Impacto do Idioma no Desempenho de Algoritmos de Classificação de Texto: Um Estudo entre Português e Inglês

Jorge N. S. Pavão[1], **Kele Belloze**[1], **Gustavo Guedes**[1]

[1]Centro Federal de Educação Tecnológica Celso Suckow da Fonseca - CEFET/RJ

jorge.pavao@aluno.cefet-rj.br, kele.belloze@cefet-rj.br,
gustavo.guedes@cefet-rj.br

Abstract. *This study examines the influence of language on the performance of machine learning algorithms in text classification tasks. Two parallel Portuguese-English corpora were used, encompassing sentiment analysis and thematic categorization of scientific abstracts. Several supervised algorithms were evaluated under different preprocessing configurations. The results show no significant variations attributable exclusively to language, indicating the robustness of these techniques across linguistic boundaries. Additionally, automatic translation was found not to impair model performance, supporting its use in multilingual scenarios.*

Resumo. *Este estudo analisa a influência do idioma no desempenho de algoritmos de aprendizado de máquina em tarefas de classificação de textos. Foram utilizados dois corpora paralelos português-inglês, abrangendo análise de sentimentos e categorização temática de resumos científicos. Diversos algoritmos supervisionados foram avaliados sob diferentes configurações de pré-processamento. Os resultados indicam que não há variações significativas atribuídas exclusivamente ao idioma, indicando a robustez das técnicas frente à mudança linguística. Adicionalmente, constata-se que a tradução automática não compromete o desempenho, o que respalda sua adoção em cenários multilíngues.*

1. Introdução

O Processamento de Linguagem Natural (PLN) compreende um conjunto de técnicas computacionais utilizadas para analisar e representar textos em linguagem natural e tem por propósito auxiliar no processamento de idiomas para o desenvolvimento de diversas tarefas ou aplicações [Liddy, 2001]. A maioria das técnicas de processamento de texto foi inicialmente desenvolvida para o inglês e posteriormente adaptada para outros idiomas. Em função dessa origem e da ampla disponibilidade de ferramentas para o inglês, é comum que tarefas como análise de sentimentos e classificação de textos sejam precedidas pela tradução do conteúdo original para o inglês, a fim de possibilitar sua execução com maior suporte tecnológico.

A literatura apresenta divergências quanto à qualidade dos resultados obtidos por meio da tradução automática. Alguns estudos apontam que a tradução pode alterar traços semânticos relevantes do texto, como os sentimentos, impactando negativamente a acurácia das tarefas subsequentes [Mirkin et al., 2015; Tebbifakhr et al., 2019; Kobellarz and Silva, 2022]. Em contrapartida, outros trabalhos demonstram que essa abordagem

pode produzir resultados competitivos, comparáveis aos obtidos com textos no idioma original [Unanue et al., 2023; Araújo et al., 2020]. Há, ainda, estudos que reconhecem simultaneamente os dois aspectos: embora a tradução automática possa modificar características relevantes do texto, os modelos aplicados aos textos traduzidos ainda assim apresentam desempenho satisfatório [Salameh et al., 2015; Mohammad et al., 2016].

Nesse cenário, torna-se evidente a necessidade de investigações adicionais que avaliem os fatores que tornam a tradução automática vantajosa em determinadas situações e prejudicial em outras, especialmente no que se refere ao idioma português. No contexto específico da classificação de textos, é pertinente questionar se algoritmos de aprendizado de máquina (AM) apresentam variações de desempenho significativas entre diferentes idiomas. Essa questão, até o momento, não foi devidamente explorada na comparação entre português e inglês, embora represente um elemento relevante a ser considerado na definição de um *pipeline* de aprendizado de máquina.

Sendo assim, o presente trabalho tem como objetivo avaliar se, em uma tarefa de classificação, o nível de acerto dos modelos de AM treinados com textos em português difere significativamente dos modelos treinados com textos em inglês. Para testar essa hipótese, foram utilizados dois conjuntos de dados, um de artigos obtidos na biblioteca de publicações científicas Scielo.org[1] e outro de análise de sentimentos criado por De Azevedo et al. [2021] e expandido por Silva et al. [2024], ambos corpus paralelos português-inglês. A utilização de dois conjuntos de dados distintos justifica-se pela necessidade de avaliar o comportamento dos modelos de classificação em contextos diversos. Essa diversidade permite verificar se o impacto do idioma se mantém consistente em tarefas com características diferentes, aumentando a robustez e generalização dos resultados obtidos. Nos dois casos, foram comparados os desempenhos dos modelos treinados em português e em inglês. Os testes contemplaram a criação de modelos com o texto original, ou seja, sem pré-processamento, bem como com remoção de pontuação, remoção de stopwords, stemming e lematização.

