Exploring Supervised Learning Models for Multi-Label Text Classification in Brazilian Restaurant Reviews

José A. de Almeida Neto[1] **, Tiago de Melo**[1]

[1]Escola Superior de Tecnologia – Universidade do Estado do Amazonas (UEA)
Manaus – AM – Brasil

{jadan.eng20,tmelo}@uea.edu.br

Abstract. *This paper investigates the use of Natural Language Processing (NLP) methods for classifying user reviews on Brazilian restaurants, exploring various pre-processing techniques to enhance supervised learning models. Among the models evaluated, the combination of Logistic Regression (LR) with the pre-processing technique of stemming proved to be most effective, achieving a micro F1-Score value of 0.89, notably in multi-label text classification. When applied to a real dataset, the model proved to be useful in identifying subtle differences in customer opinions, even within units of the same restaurant franchise.*

Resumo. *Este artigo investiga o uso de métodos de Processamento de Linguagem Natural (NLP) para classificação de comentários de clientes sobre restaurantes brasileiros, explorando diversas técnicas de pré-processamento para aprimorar modelos de aprendizado supervisionado. Entre os modelos avaliados, a combinação da Regressão Logística (LR) com a técnica de pré-processamento stemming se mostrou mais eficaz, alcançando um valor de micro F1-Score de 0,89, com destaque na classificação de texto multirrótulo. Quando aplicado a um conjunto de dados reais, o modelo conseguiu ser útil na identificação de diferenças sutis nas opiniões dos clientes, até mesmo dentro de unidades de uma mesma franquia de restaurantes.*

1. Introdução

A crescente popularidade das redes sociais e plataformas de avaliações confere maior relevância aos comentários de clientes acerca de serviços e produtos. No setor gastronômico, observa-se essa tendência, na qual tais comentários desempenham um papel crucial para o sucesso dos restaurantes [Kumar et al. 2020]. Conforme estudos recentes [Li et al. 2021, Wang et al. 2021], a visibilidade e a lucratividade nesse setor estão intrinsecamente vinculadas à reputação online do restaurante, refletida nos *feedbacks* e avaliações de seus clientes.

A identificação dos atributos mais mencionados pelos clientes torna-se um elemento de grande relevância, pois oferece aos proprietários dos estabelecimentos a possibilidade de analisar e identificar padrões, permitindo tomadas de decisão baseadas em tais informações com o objetivo de aprimorar produtos e serviços. No entanto, a análise desses dados impõe certos desafios, dentre os quais se destacam: o grande volume de dados, a presença de ambiguidade e subjetividade nos comentários (derivadas da natureza opiniativa dos mesmos) e a variação linguística, já que a expressão na internet não segue

um padrão ou formato pré-estabelecido. Além disso, um único comentário pode abordar múltiplos temas, conforme exemplificado na Figura 1. Diante deste cenário, torna-se necessário o desenvolvimento de tecnologias e estratégias de processamento de linguagem natural (*Natural Language Processing - NLP*) capazes de processar e analisar esses dados de forma eficiente.

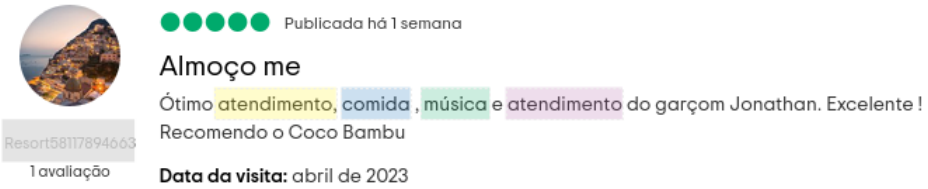

Figura 1. Exemplo de comentário postado em plataforma de avaliação.

O presente estudo busca investigar a aplicação de métodos de NLP na classificação de comentários de clientes sobre restaurantes brasileiros, além de explorar técnicas diversas que visam aprimorar o desempenho dos modelos de aprendizagem supervisionada. Para tal, foi coletado e organizado um vasto conjunto de dados contendo milhares de comentários acerca de restaurantes brasileiros. A pesquisa foca na construção de um modelo eficaz na classificação de texto multirrótulo com análise de opinião e com o objetivo de responder às seguintes perguntas de pesquisa (PP):

PP1: É possível desenvolver um modelo eficaz de aprendizagem de máquina para identificar os temas abordados em comentários sobre restaurantes no Brasil?

PP2: Quais técnicas de pré-processamento mais contribuem para o desenvolvimento do modelo eficiente abordado na PP1?

PP3: Quais seriam os resultados ao aplicar o modelo de aprendizagem mais eficiente identificado na PP1 a um extenso conjunto de dados?

Para responder à PP1, foram avaliados três modelos clássicos de aprendizagem supervisionada: *Support Vector Machine* (SVM), *Random Forest* (RF) e *Logistic Regression* (LR). Além desses, também foi empregado o AutoGluon, um dos métodos de AutoML frequentemente citados na literatura [de Oliveira and de Melo 2021, Blohm et al. 2020].

