em questão
E-ISSN 1808-5245

Aprimorando a classificação de descrições de produtos em português com a utilização de técnicas da recuperação de informação: uma abordagem de agrupamento de descrições

Gilsiley Henrique Daru [I]

[I] Universidade Federal do Paraná, Curitiba, PR, Brasil;
ghdaru@gmail.com; ORCID https://orcid.org/0000-0002-8979-0461

Gustavo Valentim Loch [II]

[II] Universidade Federal do Paraná, Curitiba, PR, Brasil;
gustavo.gvalentim@gmail.com; ORCID https://orcid.org/0000-0002-6672-8139

Daniel Felipe Pietezak [III]

[III] Neogrid, Joinville, SC, Brasil;
danielfpj98@gmail.coml; ORCID https://orcid.org/0009-0007-2802-8805

Resumo: A crescente demanda por sistemas automatizados de classificação de produtos em plataformas de e-commerce impulsionou a busca por soluções eficientes para a categorização de produtos, especialmente em português. Este estudo investiga a adaptação de técnicas clássicas de recuperação da informação, como bag-of-words, TF e TF-IDF, para a tarefa de classificar descrições curtas de produtos. A pesquisa avalia diferentes estratégias de pré-processamento e tokenização, incluindo a análise do impacto da normalização. Os resultados demonstraram que métodos simples de recuperação da informação, quando combinados com pré-processamento adequado e otimização de parâmetros, podem alcançar desempenho significativamente superior.

Palavras-chave: aprendizado de máquina; processamento de linguagem natural; classificação de texto; descrição do produto; texto curto; bag of words; frequência de termos; frequência inversa de documentos

1 Introdução

O campo da ciência da informação teve seu início logo após a Segunda Guerra Mundial, junto com uma série de novos campos de pesquisa, como a ciência da

computação, demonstrando que desde sua concepção esse ramo teria um foco altamente interdisciplinar (Saracevic, 1999). Assim sendo, pode-se definir formalmente que a área da ciência da informação engloba o estudo das regras e métodos para coletar, processar e analisar informações, muitas vezes fazendo uso de técnicas de outros ramos de pesquisa (Ershov, 1986).

Na era atual de digitalização massiva, onde mais de 2,5 quintilhões de bytes são gerados diariamente e tem perspectiva para alcançar um número de 175 quintilhões de bytes no ano de 2025 dada a inserção de novas tecnologias como o 5G para o público (Gandomi; Heider, 2015; Ishimaru, 2021). Esse grande conjunto de dados é definido formalmente como *big data*. A definição formal do termo é de uma coleção de informações que não podem ser coletados, administrados e processados por técnicas tradicionais de computadores (Wang *et. al*., 2018).

Esta nova era da informação traz consigo uma preocupação significativa para o campo da ciência da informação, pois demanda o desenvolvimento de novas metodologias capazes de lidar com esse desafio. Para superar esse obstáculo, a ciência da informação tem contribuído em conjunto com os campos da estatística e da computação para a formação do campo da ciência da computação (Marchionini, 2017).

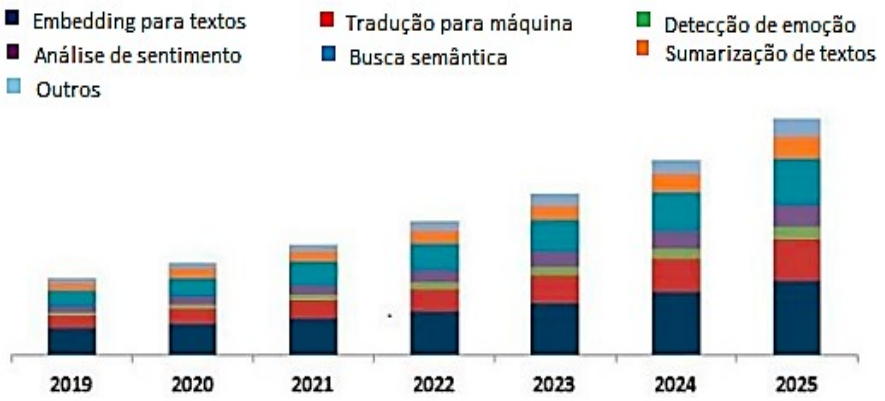
A crescente demanda de classificação de informação, para esses conjuntos de dados enormes, exige sistemas eficientes de categorização de produtos. Os desafios linguísticos específicos do português, assim como da natureza concisa e curta das descrições de produtos, que oferecem pouco contexto para análise, dificultam o desenvolvimento de metodologias eficazes (Alsmadi; Gan, 2019; Song *et. al*., 2014). A imprecisão na classificação pode levar a uma experiência frustrante para os usuários de serviços de e-commerce, resultando em dificuldades na busca por produtos (Alsmadi; Gan, 2019).

A categorização imprecisa prejudica a eficiência operacional das empresas, porque sem compreender como a informação se transforma em

percepção, conhecimento e ação dentro dos processos organizacionais, as empresas não conseguem apreciar a relevância das suas fontes de informação e das tecnologias associadas (Brandt; Vidotti, 2024). Associando-se a necessidade desta tomada de decisões baseadas em dados à grande quantidade de dados, emergem tecnologias de processamento de inteligência artificial para obtenção desses *insights*, da qual faz parte a interpretação de linguagem natural (Falcão *et al*., 2024).

A área de processamento de linguagem natural vem crescendo nos últimos anos para as mais distintas aplicações dentro do campo da ciência da informação, segundo Figura 1 (Abro; Talpur; Jumani, 2023). Em especial, tem-se que a aplicação para recuperação de informações de texto aparece como uma forte tendência entre as aplicações mais usuais.

Figura 1 - Avanço nas áreas de pesquisa de processamento de linguagem natural



Fonte: Adaptado de Abro, Talpur e Jumani (2023).

Descrição da Figura 1: Avanço nas áreas de pesquisa de processamento de linguagem natural.

Diante dessa necessidade, este estudo investiga a eficácia de algoritmos simples de recuperação da informação, combinados com técnicas de pré-processamento e otimização de parâmetros, na classificação de descrições curtas de produtos em português. A pesquisa tem como objetivo analisar o desempenho de diferentes métodos de vetorização e tokenização, incluindo a análise do

impacto da normalização. Além disso, o estudo pretende propor novas métricas de avaliação que considerem o desbalanceamento de classes, oferecendo uma visão mais abrangente do desempenho dos modelos e contribuindo para o desenvolvimento de sistemas de categorização de produtos mais eficientes em português.

2 Revisão da literatura

A extração de informações dos conjuntos de *big data* requer o emprego de uma extensa variedade de técnicas presentes na área de ciência de dados. Sobretudo, entre as abordagens em destaque, encontram-se aquelas relacionadas à *machine learning*. Esses métodos constituem uma maneira de ajudar computadores a aprender e analisar a *big data* de maneira mais eficaz e realizar previsões de problemas do mundo real (Wang *et. al*., 2018).

Esse procedimento é chamado de recuperação de informação. O processo é uma metodologia focada em obter informações relevantes para o usuário a partir de um banco de dados fornecido e avaliar seu desempenho usando determinadas métricas (Ceri *et. al.*, 2013). Como se trata de um processo intrinsecamente relacionado com previsão, tem-se que a recuperação de informação é um processo puramente estatístico desde sua concepção (Chen, 1995). Como já citado, atualmente dispõe-se de poder computacional de diferentes técnicas para realização deste procedimento.

Uma das funcionalidades do *machine learning* é realizar o procedimento de recuperação de informação de grande conjunto de dados que possuem linguagem humana em procedimentos denominados de processamento de linguagem natural (PLN) (Khurana *et. al.*, 2023). Para que os computadores possam realizar o processamento, é necessário que a linguagem humana a ser avaliada possua semântica, legibilidade, estrutura e sintaxe (Du *et. al.*, 2019).