O restante deste artigo está organizado da seguinte forma: na Seção 2 são explicados os conceitos abordados no texto. Na Seção 3 são apresentados os trabalhos relacionados. Na Seção 4 é descrita a metodologia adotada. Na Seção 5 estão os resultados obtidos e, por fim, na Seção 6 são apresentadas as considerações finais e trabalhos futuros.

2. Referencial Teórico

Para compreender o desenvolvimento deste trabalho, é importante conhecer os principais algoritmos de aprendizado de máquina e as técnicas de pré-processamento de texto. Esta seção apresenta esses conceitos, fornecendo a base teórica necessária para a análise e implementação dos modelos propostos.

2.1. Algoritmos de aprendizado de máquina

Naive Bayes é um método de classificação probabilístico, baseado no teorema de Bayes, que assume que os efeitos de um atributo sobre uma determinada classe são independentes dos valores dos outros atributos. Essa premissa é definida para simplificar o cálculo das probabilidades [Han et al., 2011].

[1]https://scielo.org/. Último acesso em 21/11/2023

Support Vector Machine (SVM) é um algoritmo que utiliza um mapeamento não linear para transformar os dados de treinamento originais em uma dimensão mais alta. Dentro desta nova dimensão, ele busca pelo hiperplano ótimo linear de separação (ou seja, uma "fronteira de decisão" que separa as tuplas de uma classe das outras). Com um mapeamento não linear apropriado para uma dimensão suficientemente alta, os dados de duas classes sempre podem ser separados por um hiperplano [Han et al., 2011].

K-Nearest Neighbor (KNN) é baseado no aprendizado por analogia, ou seja, compara uma tupla de teste com tuplas de treinamento que são similares a ela. As tuplas de treinamento são descritas por n atributos. Cada tupla representa um ponto em um espaço n-dimensional. Dessa forma, cada tupla de treinamento é armazenada em um espaço n-dimensional. Quando é fornecida uma tupla desconhecida, um classificador KNN busca no espaço pelas k tuplas de treinamento que estão mais próximas da tupla desconhecida. Essas k tuplas de treinamento são os k "vizinhos mais próximos" da tupla desconhecida. Quando os k vizinhos mais próximos são identificados, o KNN realiza a classificação da tupla desconhecida baseando-se na maioria das classes dos vizinhos [Han et al., 2011].

Random Forest é um algoritmo do tipo *ensemble* formado por muitos classificadores de árvores de decisão, de modo que a coleção de classificadores forma uma "floresta". As árvores de decisão individuais são geradas usando uma seleção aleatória de atributos em cada nó para determinar a divisão. Em outras palavras, cada árvore depende dos valores de um vetor aleatório amostrado de forma independente e com a mesma distribuição para todas as árvores na floresta. Durante a classificação, cada árvore vota e a classe mais votada é retornada [Han et al., 2011].

O último algoritmo a ser apresentado é a Regressão Logística que consiste em um algoritmo de aprendizado supervisionado usado para problemas de classificação. O algoritmo faz uso da função logística, $f(x) = \frac{1}{1+e^{-x}}$, para modelar uma variável de saída binária [Sarkar, 2019]. A principal diferença entre a regressão linear e a regressão logística é que o intervalo da regressão logística é limitado entre 0 e 1. Embora muito utilizada para classificação binária, a regressão logística também pode ser estendida para tarefas de classificação multiclasses.

2.2. Técnicas de pré-processamento

Em relação às técnicas de pré-processamento de texto utilizadas no trabalho, o *Stemming* refere-se ao processo de remover sufixos e reduzir uma palavra a uma forma básica, de modo que todas as variantes dessa palavra possam ser representadas pela mesma forma (por exemplo, 'car' e 'cars' são ambas reduzidas a 'car') [Vajjala et al., 2020].

A lematização é o processo de mapear todas as diferentes formas de uma palavra para sua forma base, ou lema. Embora isso pareça próximo à definição de *stemming*, são técnicas diferentes. Por exemplo, o adjetivo "better", quando reduzido (*stemmed*), permanece o mesmo. No entanto, na lematização, isso deve se tornar "good". A lematização requer mais conhecimento linguístico e, por este motivo, modelar e desenvolver lematizadores eficientes ainda é um problema em aberto na pesquisa em Processamento de Linguagem Natural [Vajjala et al., 2020] [Santos and Silva, 2023].

As *stopwords* são palavras que têm pouca ou nenhuma significância e geralmente são removidas do texto durante o pré-processamento para manter somente palavras com máxima relevância e contexto. *Stopwords* ocorrem com mais frequência ao se agregar um

corpus de texto com base em *tokens* singulares e verificar suas frequências. Palavras como "um", "o", "e", entre outras, geralmente são *stopwords*, mas podem não ser dependendo do contexto. Não há uma lista universal ou exaustiva de *stopwords* e frequentemente cada domínio ou idioma tem seu próprio conjunto de *stopwords* [Sarkar, 2019].

