. A precisão e autenticidade das sentenças foram asseguradas por este validador, que também categoriza as sentenças em assuntos distintos, tornando-as prontamente acessíveis por meio de uma API visando facilitar o aprendizado da Língua Inglesa.

MATERIAIS E MÉTODOS

Foram desenvolvidos três *web scrapings,* dois direcionados ao site Aulas de Inglês Grátis (ROGER, 2023) que contém uma seção para textos em Língua Inglesa e um direcionado ao site de notícias USA Today (USA TODAY, 2023) e um coletor de textos de PDFs (*Portable Document Format*) com o intuito de coletar diversas frases da Língua Inglesa. A execução do *web scraping* do USA Today aconteceu entre o dia 10 de outubro de 2023 até o dia 22 de outubro de 2023, exceto nos dias 14, 15 e 18.

Os PDFs foram retirados do *website* Project Gutenberg (HART, 1971), sendo todos de domínio público. A Tabela 1 apresenta o nome do livro e do autor de cada PDF.

Tabela 1: Nome e autores dos PDFs

Nome do Livro	Autor
Forerunners and Rivals of Christianity	F. Legge
From Trail to Railway	Albert Perry Brigham
The Retreat of the Ten Thousand	C. Witt
Nigeria its Peoples and its Problems	E. D. Morel
Little Miss Oddity	Amy E. Blanchard
Anthropology	A. L. Kroeber
The Skeleton at Home	G. J. Whyte-Melville
The House on the Marsh	Florence warden
Three Pretty Maids	Byamy E. Blanchard
Repton and its Neighbourhood	F. C. Hipkins

Fonte: HART (1971)

Os *scrapings* e o coletor de frases foram feitos com a linguagem de programação Python, na versão 3.10.7 (PYTHON, 2023). Os *web scrapings* foram desenvolvidos com a biblioteca BeautifulSoup 4.11.1 (RICHARDSON, 2023) e o coletor de textos foi feito com a biblioteca Pdfplumber 0.10.2. (SINGER-VINE, 2023).

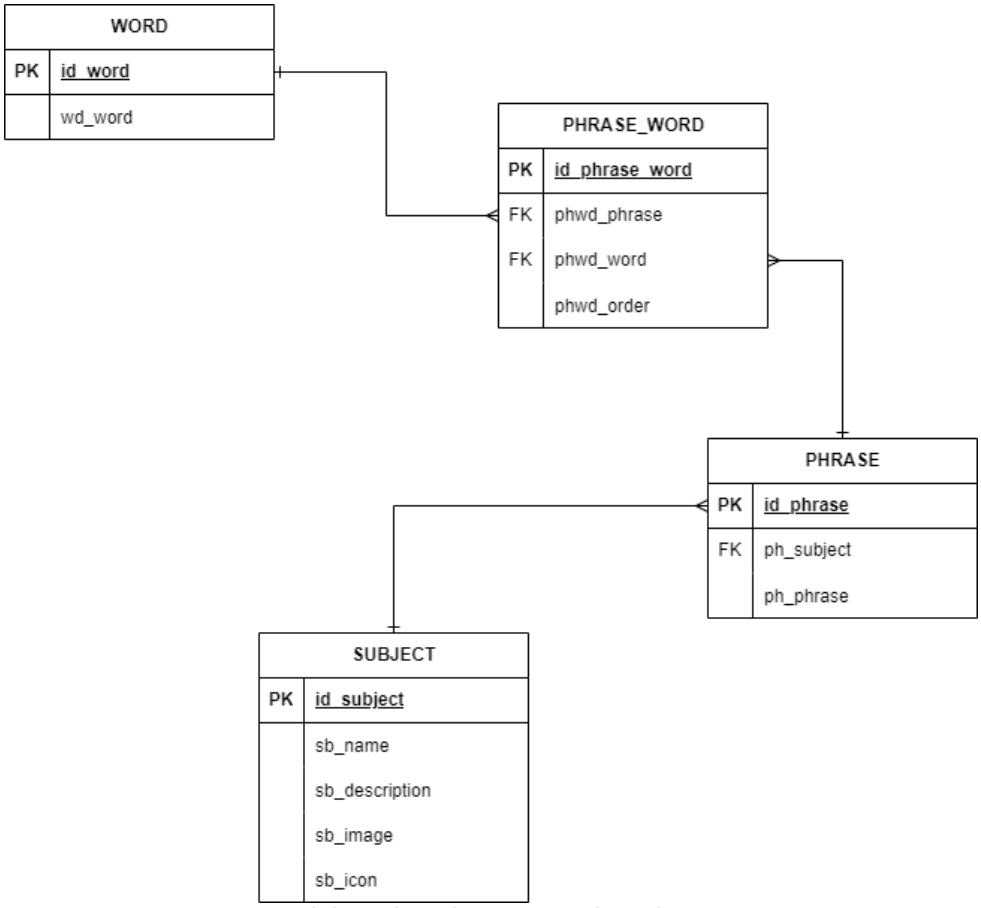
Para armazenagem das frases, foi utilizado arquivos de texto no formato CSV (*Comma-Separated Values*) e a notação de objetos JSON (*JavaScript Object Notation*) (JSON, 2023). As frases coletadas pelos *web scrapings* e pelo coletor de PDF possuem como objetivo ser validadas pelo modelo de *machine learning* e adicionadas no banco de dados.