A resposta à PP2 envolveu a aplicação dos modelos mencionados com a técnica de otimização de hiperparâmetros de busca exaustiva (*grid search*), o método de seleção de características (*SelectPercentile*) e a técnica de vetorização de texto TF-IDF (*Term Frequency-Inverse Document Frequency*). Subsequentemente, cada modelo foi executado com diversas técnicas de pré-processamento, gerando diferentes resultados que foram compilados e discutidos. Entre todos os modelos investigados, o modelo *Logistic Regression* (LR), quando combinado com a técnica de pré-processamento *stemming* e com a seleção de 100% de suas características, apresentou o melhor desempenho, alcançando um macro F1-Score de 0.81 e um micro F1-Score de 0.89.

Para responder à PP3, o modelo LR foi aplicado a um grande conjunto de comentários coletados sobre uma importante cadeia de restaurantes. Como resultado deste estudo de caso, observou-se que, mesmo pertencendo à mesma franquia, diferentes unidades são avaliadas de maneiras distintas pelos clientes. Por exemplo, notou-se que a maioria das unidades do Coco Bambu considera o serviço como o tema mais relevante.

No entanto, em algumas unidades, os clientes apontam a comida como o atributo mais relevante. Esse fato sinaliza a necessidade de uma análise mais detalhada para entender por que essas unidades apresentam essa particularidade, o que pode ser de grande interesse para os proprietários de restaurantes.

O restante do artigo está organizado da seguinte maneira: a Seção 2 descreve os principais trabalhos relacionados; a Seção 3 apresenta o conjunto de dados e os modelos e técnicas empregadas nos experimentos; a Seção 4 discute os resultados experimentais; e finalmente, a Seção 5 traz as conclusões e sugestões para trabalhos futuros.

2. Trabalhos Relacionados

Nesta seção, são citados trabalhos anteriores relacionados à classificação de texto multirrótulo com análise de opinião ou que utilizaram modelos e técnicas que foram aplicadas neste trabalho. Estudos similares que também tratam de análise de opinião relacionada a restaurantes serviram como base para a definição dos temas [Yu and Zhang 2020, de Melo 2021, He et al. 2019]. Al-Bakiri et al. [AL-Bakri et al. 2021] realizaram um estudo que explora classificação de texto multirrótulo com restaurantes para classificar restaurantes em rótulos predefinidos com base em seus recursos, utilizando o modelo de aprendizado de máquina *Support Vector Machine* (SVM). Este modelo foi utilizado neste trabalho.

Usualmente, modelos baseados em redes neurais são aplicados em problemas de análise de sentimentos [Catelli et al. 2022, Gandhi et al. 2021], todavia, também é comum a utilização de modelos clássicos de aprendizagem supervisionada, como *Random Forest* (RF), *Logistic Regression* (LR), assim como o *Support Vector Machine* (SVM) [Singh and Tripathi 2021]. Estes modelos foram utilizados neste trabalho. Priyadarshini, I. e Cotton, C [Priyadarshini and Cotton 2021] reuniram diversos modelos baseados em redes neurais para avaliação de um novo modelo para análise de sentimentos que combina *Long Short-Term Memory* (LSTM) e redes neurais convolucionais (CNN), criado a partir da aplicação da técnica de otimização de hiperparâmetros de busca exaustiva (*grid search*), que também é utilizada neste trabalho.

Outra técnica que visa melhorar o desempenho de modelos multirrótulos é a seleção de atributos. Spolaôr et al. [Spolaôr et al. 2014] realizaram um estudo que demonstrava uma melhora no desempenho em aprendizagem multirrótulo utilizando a seleção de atributos. Este método pode ser especialmente útil neste estudo, já que lida com um grande número de comentários de clientes e a seleção efetiva de atributos pode ajudar a reduzir a dimensionalidade dos dados, tornando o modelo mais eficiente. Outra técnica com o mesmo objetivo é a vetorização de texto, o TF-IDF (*Term Frequency-Inverse Document Frequency*). Diversos estudos [Das and Chakraborty 2018, Domeniconi et al. 2016] utilizam esta técnica para aplicações que envolvem classificação de texto com análise de sentimentos. TF-IDF é especialmente relevante neste contexto, pois atribui pesos aos termos com base na sua importância em um documento em relação a todo o conjunto de dados (*corpus*). Isso permite que termos que são particularmente significativos em um comentário tenham um impacto maior na classificação, o que é crucial na análise de sentimentos. Todas estas técnicas foram empregadas e avaliadas na análise da performance dos classificadores.

Adicionalmente dos modelos clássicos de aprendizagem de máquina supervisionada, o aprendizado de máquina automatizado (AutoML) é uma abordagem alternativa para este tipo de classficação [Wever et al. 2021]. Cysneiros Aragão et. al. [Cysneiros Aragão et al. 2023] realizaram um estudo que reuniu diversas ferramentas de AutoML para aplicar em problema de classificação binária, multiclasse e multirrótulo com diferentes conjunto de dados (*datasets*). Com base nisso, após o levantamento de diversos trabalhos relacionados, este estudo investigou uma alternativa que reúne modelos e técnicas citadas.