Term Frequency-Inverse Document Frequency (TF-IDF) é uma técnica que visa quantificar a importância de uma palavra específica em relação a outras palavras no documento e no *corpus*. A intuição por trás do TF-IDF é identificar palavras que ocorrem com frequência em um determinado documento, mas que são raras nos demais documentos do *corpus*. Esse padrão sugere que tais palavras são particularmente relevantes para o conteúdo específico daquele documento. A importância da palavra deve aumentar proporcionalmente à sua frequência no documento, mas, ao mesmo tempo, sua importância deve diminuir proporcionalmente à frequência da palavra nos outros documentos do *corpus* [Vajjala et al., 2020].

3. Trabalhos Relacionados

Durante a pesquisa por trabalhos relacionados, foram identificados dois estudos, que, assim como este, compararam os desempenhos de modelos treinados em português e em inglês. No entanto, nenhum deles teve como foco principal avaliar especificamente a influência do idioma no desempenho dos modelos, que é o objetivo central deste trabalho.

No estudo de Oliveira et al. [2022], utilizou-se os algoritmos KNN, SVM, *Random Forest*, *Naive Bayes*, Regressão Logística e *Stochastic Gradient Descent Classifier* (SGDC) para avaliar o desempenho de uma tarefa de classificação em português, inglês e espanhol. Apesar de o trabalho ser em parte similar, o objetivo foi diferente e pretendia avaliar se os modelos produziam o nível de desempenho desejado. Por este motivo, a metodologia adotada não permite que os resultados sejam utilizados para afirmar se o idioma influencia ou não o desempenho do algoritmo. Isso ocorre pois foram usados conjuntos de dados com informações diferentes. Por exemplo, o conjunto de dados em português era maior que os demais, logo diferenças de desempenho podem ser atribuídas aos dados de treinamento. Não há como afirmar que elas ocorrem por características do idioma.

O segundo estudo abordou a diferença entre modelos treinados em português e inglês [Pires et al., 2023]. Os autores compararam *Large Language Models*, treinados principalmente em inglês, com modelos que passaram por uma segunda etapa de treinamento para especializá-los em português. Como no artigo anterior, os objetivos eram distintos ao presente estudo e, por terem sido usados dados de treinamento diferentes, eventuais discrepâncias no resultado não podem ser atribuídas exclusivamente a características do idioma.

Na condução da pesquisa por trabalhos relacionados, ainda foram identificados dois estudos que se relacionam com este devido à criação de *corpus* paralelos português-inglês com temática semelhante. Em Soares et al. [2018] foi elaborado um conjunto de dados contendo resumos e *abstracts* de teses e dissertações disponibilizadas pela CAPES. Para o alinhamento das sentenças, foi utilizada a ferramenta Hunalign, e a qualidade do alinhamento foi validada manualmente. O *corpus* foi avaliado por meio do treinamento de sistemas de tradução automática estatística (SMT) e neural (NMT), ambos superando o desempenho do Google Translate na métrica BLEU.

Em Soares et al. [2019] foi criado um *corpus* paralelo de artigos científicos em

português, inglês e espanhol, a partir da base Scielo. O conjunto de dados desenvolvido contém textos completos, com alinhamento de sentenças realizado de forma automática e validado manualmente, atingindo mais de 98% de precisão. Além disso, foram realizados experimentos de tradução automática que demonstraram desempenho superior a trabalhos anteriores. Embora o estudo também tenha criado um *corpus* paralelo com dados da Scielo, ele enfatiza a criação e avaliação do *corpus*, e não a análise do impacto linguístico sobre o desempenho de modelos de aprendizado de máquina, como propõe este trabalho.

Apesar de possuírem temática semelhante, os dois *corpus* se diferenciam substancialmente do criado neste trabalho por passarem por alinhamento de sentença, haja vista que o objetivo era usá-los para tarefas de tradução, o que não ocorreu aqui.

Embora não tenham o objetivo de comparar a diferença de desempenho entre o português e o inglês, alguns trabalhos se aproximam deste por abordar a análise de técnicas de pré-processamento de texto em português. Oliveira and Merschmann [2021], por exemplo, avaliaram o uso das técnicas tokenização, *Part-of-Speech (PoS) Tagging*, *Stemming*, conversão para letra minúscula e extração de N-grams no contexto da análise de sentimento. Foram utilizados em conjunto com essas técnicas os algoritmos *Random Forest, Support Vector Machine* e *Multilayer Perceptron*. Os resultados mostraram que não há uma combinação melhor em todos os conjuntos de dados, e que a combinação escolhida pode afetar significativamente o desempenho preditivo do classificador.

Flores et al. [2016], por sua vez, investigam a relação entre a precisão de algoritmos de *stemming* e seu impacto na recuperação de informação (IR) em quatro idiomas: inglês, francês, português e espanhol. Em relação às diferenças entre português e inglês, os resultados indicaram que o *stemming* no português trouxe mais benefícios na recuperação de informações do que no inglês. O mesmo comportamento foi verificado no francês e espanhol, outros dois idiomas derivados do latim.