Para a criação do *dataset* de frases de treinamento e teste dos modelos de *machine learning*, foi utilizada a ferramenta ChatGPT (OPENAI, 2023) usando o LLM (*Large Language Model*) (SEJNOWSKI, 2023) GPT-3.5 da OpenAI. Dois grupos de *datasets* foram criados: o primeiro para o validador de frases e o segundo para o classificador de categoria de vocabulário. Para o validador foi criado um *dataset* de frases válidas e outro de frases inválidas. Para o classificador de categoria foram criados os *datasets.*

Todo o *dataset* do validador foi criado através do LLM GPT-3.5 e cada frase passou por uma função Python que cria uma representação das entradas do *Perceptron* em formato de *array*. O *array* retornado dessa função é inserido em outro *array* formando uma matriz. No final, o *dataset* de treino resultou em uma matriz de *arrays* de 7 posições com valores booleanos, sendo as seis primeiras as condições e a última a classificação da frase.

O MER (Modelo Entidade Relacionamento) (TEOREY; YANG; FRY, 1986) foi feito na aplicação draw.io (JGRAPH, 2023) como apresentado a Figura 3:

Figura 3 – Modelo Entidade-Relacionamento (MER) do banco de frases.



Fonte: Elaborado pelos autores (2023)

Para este intuito, a escolha de um banco de dados relacional foi efetiva, pois existem relações. O banco de dados escolhido foi o MySQL na versão 5.5.28 (ORACLE, 2023) segundo o DB-Engines Ranking (GMBH, 2023), que é um site onde é apresentado um ranking dos bancos de dados mais utilizados do mundo. Com os dados coletados foi desenvolvida uma API (*Application Programming Interface*) utilizando a linguagem de programação Python 3.10.7 com o *framework* Flask 2.2.2 (GRINBERG, 2018).

Foram utilizadas as bibliotecas mysql.connector 8.1.0 (ORACLE, 2023) para conexão com banco de dados, jsonschema 4.19.0 (BERMAN, 2013) que foi utilizado para validar os corpos das requisições enviadas pelo lado do cliente, NLTK 3.7 (*Natural Language Toolkit*) (BIRD; KLEIN; LOPER, 2009) usada para realizar o

download das palavras que fazem parte da Língua Inglesa e tokenizar os textos e as frases para criar os *datasets.* Para dizer se a frase está gramaticalmente certa, foi utilizada a biblioteca language-tool-python 2.7.1 (MORRIS, 2022).

Foi desenvolvido um algoritmo de *machine learning* para realizar a validação da frase, sendo o *Perceptron* o algoritmo aplicado. Como descrito na Figura 1, o algoritmo recebe entradas validadoras (SOUZA, 2019). Foram elaboradas condições que validam uma frase para o banco de dados, sendo essas as entradas do *Perceptron*. Foi avaliado a criação de cinco condições que possuem pesos no quesito de ratificar se a frase é válida para ser inseridas no banco de dados apresentadas na Tabela 2:

Tabela 2: Entradas do *Perceptron* e suas descrições

Entrada	Descrição
X1	Apenas possuir caracteres especiais válidos.
X2	Possuir quantidade de palavras maior do que três.
X3	Possuir a situação de dígito acompanhado de carácter alfabético ou a situação de caractere alfabético acompanhado de dígito.
X4	Possuir apenas palavras em inglês;
X5	Possuir a situação de número seguido de carácter especial seguido de letra;
X6	Possuir gramática correta.

Fonte: Elaborado pelos autores (2023)

Com as entradas consolidadas na Tabela 2, o algoritmo irá atribuir um peso de importância de cada entrada no *Perceptron*. Com esse peso, mais o potencial de repouso *bias, o machine learning* consegue chegar em uma conclusão binária de uma hipótese, neste caso, se a frase é válida ou não, de acordo com as condições criadas no *Perceptron*.