A aplicação da tecnologia baseada em PLN pode ser encontrada em diversos contextos, desde sistemas de diálogo como os *chatbots* (Hirshberg;

Manning, 2015), até a análise de grandes volumes de texto para extração de informação. Este artigo, no entanto, foca em uma aplicação específica: a classificação de descrições curtas de produtos. O objetivo é explorar uma abordagem de agrupamento de descrições para gerar documentos que representem categorias de produtos. Após essa etapa, técnicas clássicas de recuperação da informação são aplicadas para classificar uma nova descrição de produto, associando-a à categoria mais similar.

Por definição, textos curtos são aqueles que possuem até 200 caracteres (Song *et. al.*, 2014). As aplicações de classificação de textos curtos envolvem: análise de mídias sociais, avaliação de *feedback* de clientes, recuperação de informações, análise de sentimentos, detecção de *spam* e modelagem de tópicos (Alsmadi; Gan, 2019). Como exemplo desta finalidade, pode-se citar a reaquisição de um livro a partir de uma série de dados sobre ele. Segundo a literatura, existem modelos de machine learning que podem recuperar o conteúdo de um livro ao serem alimentados com textos curtos contendo informações relevantes sobre o livro ou até mesmo a imagem da capa (Buczkowski; Sobkowicz; Kozlowski, 2018). Desta forma, é possível utilizar este conjunto de metodologias para obter *insights* de dados semiestruturados.

Formalmente, a classificação de um texto curto envolve atribuir uma categoria específica a um determinado texto (Kowsari *et. al.,* 2019). Como exemplo desta metodologia, tem-se a classificação de produtos em categorias a partir de uma descrição de produto (Prabhu *et. al.,* 2021). A obtenção do agrupamento final dos atributos do texto em categorias é obtida ao realizar as seguintes etapas: pré-processamento, seleção de atributos relevantes para compor categorias (*features*) e classificação (Wang *et. al*., 2018).

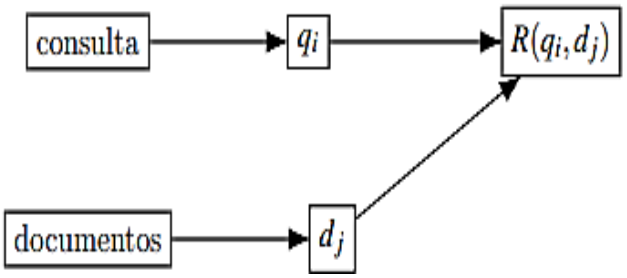
O processo de obtenção das classes para um grande conjunto de textos curtos pode ser difícil para humanos e técnicas de *machine learning* podem contribuir nesse processo, uma vez que estas podem aumentar a eficácia do processo (Wang *et. al*., 2018). Também, pode-se citar que existem diferentes métodos para realização de recuperação de informação. Dentre os mais comuns

estão: Bayes, árvore de decisão, vizinho k-próximo, regressão logística e classificador de vetor-suporte (SVC) (Jiang; Wang; Xiao, 2020). Mais recentemente, algoritmos de redes neurais têm aparecido em maior quantidade para realizar a mesma tarefa (Minaee *et. al*., 2021).

Alguns problemas surgem naturalmente ao tentar atribuir uma categoria específica a um texto curto. Entre os mais destacados estão sua brevidade, problemas relacionados à escrita e ao uso de linguagem informal (Song *et. al.*, 2014). Outro desafio conhecido para a implementação deste tipo de algoritmo é a própria língua humana, uma vez que modelos em linguagens populares como inglês são abundantes, enquanto para idiomas menos comuns são mais raros. Este desafio é notavelmente mais evidente no contexto da língua portuguesa, marcada por suas variantes e regionalismos (Song *et. al.*, 2014).

O protocolo de recuperação de informação de textos curtos segue o procedimento presente na Figura 2. Aqui, o termo d indica documento, uma consulta q e R o protocolo de consulta em si. Este modelo adaptado é amplamente utilizado em motores de busca e sistemas de gerenciamento de informações, onde a prioridade é encontrar a melhor correspondência possível para as consultas dos usuários dentro de um vasto conjunto de documentos.

Figura 2 - Fluxo de recuperação de informação para textos curtos



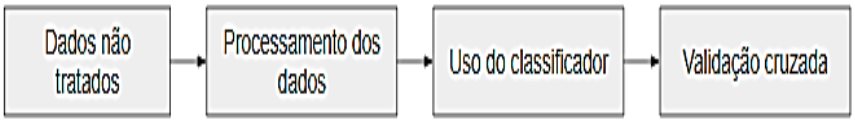
Fonte: Elaborado pelos autores.

Descrição da Figura 2: Fluxo de recuperação de informação para textos curtos.

O fluxograma presente na Figura 3 demonstra o encaminhamento necessário para geração de todo o fluxo que os dados são submetidos durante o

processo de recuperação de informação (Jiang; Wang; Xiao, 2020). De posse destes passos gerais, será explorado o que cada etapa faz no papel de recuperação de informação partindo do contexto apresentado.

Figura 3 - Procedimento para realização da classificação



Fonte: Adaptado de Jiang, Wang e Xiao, (2020).

Descrição da Figura 3: Procedimento para realização da classificação.

O primeiro passo de qualquer processo envolvendo classificação de textos curtos a partir de determinada descrição é realizar o pré-processamento dos dados. Existem algumas formas clássicas deste procedimento, estando entre elas a remoção de acentos e a conversão para minúsculas. Essas formas são procedimentos de preparação de palavras e que provocam alteração do processo de classificação (Eler *et. al*., 2018).

Os diferentes processos podem variar em termos de eficácia, sendo que esta etapa não é tão explorada quanto às demais na literatura (Naseem; Razzak; Eklund, 2021). Estudos em bases de dados similares também comprovam a linha geral de que a escolha de como a etapa de pré-processamento é realizada pode afetar o resultado final do modelo (Nafis; Awang, 2020).

Após esse tratamento prévio, as descrições de texto curto passam pelo processo denominado tokenização. Esse método corresponde a quebrar a descrição de um texto em pedaços menores chamados tokens, sendo que há diversas maneiras para realização deste procedimento (Rai; Borah, 2021).

A separação de unigramas apenas faz a separação de termo a termo, a de bigrama de dois termos agrupados e a 1-skip-2-grama realiza o agrupamento em grupos de dois, sempre pulando uma das palavras da descrição. Por fim, para

todos os casos a performance final do algoritmo também depende bastante da forma de tokenização escolhida (Rai; Borah, 2021).

Essa etapa de normalização é importante para garantir que a classificação não seja afetada por variações na escrita, como o uso de maiúsculas e minúsculas ou a presença de acentos. A tokenização por espaço, por sua vez, permite que o sistema processe cada palavra individualmente, facilitando a análise e a classificação do texto.

Para exemplificar o processo de pré-processamento e tokenização, considere a descrição de um produto: “Arroz tio joão 1 kg”. O processo se inicia com a conversão para minúsculas e remoção de acentos, logo a entrada se torna para “arroz tio joao 1 kg” Após o pré-processamento, a tokenização por espaço é realizada para dividir a descrição em tokens individuais: “arroz tio joao 1 kg” se torna [“arroz”, “tio”, “joao”, “1”, “kg”].

De posse das entradas tokenizadas, obtém-se o vocabulário de dado conjunto de dados expressos matematicamente como a quantidade de termos ser igual ao tamanho dos vetores (|w|). Tendo como outra entrada a quantidade de documentos, pode-se gerar a chamada matriz-documento através de um produto vetorial entre eles, como segue a equação na Equação 1 (Fig. 4).