A pesquisa realizada evidenciou a escassez de trabalhos voltados à identificação dos impactos que as diferenças entre os idiomas português e inglês causam nas técnicas de processamento de linguagem natural. Esse cenário indica um campo promissor para pesquisas futuras, voltadas tanto à identificação dessas influências quanto ao desenvolvimento de abordagens mais adequadas para a língua portuguesa.

4. Metodologia

A metodologia adotada neste trabalho foi estruturada em etapas objetivas, visando a criação e preparação de conjuntos de dados que possibilitassem uma análise sobre os impactos das diferenças linguísticas entre o português e o inglês no processamento de linguagem natural. A Figura 1 apresenta uma visão geral da metodologia. A seguir, são detalhadas cada uma das etapas.

4.1. Criação do conjunto de dados de resumos

O primeiro conjunto de dados utilizado neste trabalho consiste em resumos e *abstracts* dos mesmos artigos extraídos da biblioteca Scielo. Para tanto, foi criado um *script* para extrair o identificador do artigo, o resumo, o *abstract* e o tema dos trabalhos da Scielo. Foram obtidos 40 mil artigos divididos igualmente entre os oito temas existentes no site: Ciências Exatas e da Terra, Ciências Agrárias, Ciências Biológicas, Ciências Humanas, Ciências Sociais Aplicadas, Ciências da Saúde, Engenharias e Linguística, Letras e Artes.

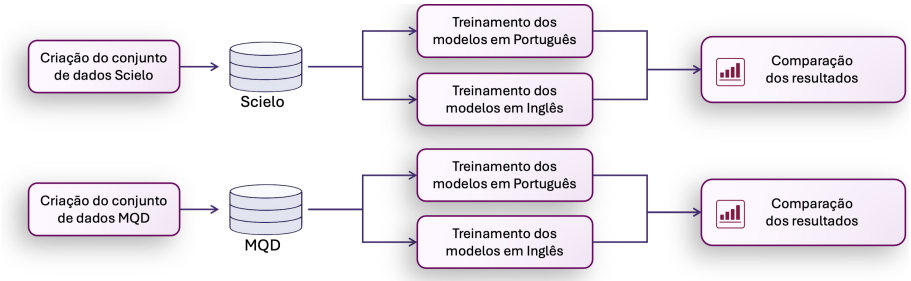

Figura 1. Metodologia para comparação dos modelos de AM em classificação de textos.

Realizou-se então, uma limpeza nos dados para excluir registros que não possuíam informações nos campos *abstract* ou resumo, bem como para ajustar o texto excluindo-se palavras e caracteres desnecessários, de modo a padronizar os dados. Por exemplo, em alguns registros, havia palavras como "Resumo" , "RESUMO: " ou o caracter "\n" indicando uma quebra de linha, e em outros não havia esses dados. Essas informações foram suprimidas para manter no conjunto de dados apenas o texto do resumo e do *abstract*. Após esse procedimento de limpeza dos dados, restaram 33.305 registros.

4.2. Preparação do conjunto de dados MQD

O segundo conjunto de dados é relativo à análise de sentimentos, nomeado MQD, criado por De Azevedo et al. [2021]. Este conjunto de dados, que originalmente possuía apenas textos em português, teve uma versão em inglês publicada por Silva et al. [2024], contendo 1.458 frases em português e suas traduções em inglês, as quais foram realizadas por meio da API do Google Translate. No experimento deste trabalho, foi necessário juntar as duas versões, agrupando os registros equivalentes para formar um *corpus* paralelo português-inglês. Não foi realizada qualquer modificação nos textos originais e na classificação dos registros.

4.3. Treinamento dos modelos

No conjunto de dados MQD, a tarefa de classificação consiste em atribuir rótulos de sentimento — negativo, neutro ou positivo — a cada texto. No conjunto de dados composto por resumos científicos da Scielo, os textos são classificados conforme as oito áreas do conhecimento mencionadas anteriormente. Em ambos os casos, foram comparados os desempenhos dos modelos treinados em português e em inglês. Em cada um dos casos de teste, foram criados dois modelos de classificação, um utilizando o texto em português e o outro em inglês. Os algoritmos avaliados foram *Naive Bayes*, *Random Forest*, *Support Vector Machine*, KNN e Regressão Logística, os quais foram selecionados porque são algoritmos consolidados e amplamente utilizados.

Considerando cada algoritmo, os seguintes testes foram realizados: (i) um modelo foi treinado sem qualquer tipo de pré-processamento, somente convertendo o conjunto de dados para letra minúscula, o que foi feito em todos os testes utilizando a biblioteca Scikit-Learn; (ii) remoção da pontuação; (iii) remoção das *stopwords*; (iv) aplicação da técnica de *stemming* utilizando a biblioteca NLTK; (v) aplicação da técnica de lematização, realizada com a biblioteca spaCy (o NLTK não efetua lematização em português); (vi)

remoção da pontuação e das *stopwords* juntamente com a realização do *stemming*; (vii) similar ao sexto teste, porém o *stemming* foi substituído pela lematização.