O algoritmo validador foi feito na linguagem de programação Python 3.10.7 e dividido em leitura dos dados, pré-processamento do *dataset* e na criação da arquitetura de rede neural MLP (*Multilayer Perceptron*). Para o algoritmo ler o *dataset*, foi utilizado a função “read_csv” disponível na biblioteca Pandas 1.5.1 (MCKINNEY, 2010). Logo em seguida, o *dataset* foi divido em duas variáveis diferentes: X para *input* do *Perceptron* e Y para o *output* do *Perceptron*. Esta divisão foi feita pelo método “iloc” disponível na biblioteca Pandas 1.5.1. Em seguida, as variáveis foram divididas em dados de treino e de teste utilizando a função “train_test_split” disponível na biblioteca

scikit-learn 1.3.0 (PEDREGOSA et al, 2011). Para o que os dados tenham um desvio padrão normal padrão, foi utilizado a técnica de pré-processamento “StandardScaler” disponível na biblioteca scikit-learn 1.3.0.

Antes de passar pela predição do modelo *Multilayer Perceptron*, os registros adquiridos através dos *web scrapings* e dos coletores de PDFs foram inseridos em uma etapa de pré-processamento que eliminará registros sem utilidade nenhuma para o objetivo do artigo, onde 3 pontos definem se um registro vai ser avaliado pelo validador feito com *Multilayer Perceptron*: Não ter URL (*Uniform Resource Locator*) dentro do texto; quando o registro for separado pelo caractere “espaço” através do método “split” da classe "str” do Python, ter pelo menos dois itens no *array* retornado; não possuir erro gramátical, avaliado através da biblioteca “language_tool_python” 2.7.1.

Para a criação da arquitetura de rede neural foi utilizado a API (*Application Programming Interface*) “Keras” para criar e treinar modelos de aprendizado profundo. A API “Keras” está disponível na biblioteca TensorFlow (ABADI et al, 2015). Na API, foram criadas três camadas densas de neurônio *Perceptron*. Nas duas primeiras camadas, são criadas camadas ocultas com seis neurônios de *Perceptron* cada, sendo utilizado a função de ativação “ReLU” (*Rectified Linear Unit*). Na última cada camada densa possui apenas um neurônio *Perceptron*, por motivo de que esta é a camada que irá realizar a classificação binária e, por esse motivo, foi utilizada a função de ativação “Sigmoid”. Foi utilizado no compilador da rede neural o otimizador “Adam” como método gradiente; para função de perda foi utilizado o “binary_crossentropy” e como métrica foi aplicado o “accuracy” (LATTARI, 2021).

As frases validadas pelo algoritmo de validação serão categorizadas em quatorze tópicos distintos, apresentados na Tabela 3:

Tabela 3: Assuntos para o classificador

Assunto
Adjectives
animals_and_their_babies
climate_and_seasons
clothing_and_accessories
colors_and_numbers
days_and_months
food_and_drinks
houses_objects_and_parts
organs_and_parts_of_human_body
places_and_means_of_transport
professions_and_family_members
school_and_study_supplies
signs_and_universe
sports_and_games

Fonte: INGLES PRATICO (2023)

Com o cadastro dos temas definidos no banco de dados, foram criados quatorze *datasets* para treinamento do classificador de temas das frases, tendo cada um deles duzentas frases criadas pelo LLM GPT-3.5. Cada *dataset* foi armazenado em um arquivo CSV e logo após importado para o código fonte onde foi mesclado formando apenas um *dataset*, totalizando em 2800 frases.

O algoritmo classificador das frases foi dividido no pré-processamento dos quatorze *datasets* e na criação do modelo de *machine learning*, sendo todas as etapas feitas na linguagem de programação Python 3.10.7. Em relação ao pré-processamento, para cada *dataset*, os dados foram divididos em duas variáveis: X para as frases e Y para as categorias. Essas variáveis foram divididas em treino e teste utilizando a função “train_text_split” disponível na biblioteca scikit-learn 1.3.0. Sobre as variáveis de treino e teste foi aplicado a classe “CountVectorizer” disponível na biblioteca scikit-learn 1.3.0. Essa classe serve para transformar uma coleção de documentos de texto em uma matriz de contagem de termos (*term frequency matrix*).

Em relação a criação do algoritmo *machine learning*, foram escolhidos três modelos de para uma comparação média de acurácia. Os Modelos escolhidos foram o *Multlayer Perceptron* (MLP), *Decision Tree* (DT) e *Random Forest* (RF) (BASILE; ALTIERI; APRO, 2022). Para todos os modelos de *machine learning* foi utilizado a biblioteca scikit-learn 1.3.0. O modelo MLP foi criado pela classe “MLPClassifier”, este

MLP foi criado com apenas uma camada densa oculta com cem neurônios *Perceptron*, sendo estipulado mil épocas para o treinamento da rede neural com o inicializador de sementes do gerador de números aleatórios. O modelo *Desicion Tree* foi criado pela classe “DecisionTreeClassifier” e o modelo *Random Forest* foi criado pela classe “RandomForestClassifier”, com cem árvores de decisão com o inicializador de sementes do gerador de números aleatórios.