Figura 4 - Equação 1**:** Obtenção da matriz termo-documento

$$A_{|V|\times|D|} = \begin{array}{c} \\ t_1 \\ t_2 \\ \vdots \\ t_i \end{array} \begin{array}{c} \begin{array}{cccc} d_1 & d_2 & \cdots & d_j \end{array} \\ \begin{bmatrix} t_{11} & t_{12} & \cdots & t_{1j} \\ t_{21} & t_{22} & \cdots & t_{2j} \\ \vdots & \vdots & \ddots & \vdots \\ t_{i1} & t_{i2} & \cdots & t_{ij} \end{bmatrix} \end{array}$$

Fonte: Elaborado pelos autores.

Descrição da Figura 4: Obtenção da matriz termo-documento.

A matriz termo-documento A da Equação 1, tem suas dimensões definidas pelo tamanho do vocabulário e pela quantidade de documentos na coleção. Se |V|

representa o número total de termos únicos no vocabulário e |D| denota o número de documentos ou classes na coleção, então a matriz A é dimensionada como |V| × |D|. Nesta matriz, cada linha representa um vetor no espaço de documentos, correspondendo a uma palavra específica do vocabulário. Cada elemento dessa linha indica a presença ou a frequência ou um peso dessa palavra em um documento específico. Analogamente, cada coluna de A pode ser interpretada como um vetor no espaço de palavras, representando um documento ou classe específica.

A sequência do processamento dos dados é dada pelo passo de transformação dos dados já separados em tokens em valores na forma de vetores para que possam ser analisados pelo computador. A modelagem adotada neste artigo será o de saco de palavras (*Bag of words*- BoW), um dos métodos mais simples para classificação de textos curtos (Deng *et. al.*, 2019). Para construção dos termos que representarão esse saco de palavras, pode-se usar abordagens binárias, de frequência de termos (TF) e de frequência de termos com o inverso de sua ocorrência (TF-IDF) (Yan *et. al.*, 2020).

A formulação binária apenas conta a presença ou ausência de determinado termo, enquanto a de frequência de termos conta a quantidade de vezes que determinado termo aparece (Yan *et. al*., 2020). A fim de tornar a medida de termos mais equilibrada, usa-se a frequência de termos com o inverso de palavras (TF-IDF) que tem por característica atenuar as aparições do termo com o inverso da quantidade de termos que ele aparece, baseando-se em uma formulação logarítmica (Jiang; Wang; Xiao, 2020; Kowsari *et. al*., 2019). É possível ainda aplicar uma normalização sobre a equação acima exposta para evitar potenciais distorções no resultado (Salton; Buckley, 1988). Os métodos podem ser sumarizados no Quadro 1.

Quadro 1 - Métodos de ponderações de termos

Método	Expressão matemática
Binário	$w_{d_i,t_j} = \begin{cases} 1, & \text{se presente} \\ 0, & \text{se ausente} \end{cases}$
TF	$w_{d_i,t_j} = tf$
TD-IDF	$w_{d_i,t_j} = tf * log\left(\frac{N}{df}\right)$

Fonte: Adaptado de Samant, Murthy e Malapati (2019).

Descrição do Quadro 1: Comparação entre métodos de *Bag of Words*.

Para exemplificar a formulação de TF-IDF, pode-se considerar o saco de palavras contendo as descrições de produto: “arroz tio joao 1 kg”, “arroz fumance parb 1kg”, “feijao carioca 1kg azulao”, “feijao 1 kg preto caldao”. Ao todo, o saco de palavras consitui-se de 13 palavras ao todo. Na Tabela 1, tem-se como ficaria a classificação para cada um dos itens, em que fij é a frequência dos termos, wij o peso de cada termo e w’ij o peso depois da normalização usando-se o módulo |wij|.

Tabela 1 - Aplicação do método TF-IDF

Código	Termo	IDF	[1]			[2]			[3]			[4]		
0	1	0,69	1	0,69	0,30	0	0	0	0	0	0	1	0,69	0,27
1	1kg	0,69	0	0	0	1	0,69	0,32	1	0,69	0,28	0	0	0
2	arroz	0,69	1	0,69	0,30	1	0,69	0,32	0	0	0	0	0	0
3	azulao	1,39	0	0	0	0	0	0	1	1,39	0,55	0	0	0
4	caldao	1,39	0	0	0	0	0	0	0	0	0	1	1,39	0,53
5	carioca	1,39	0	0	0	0	0	0	1	1,39	0,55	0	0	0
6	feij	1,39	0	0	0	0	0	0	0	0	0	1	1,39	0,53
7	feijao	1,39	0	0	0	0	0	0	1	1,39	0,55	0	0	0
8	fumace	1,39	0	0	0	1	1,39	0,63	0	0	0	0	0	0

	nce													
9	joao	1,39	1	1,39	0,60	0	0	0	0	0	0	0	0	0
10	kg	0,69	1	0,69	0,30	0	0	0	0	0	0	1	0,69	0,27
11	parb	1,39	0	1,39	0	1	1,39	0,63	0	0	0	0	0	0
12	preto	1,39	0	1,39	0	0	0	0	0	0	0	1	1,39	0,53
13	tio	1,39	1	1,39	0,60	0	0	0	0	0	0	0	0	0
$\|w_{ij}\|$			**2,30**			**2,19**			**2,50**					**2,59**

Fonte: Elaborado pelos autores**.**

Descrição da Tabela 1: Aplicação do método TF-IDF sobre as descrições “Arroz tio joão 1 kg”, “ARROZ FUMACENCE PARB 1KG”, “FEIJAO CARIOCA 1KG AZULAO”, e “Feij 1 kg Preto Caldão”.

A partir dos vetores da descrição e da categoria, pode-se realizar a comparação entre eles para configurar a classificação desta descrição. Existem diferentes formas para avaliar essa similaridade, como: distância euclidiana, coeficiente de Jaccard, correlação de Pearson e similaridade de cossenos (Deng *et. al.*, 2019). Aqui, é feito uso do método de similaridade de cossenos, visto que esta técnica apresenta maior eficácia, maior simplicidade e é normalmente utilizada para classificação de textos (Strehl; Ghosh; Mooney, 2000). A expressão matemática é visualizada a seguir na Equação 2 (Fig. 5).

Figura 5 - Equação 2**:** Cálculo da similaridade entre vetores

$$Similariade\ de\ Cosseno(u,v) = \frac{\sum_{i=1}^{n} u_i v_i}{\sqrt{\sum_{i=1}^{n} u_i^2}\sqrt{\sum_{i=1}^{n} v_i^2}}$$

Fonte: Adaptado de Deng *et. al.* (2019).

Descrição da Figura 5: Cálculo da similaridade entre vetores.

A similaridade de cosseno, presente na Equação 2, mede o ângulo entre dois vetores, sendo que quanto maior o valor obtido, maior será também a similaridade entre vetor e classificação. Assim sendo, o valor absoluto obtido

indica a probabilidade final de determinada vetor descrição estar contida em determinada categoria.

O processo de categorização também pode ser representado matricialmente sendo o vetor A matriz termo-documento e o vetor x o saco de palavras da descrição x. Por fim, as equações podem ser observadas no Quadro 2, tanto para a manipulação normalizada quanto para aquela não-normalizada.

Quadro 2 - Formulações normalizadas e não-normalizada do argmax

Método	**Expressão matemática**
Argmax	$Categoria = argmax(A^T x)$
Argmax normalizado	$Categoria = argmax(\frac{A^T}{A_.} x)$

Fonte: Elaborado pelos autores.

Descrição do Quadro 2: Formulações normalizadas e não-normalizada do argmax.