Todos os testes foram feitos na linguagem Python utilizando-se a biblioteca Scikit-Learn para criar os modelos. No *pipeline* de criação dos modelos, foi utilizado o TF-IDF e a métrica F1 Score Macro Avg para avaliação dos resultados. A fim de assegurar que o valor do F1 Score realmente reflete o desempenho do modelo e não foi influenciado pela amostra usada nos dados de treinamento, foram realizadas 30 repetições do *k-fold cross validation* na avaliação do modelo, com o valor de k igual a 5.

Importante destacar que, como o foco do trabalho foi comparar o desempenho dos modelos em português e em inglês, buscou-se que os testes fossem sempre realizados em igualdade de condições. Por este motivo, no *pipeline* de criação dos modelos não foi incluída a etapa de otimização dos hiperparâmetros. Foram mantidos os valores padrão da biblioteca Scikit-Learn, com exceção do KNN, em que foi escolhido k igual 8 no conjunto de dados de resumos e k igual 3 no conjunto de dados MQD por serem a quantidade de classes existentes em cada um desses conjuntos de dados.

5. Avaliação Experimental

Para comparar os resultados obtidos pelos dois idiomas, foi utilizado o teste de T de Student, que é um teste estatístico utilizado para comparar médias entre dois grupos. Ele é útil para avaliar a significância estatística de diferenças entre grupos e é amplamente utilizado em diversas áreas, como medicina, economia, psicologia e ciências sociais. Complementarmente, foi calculado o Cohen's D [Cohen, 1988], que é uma medida de tamanho de efeito, a qual avalia a magnitude da diferença entre as médias de dois grupos em termos de unidades de desvio padrão.

A Tabela 1 resume os resultados obtidos nos testes. Os valores negativos exibidos na tabela indicam que o modelo treinado em inglês teve um desempenho superior ao modelo treinado em português.

Tabela 1: Resultados dos experimentos da comparação dos algoritmos de AM na classificação de textos.

	Dataset MQD					**Dataset Resumos**				
Tipo de execução	**Média F1 Score PT**	**Média F1 Score EN**	**P Value**	**Diferença Estitística**	**Cohens D**	**Média F1 Score PT**	**Média F1 Score EN**	**P Value**	**Diferença Estitística**	**Cohens D**
KNN										
Sem pré-processamento	0,49802	0,51221	0,00000	True	-1,77072	0,73171	0,72554	0,00000	True	4,86120
Sem pontuação (1)	0,49802	0,50955	0,00000	True	-1,43794	0,73171	0,72555	0,00000	True	4,80950
Sem stopwords (2)	0,49717	0,50651	0,00002	True	-1,21549	0,73202	0,72494	0,00000	True	5,72374
Com Stemming (3)	0,52420	0,51724	0,00115	True	0,88351	0,73120	0,72458	0,00000	True	4,73358
Com Lematização (4)	0,51473	0,53563	0,00000	True	-2,63565	0,72659	0,72429	0,00000	True	1,47110
(1)+(2)+(3)	0,52901	0,51631	0,00000	True	1,70208	0,73210	0,72437	0,00000	True	6,04044
(1)+(2)+(4)	0,50083	0,53978	0,00000	True	-4,89178	0,73576	0,72627	0,00000	True	7,44630
Regressão Logística										
Sem pré-processamento	0,58961	0,59382	0,01925	True	-0,62170	0,78251	0,78735	0,00000	True	-7,18730
Sem pontuação (1)	0,58961	0,58712	0,11361	False	0,41478	0,78251	0,78740	0,00000	True	-7,01161
Sem stopwords (2)	0,55062	0,58519	0,00000	True	-4,46203	0,78313	0,78703	0,00000	True	-4,71369
Com Stemming (3)	0,62558	0,59707	0,00000	True	4,42549	0,78522	0,78617	0,00002	True	-1,18344
Com Lematização (4)	0,60443	0,61138	0,00033	True	-0,98675	0,78671	0,78769	0,00002	True	-1,20686
(1)+(2)+(3)	0,61231	0,59737	0,00000	True	1,89307	0,78487	0,78520	0,13103	False	-0,39548
(1)+(2)+(4)	0,58472	0,61765	0,00000	True	-4,63392	0,78631	0,78664	0,13729	False	-0,38905
Naive Bayes										
Sem pré-processamento	0,49528	0,53428	0,00000	True	-5,21690	0,65785	0,65690	0,00000	True	1,51081
Sem pontuação (1)	0,49528	0,53176	0,00000	True	-4,74258	0,65785	0,65689	0,00000	True	1,52500
Sem stopwords (2)	0,52509	0,56426	0,00000	True	-4,61065	0,66939	0,67410	0,00000	True	-7,78515
Com Stemming (3)	0,55507	0,55317	0,31388	False	0,26232	0,65693	0,65449	0,00000	True	4,07834
Com Lematização (4)	0,52886	0,55447	0,00000	True	-4,15354	0,65988	0,65727	0,00000	True	3,94297
(1)+(2)+(3)	0,58134	0,57890	0,34415	False	0,24627	0,66897	0,66953	0,00328	True	-0,79197
(1)+(2)+(4)	0,55095	0,58211	0,00000	True	-3,75688	0,67079	0,67360	0,00000	True	-4,31423
Random Forest										
Sem pré-processamento	0,53582	0,53418	0,45841	False	0,19273	0,64861	0,65833	0,00000	True	-5,14097
Sem pontuação (1)	0,53720	0,53015	0,00823	True	0,70650	0,64815	0,65793	0,00000	True	-4,08684
Sem stopwords (2)	0,53626	0,55471	0,00000	True	-2,08856	0,67391	0,68324	0,00000	True	-5,73769
Com Stemming (3)	0,57881	0,55009	0,00000	True	3,50862	0,66051	0,66153	0,04666	True	-0,52487