RESULTADO E DISCUSSÂO

O script que fez a raspagem dos dados do site USA Today e no site Aulas de Inglês Online rendeu 12 arquivos CSV. O total de registros foi de 27481. A quantidade de registros dentro de cada arquivo é apresentada na Tabela 4:

Tabela 4: Quantidade de registros por arquivo da coleta por web scraping

Nome do Arquivo	Fonte	Quantidade
2023-10-10.csv	USA Today	2392
2023-10-11.csv	USA Today	2181
2023-10-12.csv	USA Today	2247
2023-10-13.csv	USA Today	2357
2023-10-16.csv	USA Today	2432
2023-10-17.csv	USA Today	2459
2023-10-19.csv	USA Today	2432
2023-10-20.csv	USA Today	2719
2023-10-21.csv	USA Today	2128
2023-10-22.csv	USA Today	2193
110-textos-em-ingles.csv	Aulas de Inglês Online	1592
200-textos-em-ingles.csv	Aulas de Inglês Online	2349

Fonte: Elaborado pelos autores 2023

Com os dados coletados, o pré-processamento foi executado e com isso, o número de registros aumentou para 28787. No trabalho de DWIYANI; SUARJAYA e RUSJAYANTHI foi feito um classificador de letras em músicas usando o modelo *Random Forest* que também coleta dados com *web scraping* e os pré-processa antes de enviar para a predição do modelo.

Já o *script* de extração de frases de PDFs foi executado nos dez livros baixados do website Project Gutenberg e resultaram em 44174 registros no total. Cada livro rendeu a quantidade de registros exposta na Tabela 5:

Tabela 5: Registros de frases coletadas de cada PDF

Número	Nome do Livro	Quantidade de Registros	Ano de Lançamento
1	Forerunners and Rivals of Christianity	148310	1915
2	From Trail to Railway through the Appalachians	2020	1907
3	The Retreat of the Ten Thousand	1823	1896
4	Nigeria its Peoples and its Problems	3095	1912
5	Little Miss Oddity	1720	1902
6	Anthropology	8509	1923
7	The Skeleton at Home	1631	1901
8	The House on the Marsh	4377	1884
9	Three Pretty Maids	3366	1897
10	Repton and its Neighbourhood	2802	1899

Fonte: Elaborado pelos autores 2023

Apesar da quantidade de registros, nem todos são frases válidas pois em um livro existem diversos metadados, como por exemplo a data de lançamento, editora, volume, linguagem *etc*. que ficam soltos pelo arquivo e são extraídos pela ferramenta como um registro. Então, para melhorar o dataset antes de ser avaliado pelo modelo, foi utilizado o pré-processamento. Após a etapa de pré-processamento, houve uma diminuição da quantidade final de registros, eliminando 57,85% do dataset ou 25554 registros. Isso aconteceu porque os livros utilizados são antigos, sendo o mais novo deles lançado a 100 anos antes da escrita deste artigo, o que significa que graças a mudança da escrita da Língua Inglesa, o “language_tool_python”, lançado em abril de 2022, eliminou os registros que são obsoletos no que se refere ao uso da Língua atual (MORRIS, 2022).

Os *datasets* gerado pelo LLM GPT-3.5 com o objetivo de treinar o modelo MLP para classificar se a frase é válida e qual o assunto abordado teve no total 3200 registros, mas após o pré-processamento, apenas 3099 registros sobraram. A Tabela 6 mostra quantos registros foram perdidos ou ganhados após a etapa de pré-processamento de cada fonte.

Tabela 6 – Quantidade antes e depois do pré-processamento

FONTE	ANTES	DEPOIS	DIFERENÇA
Web Scraping	27481	28786	Mais 1305
PDFs	44.174	18620	Menos 25527
LLM GPT-3.5	3200	3099	Menos 101

Fonte: Elaborado pelos autores (2023)

A soma da quantidade de registros coletados pela técnica *web scraping*, pelo coletador de frases de PDFs e pelos registros gerados pelo LLM GPT-3.5 para criação do *dataset* do validador e do classificador de assuntos resultou em 50505, depois de serem pré-processados.