Para ilustrar o conceito de classificação por argmax, considere o mesmo vetor anterior como entrada: “Arroz tio joão 1 kg”. Utilizando técnicas de vetorização, essa descrição pode ser representada como um vetor numérico, por exemplo, [1, 0, 1, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 1]. Da mesma forma, as categorias “ARROZ” e “FEIJÃO” podem ser representadas por vetores numéricos, como [1, 1, 2, 0, 0, 0, 0, 0, 1, 1, 1, 1, 0, 1] e [1, 1, 0, 1, 1, 1, 1, 1, 0, 0, 1, 0, 1, 0], respectivamente. Neste exemplo, a descrição “Arroz tio joão 1 kg” apresenta uma similaridade maior com a categoria “ARROZ”, sugerindo que ela pertence a essa categoria.

Existem diferentes métricas para avaliar a performance de um algoritmo de *machine learning*. As medidas clássicas como acurácia e precisão são tidas como maneiras de medir a capacidade de predição daquele algoritmo (Wang *et. al*., 2018). Contrariamente à acurácia, o F1-Score macro é representado pela média harmônica da precisão e da revogação (Neu; Lahann; Fettke, 2022).

A acurácia pode ser definida como a quantidade percentual de amostras classificadas corretamente (Neu; Lahann; Fettke, 2022). Para duas categorias, tem-se que a acurácia pode ser matematicamente simplificada pela Equação 3 (Fig. 6). Nesta, as categorizações corretas estão no numerador da expressão, enquanto o total de exemplo disponíveis para classificação é o denominador.

Figura 6 - Equação 3**:** Métricas de desempenho: acurácia para duas categorias

$$\text{Acurácia} = \frac{\text{Verdadeiros Positivos} + \text{Verdadeiros Negativos}}{\text{Total de Exemplos}}$$

Fonte: Elaborado pelos autores.

Descrição da Figura 6: Métricas de desempenho: acurácia para duas categorias.

A acurácia, no contexto da classificação multiclasse, pode ser expressa matematicamente como a razão entre o número de exemplos corretamente classificados em cada classe e o número total de exemplos disponíveis. Essa métrica, representada pela Equação 3 (Fig. 7), mede o desempenho geral do modelo, considerando todas as classes e seu número de elementos.

Figura 7 - Equação 3: Métricas de desempenho: Acurácia multiclasse

$$ACC = \frac{\sum_{i=1}^{C} TP_{ii}}{\sum_{i=1}^{C} \sum_{j=1}^{C} C_{ij}}$$

Fonte: Elaborado pelos autores.

Descrição da Figura 7: Métricas de desempenho: Acurácia multiclasse.

A acurácia é, portanto, uma medida de avaliação global para verificação de adequação ou não de determinada descrição dentro de certa categoria. Essa métrica pode possuir problemas de desbalanceamento quando existem classes

dominantes, ou seja, com mais elementos do que outras (Takakashi *et. al.*, 2021). Para contornar esse problema, usa-se a métrica F1-Score macro.

O F1-Score Macro é uma métrica que fornece uma avaliação equilibrada do desempenho do modelo em cenários de classificação multiclasse, especialmente em situações onde as classes estão desbalanceadas. Calculado como a média harmônica das médias simples de precisão e revocação para cada classe (Takakashi *et. al*., 2021), o F1-Score Macro atribui pesos iguais a todas as classes, evitando que classes dominantes com um grande número de exemplos influenciam desproporcionalmente o resultado.

A Equação 4 (Fig. 8) demonstra matematicamente o método de cálculo do F1-Score macro.

Figura 8 - Equação 4: Métricas de desempenho: F1-Score Macro

$$\text{F1-Score} = 2 \times \frac{\text{Precisão} \times \text{Revocação}}{\text{Precisão} + \text{Revocação}}$$

Fonte: Elaborado pelos autores.

Descrição da Figura 8: Métricas de desempenho: F1-Score Macro.

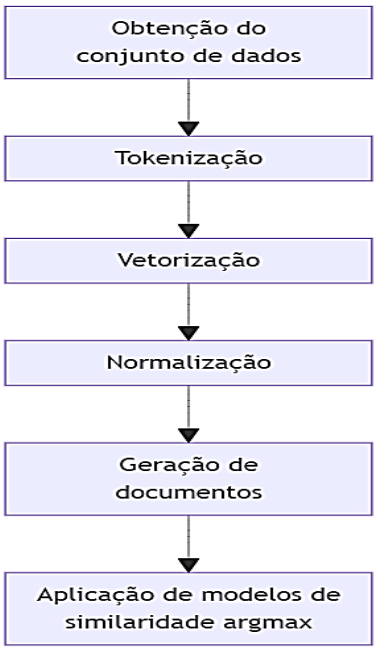
As métricas acima citadas são utilizadas especialmente para problemas de classificação (Neu; Lahann; Fettke, 2022). Ressalta-se que os resultados de algoritmos de *machine learning*, para serem implementados, também devem ser avaliados em termos de interpretabilidade, estabilidade, robustez e facilidade de uso (Wang *et. al*., 2018).

Para este artigo, começou-se escolhendo o tipo de pré-processamento a ser usado, passando pela transformação destes em vetores numéricos, aplicando os métodos previamente citados para finalmente avaliar a eficácia em termos de acurácia e F1-Score macro.

3 Metodologia

Este estudo utilizou uma abordagem quantitativa com um design experimental para avaliar o desempenho de um novo método de classificação de descrições curtas de produtos em português, focado em contribuir para o desenvolvimento de sistemas mais eficientes de categorização de produtos. O conjunto de dados Retail Product Description-Ptbr, composto por 250.365 descrições de produtos e suas respectivas categorias, coletadas de 18 dos maiores varejistas no Brasil (Darú *et al*., 2022), foi utilizado como base para a análise. O procedimento experimental pode ser sumarizado, conforme demonstra Figura 9.

Figura 9 - Fluxo de trabalho para este artigo



Fonte: Elaborado pelos autores.

Descrição da Figura 9: Fluxo de trabalho para este artigo.

As descrições dos produtos foram normalizadas, com a remoção de acentos e conversão para minúsculas. A tokenização por espaço foi então realizada para dividir a descrição em palavras individuais. Por exemplo, a descrição “Arroz tio joão 1.kg” se torna [“arroz”, “tio”, “joao”, “1”, “kg”] após o pré-processamento e a tokenização, como já apresentado anteriormente.

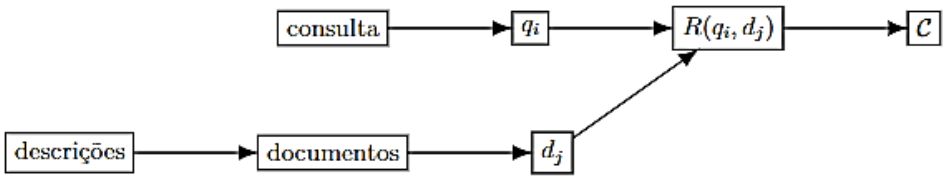
Os tokens foram convertidos em representações vetoriais utilizando os métodos bag-of-words, TF e TF-IDF (Samant; Murthy; Malapti, 2019). Esses métodos permitem representar o texto como um vetor numérico, onde cada elemento corresponde à frequência de um determinado termo na descrição.

É proposto um novo método de classificação, combinando a técnica tradicional de Recuperação da Informação (RI) com uma etapa inovadora de agrupamento. As descrições de produto são agrupadas de acordo com a mesma classe, criando um único documento representativo para cada categoria.

No método proposto, cada categoria de produto, formada pelo agrupamento de descrições semelhantes, é representada por um único documento d. Uma nova descrição a ser classificada é a consulta q. A função R calcula a similaridade entre a consulta q e cada documento d do conjunto D. Após a aplicação da função de recuperação R, o argmax é utilizado para selecionar o documento d com maior similaridade com a consulta q. Esse documento representa uma classe a qual é atribuída a descrição "consultada".