	Dataset MQD					Dataset Resumos				
Tipo de execução	**Média F1 Score PT**	**Média F1 Score EN**	**P Value**	**Diferença Estitística**	**Cohens D**	**Média F1 Score PT**	**Média F1 Score EN**	**P Value**	**Diferença Estitística**	**Cohens D**
Com Lematização (4)	0,55009	0,55652	0,00279	True	-0,80659	0,66591	0,66035	0,00000	True	2,94871
(1)+(2)+(3)	0,59071	0,57614	0,00000	True	1,50562	0,67786	0,68316	0,00000	True	-3,22757
(1)+(2)+(4)	0,55492	0,59560	0,00000	True	-4,90792	0,67366	0,68185	0,00000	True	-5,14121
SVM										
Sem pré-processamento	0,59108	0,59127	0,93022	False	-0,02271	0,77772	0,78152	0,00000	True	-3,27542
Sem pontuação (1)	0,59108	0,58448	0,00489	True	0,75558	0,77772	0,78153	0,00000	True	-3,31848
Sem stopwords (2)	0,56268	0,58285	0,00000	True	-2,30058	0,77677	0,77927	0,00000	True	-2,59460
Com Stemming (3)	0,62680	0,59973	0,00000	True	3,49216	0,77613	0,78002	0,00000	True	-3,65799
Com Lematização (4)	0,60748	0,61281	0,00507	True	-0,75216	0,77938	0,78216	0,00000	True	-2,63841
(1)+(2)+(3)	0,60760	0,59372	0,00000	True	1,67606	0,77410	0,77729	0,00000	True	-2,76291
(1)+(2)+(4)	0,59111	0,61854	0,00000	True	-2,92796	0,77783	0,77962	0,00000	True	-1,79961

Considerando que segundo Cohen [1988], o valor de D maior ou igual a 0,8 indica um grande efeito, é possível perceber primeiramente que em quase todos os casos em que a diferença entre os valores obtidos em cada idioma foi estatisticamente significante, o efeito dessa diferença foi grande e em alguns poucos casos foi médio. Essa constatação reforça que, em todos os casos em que foram observadas diferenças estatísticas, essas diferenças foram de fato relevantes.

Partindo desse princípio, os resultados podem ser interpretados sob mais de uma perspectiva. Avaliando sob o ponto de vista dos algoritmos utilizados, nota-se que, embora o Teste T tenha indicado que houve diferença estatística na maior parte dos casos, não ocorreram padrões que indicassem vantagem de um idioma em relação ao outro na maioria dos algoritmos. Por exemplo, no caso do KNN, em todos os testes houve diferença estatística entre os modelos treinados em inglês e português. No entanto, enquanto no conjunto de dados Scielo o português obteve desempenho superior em todos os testes, no conjunto de dados MQD o inglês obteve desempenho superior em cinco dos sete testes. Ou seja, não há prevalência de um idioma em relação ao outro que permaneça constante nos dois conjuntos de dados.

Um comportamento semelhante pode ser visto nos demais algoritmos. Considerando apenas os resultados dos testes em que houve diferença estatística, tem-se que com o *Naive Bayes* o português foi melhor em quatro testes no conjunto de dados Scielo, porém foi pior em cinco testes no MQD. Com o SVM, o inglês foi melhor em todos os testes no Scielo, mas só foi melhor em metade no MQD. O *Random Forest* seguiu a mesma linha, enquanto o inglês apresentou melhor desempenho em seis testes realizados no conjunto de dados Scielo, no MQD foi melhor em apenas três.