O modelo MLP responsável por fazer a validação das frases foi treinado com o total de 1200 frases: 600 frases válidas e 600 inválidas e resultou na acurácia de 0.93. A matriz de confusão é apresentada na Tabela 7:

Tabela 7 – Matriz de Confusão do Validador de Frases (FV = Frase Válida; FI = Frase Inválida)

	FV	FI
FV	111	8
FI	1	120

Fonte: Elaborado pelos autores (2023)

A matriz de confusão apresentada reflete o desempenho de um modelo de classificação binária nas classes FV (Falso Positivo) e FI (Falso Negativo). Foram acertadas 111 previsões de FV (Verdadeiros Positivos) e 120 previsões de FI (Verdadeiros Negativos), enquanto apenas 1 previsão de FI se mostrou incorreta (Falso Negativo), juntamente com 8 previsões equivocadas de FV (Falsos Positivos).

Nesse contexto, uma alta acurácia na identificação da classe FV é notável, sendo possível afirmar que as previsões corretas superaram consideravelmente as previsões incorretas em ambas as classes. Entretanto, a necessidade de considerar métricas adicionais, tais como *precision*, *recall*, F1-score, *mean square error* (MSE), para uma avaliação abrangente do desempenho do modelo é ressaltada. A Tabela 8 mostra os resultados das métricas que foram utilizadas.

Tabela 8 - Métricas utilizadas para avaliar o modelo MLP

Métrica	Resultado
Acurácia	0.93
Precisão	0.89
Recall	0.97
F1-Score	0.93
MSE	0.07

Fonte: Elaborado pelos autores (2023)

Os valores de acurácia (0,93), precisão (0,89), recall (0,97) e F1-Score (0,93) sugerem que o modelo de *Multilayer Perceptron* (MLP) realizou de maneira efetiva a tarefa de classificar frases como válidas ou inválidas. A alta acurácia indica que a maioria das previsões da fase de teste estão corretas, enquanto a precisão e o recall estão equilibrados, o que é um resultado sólido para uma tarefa de classificação. O F1-Score de 0,93 destaca a eficácia do modelo em encontrar um equilíbrio entre a precisão e o *recall* (CARBONELL-RIVEIRA et al, 2020). O MSE baixo de (0.07) também demonstra que o algoritmo teve um desempenho adequado no que se refere a predição dos testes. Globalmente, os números mostram um desempenho sólido do modelo na tarefa.

No total, 50505 possíveis frases passaram pelo validador. Destas, apenas 669 foram consideradas como “não frases” e descartadas, sendo que outras 49836 foram classificadas como frases. Isso significa que o modelo MLP descartou apenas 1,32% do *dataset* pré-procesado.

Tabela 9 - Acurácia dos modelos *Decision Tree* (DT), *Random Forest* (RF) e *Multilayer Perceptron* (MLP) para classificação de assunto

NOME DAS CLASSES	DT	RF	MLP
adjectives	0.28	0.55	0.64
animals_and_their_babies	0.85	0.91	0.94
climate_and_seasons	0.88	0.94	0.98
clothing_and_accessories	0.67	0.84	0.92
colors_and_numbers	0.58	0.81	0.94
days_and_months	0.68	0.81	0.98
food_and_drinks	0.68	0.84	1.00
houses_objects_and_parts	0.61	0.94	0.97
organs_and_parts_of_human_body	0.72	0.74	0.92
places_and_means_of_transport	0.33	0.47	0.67
professions_and_family_members	0.75	0.94	0.97
school_and_study_supplies	0.45	0.73	0.76

signs_and_universe	0.50	0.97	0.97
sports_and_games	0.42	0.68	0.77
Acurácia do modelo	0.56	0.79	0.89

Fonte: Elaborado pelos autores (2023).

Nesta comparação dos três modelos, o *Multilayer Perceptron* teve a acurácia maior, sendo o modelo definitivo deste classificador. A tabela mostrou que a classe "*adjectives*" obteve menor acurácia com relação às outras classes em todos os modelos de *machine learning* e isso acontece pelo motivo de todas as frases do *dataset* gerado pela junção dos 14 conterem adjetivos, logo foi a classe mais confusa para os modelos aprenderem.