Um exemplo de aplicação pode ser visto ao estudar descrições como "Arroz branco 1kg", "Arroz integral 1kg" e "Arroz parbolizado 1kg" são agrupadas em um único documento representativo da categoria "Arroz". Ao analisar uma nova descrição como "Arroz branco 1kg", a função de recuperação R calcula a similaridade da descrição com cada documento de categoria, e a metodologia argmax, após analisar a similaridade, escolhe o documento "Arroz" como o mais similar, classificando a descrição como pertencente à esta categoria. Um resumo desse processo é visualizado na Figura 10.

Figura 10 - Fluxo de recuperação de informação



Fonte: Elaborado pelos autores.
Descrição da Figura 10: Fluxo de recuperação de informação.

Ao todo foram avaliadas 12 diferentes combinações, conforme apresentado no Quadro 3.

Quadro 3 - Sumarização das combinações aplicadas sobre os dados

ID	Tokenização	Tipo	Normalização
ARGMAX1BI	Unigrama	Binário	Nenhuma
ARGMAXX1BINORM	Unigrama	Binário	Normalizado
ARGMAX1TF	Unigrama	TF	Nenhuma
ARGMAX1TFNORM	Unigrama	TF	Normalizado
ARGMAX1TFIDF	Unigrama	TFIDF	Nenhuma
ARGMAX1TFIDFNORM	Unigrama	TFIDF	Normalizado
ARGMAX2BI	Bigrama	Binário	Nenhuma
ARGMAXX2BINORM	Bigrama	Binário	Normalizado
ARGMAX2TF	Bigrama	TF	Nenhuma
ARGMAX2TFNORM	Bigrama	TF	Normalizado
ARGMAX2TFIDF	Bigrama	TFIDF	Nenhuma
ARGMAX2TFIDFNORM	Bigrama	TFIDF	Normalizado

Fonte: Elaborado pelos autores.

Descrição do Quadro 3: Sumarização das combinações aplicadas sobre os dados.

O desempenho dos modelos foi avaliado utilizando a acurácia e o F1-Score Macro (Takakashi *et. al*., 2021). A acurácia mede a proporção de exemplos corretamente classificados, enquanto o F1-Score Macro avalia o desempenho do modelo em relação a todas as classes, considerando o desbalanceamento.

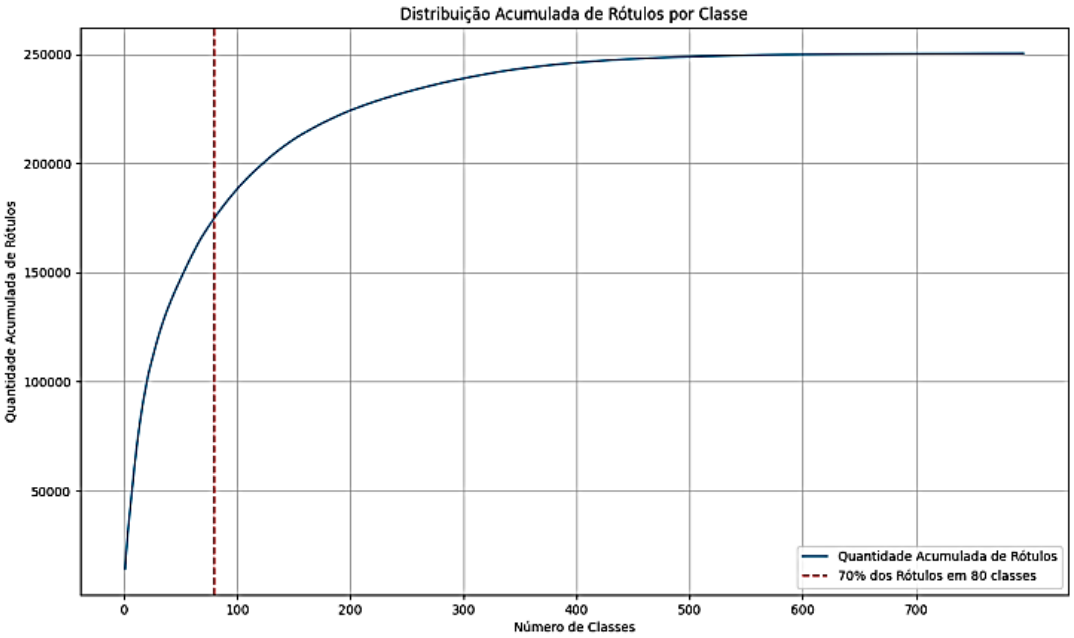
4. Resultados e discussões

Este estudo avaliou o desempenho de um novo método de classificação de descrições curtas de produtos em português, baseado em técnicas de recuperação

da informação e agrupamento de descrições. A análise dos resultados, com base no conjunto de dados Retail Product Description-Ptbr, revelou padrões e *insights* importantes para o desenvolvimento de sistemas de categorização.

A análise exploratória dos dados indicou um desbalanceamento significativo entre as categorias de produtos. A Figura 11 mostra que 70% das descrições de produtos estão concentradas em apenas 80 classes de produtos. Esse desbalanceamento é um desafio comum em tarefas de classificação de texto e pode impactar negativamente o desempenho dos modelos.

Figura 11 - Distribuição das classes

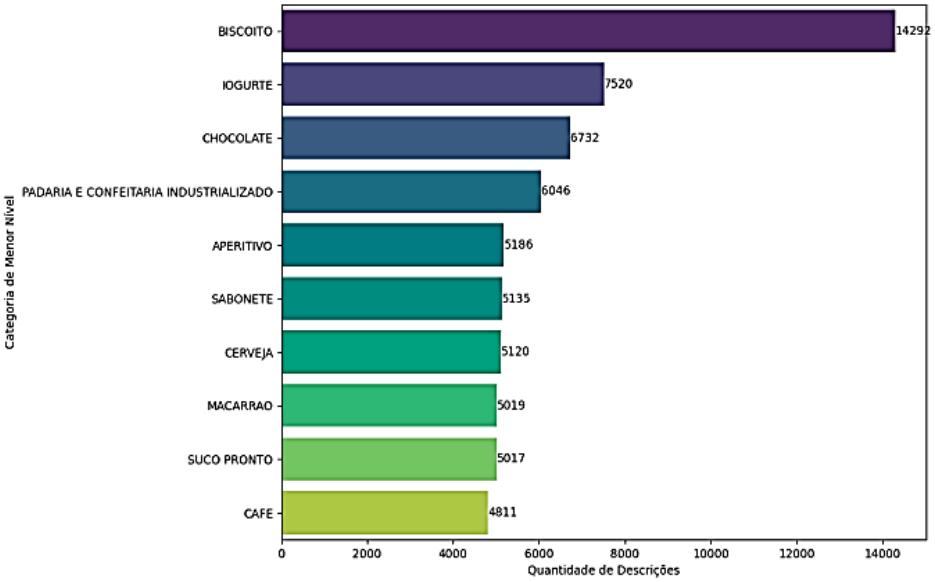


Fonte: Elaborado pelos autores.

Descrição da Figura 11: Distribuição das classes.

A Figura 12 ilustra a concentração de descrições em algumas categorias específicas, com a classe “BISCOITO” representando 7% do total de amostras. Essa concentração reforça a necessidade de métodos de classificação robustos que possam lidar com o desbalanceamento de classes.