O único algoritmo que apresentou uma pequena vantagem de um idioma em relação ao outro foi a Regressão Logística. Os modelos treinados com textos em inglês foram melhores em cinco testes no Scielo e em quatro no MQD. Nota-se que foi uma pequena superioridade visto que foram realizados sete testes com cada conjunto de dados.

Ao analisar sob a perspectiva das técnicas de pré-processamento é possível perceber que na maioria dos casos não houve um idioma que tenha se destacado em relação ao outro. A Tabela 2 consolida os resultados e apresenta a quantidade de testes em que cada idioma foi melhor.

Tabela 2: Consolidação dos resultados por técnicas de pré-processamento.

	Dataset MQD			Dataset Resumos		
Tipo de execução	**PT**	**EN**	**Sem Diferença Estatística**	**PT**	**EN**	**Sem Diferença Estatística**
Sem pré-processamento	0	3	2	2	3	0
Sem pontuação (1)	2	2	1	2	3	0
Sem stopwords (2)	0	5	0	1	4	0
Com Stemming (3)	4	0	1	2	3	0
Com Lematização (4)	0	5	0	3	2	0
(1)+(2)+(3)	4	0	1	1	3	1
(1)+(2)+(4)	0	5	0	1	3	1

Nos testes realizados sem pontuação (1), com *stemming* (3), com lematização (4) e no teste em que foram combinadas a remoção da pontuação e das *stopwords* com *stemming* (1)+(2)+(3), nenhum dos idiomas foi melhor nos dois conjuntos de dados. No teste sem pré-processamento e no teste em que foram combinadas a remoção da pontuação e das *stopwords* com a lematização (1)+(2)+(4), o inglês foi melhor em ambos, porém com uma diferença pequena no conjunto de dados de resumos.

O único teste em que foi possível identificar uma prevalência relevante de um idioma foi no caso da remoção das *stopwords* (2), onde o inglês foi melhor nos cinco testes realizados com o MQD e em quatro com o Scielo.

Por fim, vale destacar que, no MQD, os textos em inglês foram gerados por meio de tradução automática a partir dos textos em português. Nesse conjunto de dados, os modelos treinados com textos em inglês superaram os textos em português em 20 testes, enquanto o inverso ocorreu em apenas 10. Diante disso, conclui-se que a tarefa de classificação em textos traduzidos produziu resultados superiores à mesma tarefa aplicada aos textos originais, ou seja, a tradução automática não prejudicou o desempenho dos modelos.

6. Considerações Finais

Este estudo apresentou uma comparação de modelos de aprendizado de máquina na classificação de textos, considerando os idiomas inglês e português, com o objetivo de avaliar se o idioma influencia na tarefa de classificação. Os resultados indicam que os algoritmos testados apresentaram desempenhos semelhantes tanto para textos em português quanto em inglês, não sendo observadas diferenças significativas que favorecessem um idioma em relação ao outro nos algoritmos *KNN*, *Random Forest*, *SVM* e *Naive Bayes*. Apenas a Regressão Logística apresentou uma pequena superioridade do idioma inglês nos dois conjuntos de dados. Adicionalmente, as técnicas de pré-processamento de texto empregadas não indicaram, de forma consistente, vantagem para os textos em inglês comparados aos textos correspondentes em português, tendo sido notada uma pequena superioridade para o inglês apenas na remoção de *stopwords*.

Os resultados obtidos indicam que, nos contextos avaliados, a utilização do idioma português para o treinamento de modelos de aprendizado de máquina em tarefas de classificação de texto não resulta em impacto significativo no desempenho. Adicionalmente, a aplicação de traduções automáticas mostrou-se uma estratégia viável, sem comprometer o nível de acerto dos modelos. Em um dos conjuntos de dados, os modelos treinados em inglês (a partir de textos traduzidos) superaram os modelos treinados

diretamente nos textos originais em português em diversos cenários. Os resultados encontrados nesse trabalho estão mais alinhados com a perspectiva apresentada por Unanue et al. [2023] e Araújo et al. [2020], segundo a qual, mesmo com possíveis alterações no conteúdo, os modelos treinados com textos traduzidos podem atingir níveis de desempenho competitivos. Esses achados fornecem subsídios para pesquisas futuras que busquem aprofundar a compreensão sobre a influência do idioma em diferentes tarefas de processamento de linguagem natural.

É importante destacar que os resultados obtidos são válidos apenas para as tarefas e domínios analisados e podem não se aplicar a tarefas linguisticamente mais sensíveis, como detecção de ironia ou análise de postura. A avaliação do impacto do idioma nesses contextos se mostra como uma boa oportunidade de trabalhos futuros.