Com a execução do treino e das predições, o modelo MLP responsável por classificar o assunto das frases obteve um número de itens para cada classe que são apresentados na Tabela 10:

Tabela 10 – Quantidade de frases por classe de assunto

Nome da Classe	Descrição	Quantidade
adjectives	Adjetivos	2730
animals_and_their_babies	Animais e seus filhotes	1354
climate_and_seasons	Clima e estações do ano	589
clothing_and_accessories	Vestimenta e acessórios	2429
colors_and_numbers	Cores e números	2974
days_and_months	Dias e meses	3859
food_and_drinks	Comidas e Bebidas	555
houses_objects_and_parts	Objetos da casa e suas partes	1271
organs_and_parts_of_human_body	Órgãos e partes do corpo humano	9388
places_and_means_of_transport	Lugares e meios de transporte	2119
professions_and_family_members	Profissões e membros da família	817
school_and_study_supplies	Escola e materiais de estudo	11816
signs_and_universe	Signos e o universo	4403
sports_and_games	Esportes e jogos	6201

Fonte: Elaborado pelos autores (2023)

A classe "school_and_study_supplies" foi a que apresentou o maior número de registros, representando aproximadamente 23,4% das frases, enquanto a classe "food_and_drinks" registrou a menor quantidade, correspondendo a cerca de 1,1%.

Embora os resultados indiquem que um desempenho adequado foi alcançado pelo algoritmo, a elevada quantidade de frases na classe associada a escola e materiais de estudo revela que o *dataset* não possuía um número de registros suficiente para que o objetivo deste modelo pudesse ser plenamente atingido.

O *dataset* feito para o treino do MLP responsável por classificar os assuntos das frases engloba apenas 200 frases para cada assunto, isso significa que uma quantidade de palavras muito maior do contexto de cada classe não foi vista pelo modelo como exemplo. Além disso, como se trata de um modelo de aprendizagem profunda, é necessário um *dataset* grande o suficiente para o objetivo do projeto (CEBRAL et al, 2022).

Um *dataset* maior com dezenas de milhares de frases para cada assunto definido treinaria com mais eficácia o modelo para decidir qual seria a melhor classificação de uma frase. Isso acontece porque a técnica *"*bag of words*"* utiliza a frequência das palavras presentes no texto para criar o *dataset*. Cada assunto possui seus jargões e termos, o que é um ponto de muita relevância para o modelo entender qual assunto a frase aborda, mas as palavras que formam o contexto também são diferentes em cada assunto.

CONCLUSÃO

A utilização do modelo de *machine learning Multilayer Perceptron* para a validação e classificação de frases foi executada com sucesso, provando assim a eficácia e robustez do modelo implementado. O trabalho também demonstrou que o modelo *Decision Tree* e o *Random Forest* não foram tão eficazes quanto o modelo *Multilayer Perceptron* para classificar o assunto exposto na frase. A capacidade de generalização do *Multilayer Perceptron* se destacou ao lidar com uma ampla gama de dados textuais, o que demonstrou seu potencial em aplicações práticas no *Big Data* em Linguística. Dessa forma, o *Multilayer Perceptron* apresentou-se como uma abordagem eficiente para a classificação de frases e com grande potencial para a aprendizagem de idiomas.

Para aprimorar a pesquisa, sugere-se expandir os conjuntos de treinamento e usar a técnica "bag of words" no classificador de frases válidas. Isso aumentará a diversidade de exemplos e simplificará a representação de dados, melhorando a eficiência do modelo MLP e sua precisão. Essas medidas têm o potencial de tornar o

modelo mais útil em aplicações práticas de processamento de linguagem natural e análise textual. Outro ponto interessante seria a criação de um *web scraping* para cada assunto, buscando *blogs* que os abordam. Isso faria com que o modelo aprendesse a classificar a frase pela forma que foi escrita a frase já que haveria dezenas de milhares de exemplos.

REFERÊNCIAS BIBLIOGRAFICAS

ABADI, M et al. **TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems**. 2015. Disponível em: https://www.tensorflow.org/. Acesso em: 12.out.2023.

ALI, AM.; GHALEB, Fuad A.; AL-RIMY, Blander Ali Saleh; ALSOLAMI, Fawaz Jaber; KHAN, Asif Irshad. **Deep Ensemble Fake News Detection Model Using Sequential Deep Learning Technique**. 2022. Disponível em: https://doaj.org/article/1493c146446140bab2b877ec31b9a0d9. Acesso em: 30.out.2023.

ALSHAHRI, AH.; ELBISY, MS. **Assessment of Using Artificial Neural Network and Support Vector Machine Techniques for Predicting Wave-Overtopping Discharges at Coastal Structures**. 2023. Diponível em: https://www.mdpi.com/2077-1312/11/3/539. Acesso em: 24.out.2023.

ALZUBI, J.; NAYYAR, A.; KUMAR, A. **Machine Learning from Theory to Algorithms: An Overview**. 2018. Disponível em: https://iopscience.iop.org/article/10.1088/1742-6596/1142/1/012012/pdf. p.1. Acesso em: 31.out.2023.