Figura 12 - Principais classes baseada na quantidade de descrições

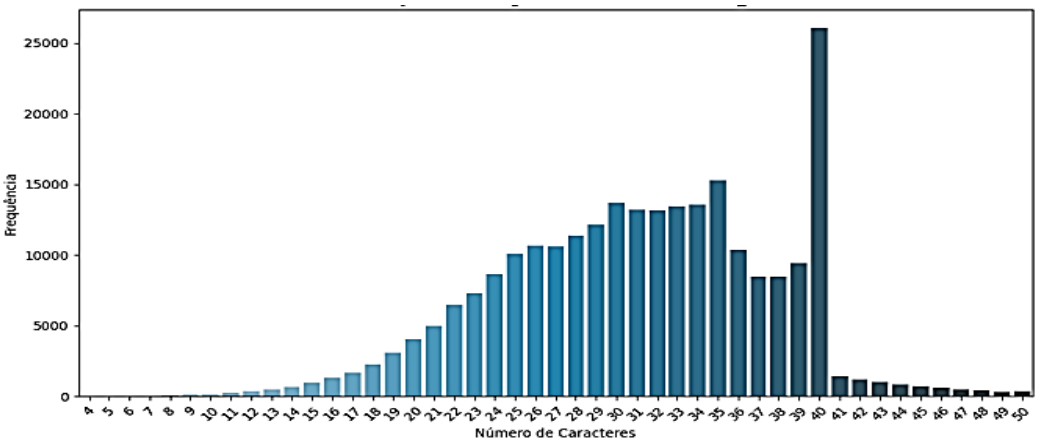


Fonte: Elaborado pelos autores.

Descrição da Figura 12: Principais classes baseada na quantidade de descrições.

A Figura 13 apresenta a distribuição de caracteres por descrição, mostrando que a maioria das descrições possui entre 20 e 40 caracteres. Esse padrão indica que as descrições são projetadas para se encaixar em cupons fiscais, refletindo a importância da concisão e da clareza nas descrições de produtos.

Figura 13 - Frequência de palavras com número de caracteres



Fonte: Elaborado pelos autores.

Descrição da Figura 13: Frequência de palavras com número de caracteres.

Os resultados da classificação de descrições de produtos, avaliados utilizando as métricas de acurácia e F1-Score Macro, demonstraram o potencial do método proposto para lidar com o desbalanceamento de classes. A Tabela 2 resume as principais estatísticas dos testes realizados.

Tabela 2 - Sumarização do método binário

Método	NGrama	Normalização	Acurácia			F1 Score Macro		
			Média	CV	#	Média	CV	#
Binário	[1,1]	Nenhuma	75,31	0,44	6	58,77	3,15	3
Binário	[1,1]	L2	19,07	2,40	12	25,39	2,04	11
Binário	[1,2]	Nenhuma	89,56	0,21	1	70,09	1,91	1
Binário	[1,2]	L2	31,30	1,48	11	33,82	1,77	8
TF	[1,1]	Nenhuma	63,76	0,41	10	23,44	2,05	12
TF	[1,1]	L2	77,05	0,24	4	52,31	2,46	5
TF	[1,2]	Nenhuma	66,68	0,28	9	27,08	1,69	10
TF	[1,2]	L2	79,65	0,29	3	55,56	2,61	4
TFIDF	[1,1]	Nenhuma	70,64	0,28	8	32,10	2,25	9
TFIDF	[1,1]	L2	78,37	0,32	5	55,20	2,47	6
TFIDF	[1,2]	Nenhuma	74,57	0,33	7	37,67	1,53	7
TFIDF	[1,2]	L2	82,76	0,29	2	59,83	2,52	2

Fonte: Elaborado pelos autores.

Descrição do Tabela 2: A tabela resume as médias e o coeficiente de variação da acurácia e do F1 Score Macro, além da posição de cada configuração no ranking de desempenho.

A configuração binária [1, 2] sem normalização (bigrama) apresentou o melhor desempenho em relação à acurácia, alcançando uma média de 89.56%. O método binário com a aplicação de normalização L2 e unigrama, por outro lado, obteve o pior desempenho, com uma média de acurácia de apenas 19.07%. O

método binário[1, 2] sem normalização (bigrama) também obteve o melhor desempenho em relação ao F1-Score Macro, com uma média de 70.09%.

A análise do método binário, utilizando o método argmax, revelou um desempenho significativamente melhor quando aplicado sem a normalização L2 e com a inclusão de bigramas. A combinação de bigramas e a ausência de normalização resultou na maior média de acurácia (89.56%) e F1-Score Macro (70.09%). Por outro lado, a configuração com normalização L2 e unigramas apresentou os piores resultados, com as menores médias de acurácia (19.07%) e F1-Score Macro (25.39%). Ambos os dados são visualizados na Tabela 3.

Tabela 3 - Sumarização do método binário

Método	NGrama	Normalização	Acurácia			F1 Score Macro		
			Média	CV	#	Média	CV	#
Binário	[1,1]	Nenhuma	75,31	0,44	2	58,77	3,15	2
Binário	[1,1]	L2	19,07	2,40	4	25,39	2,04	4
Binário	[1,2]	Nenhuma	89,56	0,21	1	70,09	1,91	1
Binário	[1,2]	L2	31,30	1,48	3	33,82	1,77	3

Fonte: Elaborado pelos autores.

Descrição do Tabela 3: A tabela resume as médias e o coeficiente de variação da acurácia e do F1 Score Macro, além da posição de cada configuração do método binário ranking de desempenho.

Essa diferença no desempenho pode ser explicada pela capacidade dos bigramas em capturar contextos relevantes, enquanto a normalização L2, em um cenário binário, pode penalizar classes com uma maior diversidade de termos. A normalização L2 tende a reduzir o peso dos termos.

A análise do método Term Frequency (TF), utilizando o método argmax, demonstrou que a configuração com bigramas e normalização L2 apresentou o melhor desempenho, alcançando as maiores médias de acurácia (79.65%) e F1-Score Macro (55.56%). Ambos os valores estão resumidos na Tabela 4.

Tabela 4 - Sumarização do método TF

Método	NGrama	Normalização	Acurácia			F1 Score Macro		
			Média	CV	#	Média	CV	#
TF	[1,1]	Nenhuma	63,76	0,41	4	23,44	2,05	4
TF	[1,1]	L2	77,05	0,24	2	52,31	2,46	2
TF	[1,2]	Nenhuma	66,68	0,28	3	27,08	1,69	3
TF	[1,2]	L2	79,65	0,29	1	55,56	2,61	1

Fonte: Elaborado pelos autores.

Descrição do Tabela 4: A tabela resume as médias e o coeficiente de variação da acurácia e do F1 Score Macro, além da posição de cada configuração do método TF no ranking de desempenho.

Essa configuração sugere que, para o método TF, a combinação de bigramas e normalização L2 é a mais eficiente para a classificação de descrições curtas de produtos. Os bigramas permitem capturar relações contextuais mais complexas, enquanto a normalização L2 ajuda a reduzir a influência de termos frequentes, evitando que termos comuns distorçam o processo de classificação.

A análise do método TF-IDF, utilizando o método Argmax, demonstrou que a configuração com bigramas e normalização L2 obteve os melhores resultados, atingindo as maiores médias de acurácia (82.76%) e F1-Score Macro (59.83%). Essa configuração indica que, para o método TF-IDF, a combinação de bigramas e normalização L2 é ideal para a classificação de descrições curtas de produtos. Os dados estão sumarizados na Tabela 5.

Tabela 5 - Sumarização do método TF-IDF

Método	NGrama	Normalização	Acurácia			F1 Score Macro		
			Média	CV	#	Média	CV	#
TFIDF	[1,1]	Nenhuma	70,64	0,28	4	32,10	2,25	4
TFIDF	[1,1]	L2	78,37	0,32	2	55,20	2,47	3
TFIDF	[1,2]	Nenhuma	74,57	0,33	3	37,67	1,53	2
TFIDF	[1,2]	L2	82,76	0,29	1	59,83	2,52	1

Fonte: Elaborado pelos autores.

Descrição do Tabela 5: A tabela resume as médias e o coeficiente de variação da acurácia e do F1 Score Macro, além da posição de cada configuração do método TF-IDF no ranking de desempenho.

A normalização L2, ao penalizar termos excessivamente frequentes, ajuda a equilibrar a influência dos termos nas descrições, permitindo que a classificação capture nuances linguísticas mais sutis presentes em textos curtos. A inclusão de bigramas, por sua vez, amplia a análise contextual, considerando as relações entre palavras adjacentes, o que contribui para um desempenho mais preciso.