Uma escolha metodológica que poderia ser interpretada como uma possível limitação do trabalho, é o fato de nos experimentos não ter sido realizada a etapa de otimização dos hiperparâmetros, o que poderia enfraquecer o desempenho dos modelos e impactar as conclusões. Essa decisão foi tomada com o objetivo de realizar os experimentos em igualdade de condições para os dois idiomas, visando captar apenas a influência dos idiomas nos resultados. Nesse contexto, tem-se como oportunidade de trabalhos futuros a repetição dos experimentos incluindo a etapa de otimização dos hiperparâmetros, a fim de avaliar se haveria mudança nos resultados.

Por fim, embora os resultados pareçam indicar que pode não ser tão necessário o desenvolvimento de ferramentas de PLN específicas para o português, tendo em vista os bons resultados alcançados com textos traduzidos, o trabalho não permite chegar a essa conclusão, pois os experimentos foram realizados com um número reduzido de conjuntos de dados e em tarefas específicas. Seriam necessários novos estudos explorando diferentes tarefas, conjuntos de dados e contextos para se chegar a tal conclusão.

Referências

Araújo, M., Pereira, A., and Benevenuto, F. (2020). A comparative study of machine translation for multilingual sentence-level sentiment analysis. *Information Sciences*, 512:1078–1102.

Cohen, J. (1988). *Statistical Power Analisys for Behavioral Sciences*. Lawrence Erbaum Associates, 2nd edition.

De Azevedo, G., Pettine, G., Feder, F., Portugal, G., Mendes, C. O. S., Ribeiro, R. C., Mauro, R. C., Junior, F. P., and Guedes, G. (2021). Nat: Towards an emotional agent.

Flores, F. N., Moreira, and P., V. (2016). Assessing the impact of stemming accuracy on information retrieval – a multilingual perspective. *Information Processing and Management*, 52(5):840 – 854.

Han, J., Kamber, M., and Pei, J. (2011). *Data Mining: Concepts and Techniques*. Morgan Kaufmann, 3rd edition.

Kobellarz, J. and Silva, T. (2022). Should we translate? evaluating toxicity in online comments when translating from portuguese to english. In *ACM International Conference Proceeding Series*, pages 89–98.

Liddy, E. (2001). Natural language processing. In *Encyclopedia of Library and Information Science*. Marcel Decker, Inc.

Mirkin, S., Nowson, S., Brun, C., and Perez, J. (2015). Motivating personality-aware machine translation. In *Conference Proceedings - EMNLP 2015:*, pages 1102–1108.

Mohammad, S., Salameh, M., and Kiritchenko, S. (2016). How translation alters sentiment. *Journal of Artificial Intelligence Research*, 55:95–130.

Oliveira, D. F., Nogueira, A. S., and Brito, M. A. (2022). Performance comparison of machine learning algorithms in classifying information technologies incident tickets. *AI*, 3:601–622.

Oliveira, D. N. and Merschmann, L. H. d. C. (2021). Joint evaluation of preprocessing tasks with classifiers for sentiment analysis in brazilian portuguese language. *Multimedia Tools and Applications*, 80(10):15391 – 15412.

Pires, R., Abonizio, H., Almeida, T. S., and Nogueira, R. (2023). Sabiá: Portuguese large language models. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 14197 LNAI:226 – 240.

Salameh, M., Mohammad, S., and Kiritchenko, S. (2015). Sentiment after translation: A case-study on arabic social media posts. In *Conference Proceedings - NAACL HLT 2015*, pages 767–777.

Santos, L. d. F. and Silva, M. V. d. (2023). The effect of stemming and lemmatization on portuguese fake news text classification. *arXiv preprint arXiv:2310.11344*.

Sarkar, D. (2019). *Text Analytics with Python*. Apress Berkeley, 2nd edition.

Silva, E., Silva, G., and Belloze, K. (2024). Abordagens baseadas em ontologias para análise de sentimentos em português do brasil. Dissertação de mestrado, Centro Federal de Educação Tecnológica Celso Suckow da Fonseca - CEFET/RJ.

Soares, F., Moreira, V. P., and Becker, K. (2019). A large parallel corpus of full-text scientific articles. page 3459 – 3463.

Soares, F., Yamashita, G. H., and Anzanello, M. J. (2018). A parallel corpus of theses and dissertations abstracts. In *Computational Processing of the Portuguese Language*, pages 345–352, Cham. Springer International Publishing.

Tebbifakhr, A., Bentivogli, L., Negri, M., and Turchi, M. (2019). Machine translation for machines: The sentiment classification use case. In *Conference Proceedings EMNLP-IJCNLP 2019*, pages 1368–1374.

Unanue, I., Haffari, G., and Piccardi, M. (2023). T3l: Translate-and-test transfer learning for cross-lingual text classification. *Transactions of the Association for Computational Linguistics*, 11:1147–1161.

Vajjala, S., Majumder, B., Gupta, A., and Surana, H. (2020). *Practical Natural Language Processing, A Comprehensive Guide to Building Real-World NLP Systems*. O'Reilly Media, Inc.