BASILE, AL.; ALTIERI, MP.; APRO, PH. **O uso de machine learning na classificação de textos com ênfase em fake news**. 2022. Disponível em: https://dspace.mackenzie.br/items/14bf7003-f726-4c39-becf-0a575fdf6747. p.2. Acesso em: 11.out.2023.

BERMAN, J. Jsonschema 4.19.1 **Documentation**. 2013. Disponível em: https://python-jsonschema.readthedocs.io/en/stable/. Acesso em: 25.set.2023.

BIRD, S.; KLEIN, E.; LOPER, E. **Natural language processing with Python: analyzing text with the natural language toolkit**. 2009. Disponível em: https://www.nltk.org/. Acesso em: 04.nov.2023.

CARBONELL-RIVEIRA, JP.; ESTORNELL, J.; RUIZ, LA.; TORRALBA, J. CRESPO-PEREMARCH, P. **Classification of UAV-BASED Photogrammetric Point Clouds of Riverine Species Using Machine Learning Algorithms: A Case Study in the Palancia River, Spain.** 2020. Disponível em: https://isprs-

archives.copernicus.org/articles/XLIII-B2-2020/659/2020/isprs-archives-XLIII-B2-2020-659-2020.pdf. Acesso em: 01.nov.2023.

CEBRAL, LJC.; SÁNCHEZ, JM.; FERNÁNDEZ, ER.; SÁNCHEZ, PA. **Heuristic Generation of Multispectral Labeled Point Cloud Datasets for Deep Learning Models**. 2022. Disponível em: https://isprs-archives.copernicus.org/articles/XLIII-B2-2022/571/2022/isprs-archives-XLIII-B2-2022-571-2022.pdf. p.571. Acesso em: 04.nov.2023.

DWIYANI, LKD.; SUARJAYA, IMAD.; RUSJAYANTHI, NKD. **Classification of Explicit Songs Based on Lyrics Using Random Forest Algorithm**. 2023. Disponivel em: https://journal-isi.org/index.php/isi/article/view/491. Acesso em: 24.out.2023.

GMBH. **DB-Engines Ranking**. 2023. Disponível em: https://db-engines.com/en/ranking. Acesso em: 28.set.2023.

GRINBERG, M. **Flask web development: developing web applications with python**. 2018. Disponível em: https://flask.palletsprojects.com/en/3.0.x/. Acesso em: 08.nov.2023.

HART, M. **Project Gutenberg**. 1971. Disponível em: https://www.gutenberg.org/. Acesso em: 21.out.2023

INGLES PRATICO. **Listas de Vocabulário em inglês**. 2023. Disponível em: https://inglespratico.com.br/vocabulario-em-ingles/. Acesso em: 17.out.2023

JGRAPH. **Draw.io Documentation**. 2023. Disponível em: https://www.drawio.com/doc/. Acesso em: 10.out.2023.

JSON. **Introducing JSON**. 2023. Disponível em: https://www.json.org/json-en.html. Acesso em: 21.set.2023.

JUNIOR, AC.; SILVA, CF.; LOPES, RP. **Predictive Analysis of COVID-19 Symptoms in Social Networks through Machine Learning**. 2022. Disponível em: https://doaj.org/article/458dada6d68541b3a0885ce037a59520. Acesso em: 30.out.2023.

LATTARI, L.G.. A primeira REDE NEURAL PROFUNDA: Perceptron Multicamada. Redes Neurais e Deep Learning 06. Universo Discreto. YOUTUBE. 4 ago. 2021. Disponível em: https://www.youtube.com/watch?v=e5nC31i7nVY&list=PL-t7zzWJWPtygNTsgC_M8c9a-p5biCjho&index=6. Acesso em: 12.out.2023.

MAEDA, AC.; MORAIS, SMW. **Chatbot baseado em Deep Learning: um Estudo para Língua Portuguesa**. 2017. Disponível em: https://www.researchgate.net/profile/Andherson-Maeda/publication/323675753_Chatbot_baseado_em_Deep_Learning_um_Estudo_

para_Lingua_Portuguesa/links/5aa33a5daca272d448b6cb80/Chatbot-baseado-em-Deep-Learning-um-Estudo-para-Lingua-Portuguesa.pdf. Acesso em: 23.out.2023.

MCCULLOCH, WS.; PITTS, W. **A Logical Calculus of the Ideas Immanet in Nervous Activity**. 1943. Disponível em: https://www.cs.cmu.edu/~./epxing/Class/10715/reading/McCulloch.and.Pitts.pdf. Acesso em: 30.out.2023.

MITCHELL, R. **Web Scraping With Python: Colleting More Data From the Modern Web**. 2018. Disponível em: https://edu.anarcho-copy.org/Programming%20Languages/Python/Web%20Scraping%20with%20Python,%202nd%20Edition.pdf. p.9-10. Acesso em: 12.out.2023.