Os resultados demonstraram que os melhores desempenhos foram observados com a utilização de bigramas. No método binário, a ausência de normalização L2 foi benéfica, enquanto nos métodos TF e TF-IDF, a aplicação da normalização L2 mostrou-se essencial.

5 Conclusões

Este estudo investigou a eficácia de um novo método de classificação de descrições curtas de produtos em português, combinando técnicas de recuperação da informação e agrupamento de descrições. A análise demonstrou o potencial desse método para melhorar a precisão e a eficiência da categorização de produtos, especialmente em cenários com desbalanceamento de classes.

Os resultados confirmaram a hipótese de que algoritmos simples de recuperação da informação, quando combinados com técnicas adequadas de pré-processamento e otimização de parâmetros, podem alcançar níveis de desempenho elevados. A pesquisa evidenciou a importância da normalização L2, do uso de bigramas e da consideração do desbalanceamento de classes na classificação de textos curtos.

As principais contribuições deste estudo incluem a análise do desempenho de diferentes algoritmos de recuperação da informação em português, a proposição de um novo método de classificação baseado em agrupamento e a demonstração da importância de técnicas de pré-processamento e otimização para aumentar a precisão e a eficiência. A pesquisa também destacou a necessidade de considerar o desbalanceamento de classes no desenvolvimento de sistemas de classificação de texto.

Apesar dos resultados promissores, este estudo apresenta algumas limitações. A avaliação foi limitada a descrições curtas de produtos e a um conjunto específico de técnicas. Pesquisas futuras podem expandir o escopo desta investigação, aplicando os métodos avaliados a diferentes tipos de textos, explorando novas técnicas de recuperação da informação e machine learning, e investigando o impacto do desbalanceamento de classes em diferentes cenários. A implementação de modelos mais avançados, como redes neurais profundas, também pode ser explorada para comparar o desempenho com os métodos mais simples analisados neste estudo.

Este estudo fornece uma base sólida para o desenvolvimento de sistemas de classificação de texto mais eficientes e precisos, contribuindo para o avanço do processamento de linguagem natural e das aplicações de recuperação da informação.

Referências

ABRO, A. A.; TALPUR, S. H.; JUMANI, A. K. A. Natural language processing challenges and issues: a literature review. **Gazi University Journal of Science**,

Istanbul, v. 36, n. 4, p. 1522-1536, 2023. Disponível em: https://doi.org/10.35378/gujs.1032517. Acesso em: 1 jul. 2024.

ALSMADI, I.; GAN, K. Review of short-text classification. **International Journal of Web Information Systems**, Leeds, v. 15, n. 2, p. 155-182, 2019. Disponível em: https://doi.org/10.1108/IJWIS-12-2017-0083. Acesso em: 1 jul. 2024.

BRANDT M.; VIDOTTI, S. Arquitetura da informação para processamento de negócio e modelagem de banco de dados: aproximações possíveis. **Em Questão**, Porto Alegre, v. 30, p. 1-22, 2024. Disponível em: https://doi.org/10.1590/1808-5245.30.131304. Acesso em: 1 jul. 2024.

BUCZKOWSKI, P.; SOBKOWICZ, A.; KOZLOWSKI, M. Deep learning approaches towards book covers classification. *In*: INTERNATIONAL CONFERENCE ON PATTERN RECOGNITION APPLICATIONS AND METHODS, 7., 2018, Madeira, Portugal. **Proceedings** […]. Setúbal: SciTePress, 2018. p. 309-316. Disponível em: https://doi.org/10.5220/0006556103090316. Acesso em: 1 jul. 2024.

CERI, S.; BOZZON, A.; BRAMBILLA, M.; VALLE, E.; FRATERNALLI, P. QUARTERONI, S. An introduction to information retrieval. *In*: **Web Information Retrieval. Data-Centric Systems and Applications**. Berlin: Springer: 2013. Disponível em: https://doi.org/10.1007/978-3-642-39314-3_1. Acesso em: 4 jul. 2024.

CHEN, H. Machine learning for information retrieval: neural networks, symbolic learning, and genetic algorithms. **Journal of American Society for Information Science**, New Jersey, v. 46, n. 3, p. 194-216, 1995. Disponível em: https://doi.org/10.1002/(SICI)1097-4571(199504)46:3<194::AID-ASI4>3.0.CO;2-S. Acesso em: 1 jul. 2024.

DARU, G.; MOTTA, F.; CASTELO, A.; LOCH, G. Short text classification applied to item description: some methods evaluation. **Semina: Ciências Exatas e Tecnológicas**, Londrina, v. 43, n. 2, p. 186-198, 2022. Disponível em: https://doi.org/10.5433/1679-0375.2022v43n2p189. Acesso em: 4 jul. 2024.

DENG, X.; LI, Y.; WENG, J.; ZHANG, J. Feature selection for text classification: a review. **Multimedia Tools and Applications**, New York, v. 78, p. 3797-3816, 2019. Disponível em: https://doi.org/10.1007/s11042-018-6083-5. Acesso em: 1 jul. 2024.

DU, J.; RONG, J.; MICHALSKA, S.; WANG, H.; ZHANG, Y. Feature selection for helpfulness prediction of online products reviews: an empirical

study. **Plos One**, San Francisco, p. 1-26, 23 Dec. 2019. Disponível em: https://doi.org/10.1371/journal.pone.0226902. Acesso em: 4 jul. 2024.

ELER, D.; GROSA, D.; POLA, I.; GARCIA, R.; CORREIA, R.; TEIXEIRA, J. Analysis of document pre-processing effects in text and opinion mining. **Information**, Basileia, v. 9, n. 4, p. 100, 2018. Disponível em: https://doi.org/10.3390/info9040100. Acesso em: 1 jul. 2024.

ERSHOV, A. What is information science? A Lesson for the Teacher **Soviet Education,** London, v. 28, n. 10-11, p. 51-54, 1986. Disponível em: https://doi.org/10.2753/RES1060-939328101151. Acesso em: 1 jul. 2024.

FALCÃO, L.; LOPES, B.; SOUZA, R.; BARBOSA, R. Uso de deep learning para a construção de um modelo de recuperação da informação aplicado ao sistema de mineração no Brasil. **Em Questão**, Porto Alegre, v. 30, p. 1-30, 2024. Disponível em: https://doi.org/10.1590/1808-5245.30.135550. Acesso em: 1 jul. 2024.

GANDOMI, A.; HAIDER, M. Beyond the hype: Big data conccepts, methods, and analytics. **International Journal of Information Management**, Amsterdam, v. 35, n. 2, p. 137-144, 2015. Disponível em: https://doi.org/10.1016/j.ijinfomgt.2014.10.007. Acesso em: 1 jul. 2024.

HIRSCHBERG, J.; MANNING, C. Advances in natural language processing. **Science**, Washington, v. 349, n. 6245, p. 261-266, 2015. Disponível em: https://doi.org/10.1126/science.aaa8685. Acesso em: 1 jul. 2024.

ISHIMARU, K. "Memory" for Sustainable Society. *In*: INTERNATIONAL WORKSHOP ON JUNCTION TECHNOLOGY, 20., 2021, Kyoto. **Proceedings** […]. New York: IEEE, 2021. Disponível em: https://doi.org/10.23919/IWJT52818.2021.9609367. Acesso em: 1 jul. 2024.

JIANG, H.; WANG, W.; XIAO, Y. Explaining a bag of words with hierarchical conceptual labels. **World Wide Web**, New York, v. 23, p. 1693-1713, 2020. Disponível em: https://doi.org/10.1007/s11280-019-00752-3. Acesso em: 1 jul. 2024.

KHURANA, D.; KOLI, A.; KHATTER, K.; SINGH, S. Natural language processing state of the art, current trends and challenges. **Multimedia Tools and Applications**, New York, v. 82, n. 3, p. 3713-3744, 2023. Disponível em: https://doi.org/10.1007/s11042-022-13428-4. Acesso em: 1 jul. 2024.

KOWSARI, K.; MEIMANDI, K.; HEIDARSAFA, M.; MENDU, S.; BARNES, L.; BROWN, D. Text classification algorithms: a survey. **Information**, Basileia,

v. 10, n. 4, p. 150, 2019. Disponível em: https://doi.org/10.3390/info10040150. Acesso em: 1 jul. 2024.

MARCHIONINI, G. Information science roles in the emerging field of data science. **Journal of Data and Information Science**, Boston, v. 1, n. 2, p. 1-6, 2017. Disponível em: https://doi.org/10.20309/jdis.201609. Acesso em: 1 jul. 2024.

MINAEE, S.; KALCHBRENNER, N.; CAMBRIA, E.; NIKZAD, N.; CHENAGHLU, M.; GAO, J. Deep learning-based text classification: a comprehensive review. **ACM Computing Surveys (CSUR)**, New York, v. 54, n. 3, p. 1-40, 2021. Disponível em: https://doi.org/10.1145/3439726. Acesso em: 1 jul. 2024.

NAFIS, N.; AWANG, S. The impact of pre-processing and feature selection on text classification. *In*: ZAKARIA, Z., AHMAD, R. (ed.). **Advances in Eletronics Engineering**, New York: Springer, 2020. p. 269-280.

NASEEM, U.; RAZZAK, I.; EKLUND, P. A survey of pre-processing techniques to improve short-text quality: a case study on hate speech detection on Twitter. **Multimedia Tools and Applications**, New York, v. 80, p. 35239-35266, 2021. Disponível em: https://doi.org/10.1007/s11042-020-10082-6. Acesso em: 1 jul. 2024.

NEU, D.; LAHANN, J.; FETTKE, P. A systematic literature review on state-of-the-art deep learning methods for process prediction. **Artificial Intelligence Review**, New York, v. 55, p. 801-827, 2022. Disponível em: https://doi.org/10.1007/s10462-021-09960-8. Acesso em: 1 jul. 2024.

PRABHU, Y.; KANNAN, A.; AGASTYA, A.; GOGINENI, M.; VARMA, M. Extreme Text Classification. *In*: ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 24., 2021, London. **Proceedings** […]. New York: ACM, 2021. p. 1322-1330.

RAI, A.; BORAH, S. Study of various methods for Tokenization. *In*: MANDAL, J.; MUKHOPADHYAY, S.; ROY, A. (ed.) **Applications of Internet of Things**. Singapore: Springer, 2021. p. 193-200. Disponível em: https://doi.org/10.1007/978-981-15-6198-6_18. Acesso em 1 jul. 2024.

SALTON, G; BUCKLEY, C. Term-weighting approaches in automatic text retrieval. **Information Processing & Management**, Amsterdam, v. 24, n. 5, p. 513-523, 1988. Disponível em: https://doi.org/10.1016/0306-4573(88)90021-0. Acesso em: 1 jul. 2024.

SAMANT, S.; MURTHY, M. V. R.; MALAPATI, A. Comparison of term weighting schemes for text classification. **International Journal of Information Technology and Computer Science**, Hong Kong, v. 11, n. 8, p. 43-50, 2019. Disponível em: https://doi.org/10.5815/ijitcs.2019.08.06. Acesso em: 1 jul. 2024.

SARACEVIC, T. Information Science. **Journal of American Society for Information Science**, New Jersey, v. 50, n. 12, p. 1051-1063, 1999. Disponível em: https://doi.org/10.1002/(SICI)1097-4571(1999)50:12<1051::AID-ASI2>3.0.CO;2-Z. Acesso em: 1 jul. 2024.

SONG, G.; YE, Y.; DU, X.; HANG, X.; BIE, S. Short text classification: a survey. **Journal of Multimedia**, Oulu, v. 9, n. 5, p. 635-643, 2014. Disponível em: https://doi.org/10.4304/jmm.9.5.635-643. Acesso em: 1 jul. 2024.

STREHL, A; GHOSH, J.; MOONEY, R. Impact of similarity on web-page clustering. *In*: WORKSHOP ON ARTIFICIAL INTELLIGENCE FOR WEB SEARCH, July 2000, Boston. **Proceedings** […]. Washington: Association for the Advancement of Artificial Intelligence, 2000.

TAKAKASHI, K.; YAMAMOTO, K.; KUCHIBA, A.; KOYAMA, T. Confidence interval for micro-averaged F1 and macro-averaged F1 scores. **Applied Intelligence**, New York, v. 52, p. 4961-4972, 2022. Disponível em: https://doi.org/10.1007/s10489-021-02635-5. Acesso em: 1 jul. 2024.

WANG, Y. ZHOU, Z.; JIN, S.; LU, M. Comparisons and selections of features and classifiers for short text classification. **IOP Science**, Bristol, v. 261, p. 12018, 2018. Disponível em: https://doi.org/10.1088/1757-899X/261/1/012018. Acesso em: 1 jul. 2024.

YAN, D.; LI, K.; GU, S.; YANG, L. Network-based bag-of-words model for text classification. **IEEE Access**, New York, v. 8, p. 82641-82652, 2020. Disponível em: https://doi.org/ 10.1109/ACCESS.2020.2991074. Acesso em: 1 jul. 2024.

Product description classification in portuguese: performance assessment of machine learning algorithms, preprocessing and attribute extraction

Abstract: The growing demand for automated product classification systems in e-commerce platforms has fueled the search for efficient solutions for product categorization, particularly in Portuguese. This study investigates the adaptation of classical information retrieval techniques, such as bag-of-words, TF, and TF-

IDF, for the task of classifying short product descriptions. The research evaluates different preprocessing and tokenization strategies, including analyzing normalization impact. The results show that simple information retrieval methods, when combined with appropriate preprocessing and parameter optimization, can achieve significantly superior performance.

Keywords: machine learning; natural language processing; text classification; product description; short text; bag of words; term frequency; inverse document frequency

Recebido: 15/03/2024

Aceito: 24/06/2024

Declaração de autoria

Concepção e elaboração do estudo: Gilsiley Henrique Darú, Gustavo Valentim Loch

Coleta de dados: Gilsiley Henrique Darú

Análise e interpretação de dados: Gilsiley Henrique Darú

Redação: Gilsiley Henrique Darú, Daniel Felipe Pietezak

Revisão crítica do manuscrito: Gilsiley Henrique Darú, Gustavo Valentim Loch

Como citar

DARU, Gilsiley Henrique; LOCH, Gustavo Valentim; PIETEZAK, Daniel Felipe. Aprimorando a classificação de descrições de produtos em português com a utilização de técnicas da recuperação de informação: uma abordagem de agrupamento de descrições. **Em Questão**, Porto Alegre, v. 30, e-139205, 2024. DOI: https://doi.org/10.1590/1808-5245.30.139205

Parecer(es) aberto(s):

https://doi.org/10.1590/1808-5245.30.139205A

https://doi.org/10.1590/1808-5245.30.139205B

Gilsiley Henrique Daru, Gustavo Valentim Loch, Daniel Felipe Pietezak

Aprimorando a classificação de descrições de produtos em português com a utilização de técnicas da recuperação de informação: uma abordagem de agrupamento de descrições
Product description classification in portuguese: performance assessment of machine learning algorithms, preprocessing and attribute extraction

Em Questão
vol. 30, e-139205, 2024
Universidade Federal do Rio Grande do Sul,
ISSN: 1807-8893
ISSN-E: 1808-5245

DOI: https://doi.org/10.1590/1808-5245.30.139205