MORRIS, J. **language_tool_python: a grammar checker for Python**. 2022. Disponível em: https://github.com/jxmorris12/language_tool_python. Acesso em: 14.out.2023.

NASCIMENTO, Felipe Thiago de Oliveira. **A importância do Big Data nas Organizações**. 2018. Disponível em: https://www.cin.ufpe.br/~tg/2018-2/TG_SI/fton.pdf. p.13. Acesso em: 11.out.2023.

NORIEGA, L. Multilayer Perceptron Tutorial. 2005. Disponível em: https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=4c8339b893423f1e14e34cc1543faee4e5ee4244. p.1. Acesso em: 25.set.2023.

OLATUNJI, OO.; AKINLABI, S.; MADUSHELE, N.; ADEDEJI, Paul A.; FELIX, I. **Multilayer perceptron artificial neural network for the prediction of heating value of municipal solid waste**. 2019. Disponível em: https://www.aimspress.com/fileOther/PDF/energy/energy-07-06-944.pdf. P.947. Acesso em: 24.out.2023.

ORACLE. **MySQL Documentation**. 2023. Disponível em: https://dev.mysql.com/doc/. Acesso em: 10.out.2023.

ORACLE. **Introduction to MySQL Connector/Python**. 2023. Disponível em: https://dev.mysql.com/doc/connector-python/en/connector-python-introduction.html. Acesso em: 25.set.2023.

PEDREGOSA, F et al. **Scikit-learn: Machine Learning in Python**. 2011. Disponível em: https://scikit-learn.org/0.21/documentation.html. Acesso em: 11.out.2023.

PYTHON. **The Python tutorial**. 2023. Disponível em: https://docs.python.org/3/tutorial/index.html. Acesso em: 19.set.2023.

RAMCHOUN, H.; IDRISSI, MAJ.; GHANOU, Y.; ETTAOUIL, M. **Multilayer Perceptron: Architecture Optimization and Training**. 2016. Disponível em:

https://reunir.unir.net/bitstream/handle/123456789/11569/ijimai20164_1_5_pdf_30533.pdf?sequence=1&isAllowed=y. p.27. Acesso em: 23.out.2023.

RICHARDSON, L. **Beautiful Soup Documentation**. 2015. Disponível em https://www.crummy.com/software/BeautifulSoup/bs4/doc/. Acesso em: 21 de setembro de 2023.

ROGER, J. **200 textos em inglês**. 2023. Disponível em: https://aulasdeinglesgratis.net/category/200-textos-em-ingles/. Acesso em: 10.out.2023.

SEJNOWSKI, TJ. **Large Language Models and the Reverse Turing Test**. 2023 Disponível em: https://direct.mit.edu/neco/article/35/3/309/114731/Large-Language-Models-and-the-Reverse-Turing-Test. p.310. Acesso em: 10.out.2023.

SILVA, Ivan Nunes da; SPATTI, DH.; FLAUZINO, RA. **Redes neurais artificiais para engenharia e ciências aplicadas**. 2010. Disponível em: https://repositorio.usp.br/item/002158384. Acesso em: 24.out.2023.

SINGER-VINE, J. **Pdfplumber**. 2023. Disponível em: https://pydigger.com/pypi/pdfplumber. Acesso em: 10.out.2023.

SOUZA, Henrique Alfredo de. **Treinamento de Redes Neurais com Arquitetura Multilayer Perceptron em FPGA**. 2019. Disponível em: https://repositorio.ufsc.br/bitstream/handle/123456789/199154/TCC_HenriqueSouza.pdf?isAllowed=y&sequence=1. p.7-10. Acesso em: 23.out.2023.

ROSENBLATT, F. **Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms**. 1962. Disponível em: https://babel.hathitrust.org/cgi/pt?id=mdp.39015039846566&seq=1. Acesso em: 30.out.2023.

TEOREY, TJ.; YANG, D.; FRY, JP. **A Logical Design Methodology for Relational Databases Using the Extended Entity-Relationship Model**. 1986. Disponível : https://dl.acm.org/doi/pdf/10.1145/7474.7475. p.2. Acesso em: 25.set.2023.

USA TODAY. **Breaking News And Latest News Today**. 10-15 out. 2023. Disponivel em: https://www.usatoday.com/. Acesso em: 10.out.2023.

VALENZUELA, J. **El big data en los estudios del lenguaje**. 2022. Disponível em: https://www.raco.cat/index.php/Elies/article/view/403752. p.258. Acesso em: 14.out.2023.

APÊNDICE

Repositório disponível no *link*: https://github.com/danielsbp/PhraseFactory