3. Materiais e Métodos

3.1. Coleta de Dados

Neste estudo, implementou-se um *crawler* em Python para coletar comentários postados em plataformas populares de avaliação *online* como Facebook, Instagram, Google Review e TripAdvisor. Foram coletados 4.000 comentários de restaurantes, referentes ao período de 2000 a 2022. A escolha deste período deve-se à ampla adoção dessas plataformas durante esses anos.

3.2. Tema das Opiniões

Este estudo considerou que as sentenças dos usuários estão relacionadas a um conjunto pré-definido de temas. A definição dos temas teve como base outros estudos [Yu and Zhang 2020, de Melo 2021, He et al. 2019]. Cada sentença foi assinalada aos seguintes temas: ambiente, bebida, comida, geral, localização, preço, serviço e outros. Sentenças relacionadas à decoração, interior, música, espaço, iluminação ou lotação foram identificadas como ambiente. Sentenças relacionadas à cerveja, vinho ou *drinks* foram identificadas como bebida. Sentenças relacionadas à alimentação, sobremesa, cardápio ou pratos foram identificadas como comida. Sentenças relacionadas a transporte ou o acesso ao restaurante foram identificadas como localização. Sentenças relacionadas ao custo do serviço ou dos produtos dos restaurantes foram identificadas como preço. Sentenças relacionadas às atitudes dos atendentes, presteza, interação com o gerente ou atraso foram identificadas como serviço. Sentenças relacionadas ao restaurante em si, tais como "excelente restaurante", foram identificadas como geral. Finalmente, sentenças que não se enquadravam em nenhum dos temas anteriores foram classificadas como outros.

A Tabela 1 apresenta as quantidades de ocorrências de cada tema no conjunto de dados. Observa-se que o tema de comida é o mais frequentemente comentado, com 2.091 ocorrências, seguido pelo tema de serviço, com 1.406 ocorrências, e ambiente, com 1.257 ocorrências. A partir destes dados, é possível ainda observar que o tema outros, que representa as sentenças que não puderam ser identificadas em um dos temas previamente definidos, possui apenas 1,4% do número total de ocorrências. Isto evidencia que os temas pré-definidos possuem uma alta cobertura na classificação das sentenças.

3.3. Tarefa de Identificação de Temas

Este estudo foi conduzido a partir de um conjunto de dados que foi manualmente anotado pelos autores. O processo de anotação seguiu os seguintes passos: inicialmente, um dos autores anotou cada sentença conforme as definições de temas descritas na Subseção 3.2. Em seguida, o segundo autor revisou as anotações. Divergências foram discutidas até que um consenso fosse alcançado entre os autores.

Tabela 1. Tabela de ocorrências de cada tema.

Tema	Quantidade de ocorrências
comida	2.091
serviço	1.406
ambiente	1.257
geral	1.193
preço	784
bebida	404
localização	201
outros	106
total	**7.442**

A tarefa de identificação de temas é uma tarefa de classificação multirrótulo, o que significa que cada sentença pode ser relacionada a mais de um tema simultaneamente. Os temas foram indicados por 0 ou 1, demonstrando a ausência ou presença respectivamente de cada tema em cada sentença. O principal objetivo da tarefa é identificar a ocorrência de temas nos comentários dos usuários sobre os restaurantes, conforme definido a seguir.

Seja $C = \{c_1, \ldots, c_n\}$ um conjunto de comentários sobre um restaurante $\mathcal{R}$, onde cada $c \in C$ é formado por um conjunto de sentenças $S = \{s_1, \ldots, s_m\}$, onde cada sentença $s \in S$ pode estar associada ao conjunto de temas $T = \{$*comida*, *serviço*, *ambiente*, *geral*, *preço*, *bebida*, *localização*, *outros*$\}$. A tarefa de identificação de temas consiste em uma função $f : S \rightarrow 2^T$, onde cada sentença $s_i \in S$ é identificada como associada a um ou mais temas de T.

A tarefa deste estudo foi modelada como um problema de classificação multirrótulos porque é possível ocorrer a presença de mais de um tema em uma única sentença. Foi adotado o método de relevância binária de [Zhang and Zhou 2013] para classificação de multirrótulos, que transforma o problema de multirrótulos em múltiplos problemas binários separados e independentes.

Mais formalmente, a estratégia proposta pode ser descrita como a aplicação de um conjunto de classificadores binários, onde cada classificador $f_{T_j} : s_i \rightarrow \{0, 1\}$ é associado a um tema específico $T_j \in T$. No treinamento de cada f_{T_j}, as sentenças referentes a T_j são consideradas exemplos positivos, enquanto todas as outras sentenças são tratadas como exemplos negativos. Após o treinamento dos classificadores, a função f é definida como:

$$f(s_i) = \{T_j \in T \mid f_{T_j}(s_i) = 1\} \tag{1}$$

Isto significa que cada sentença s_i é identificada para cada tema T_j, onde f_{T_j} aplica-se para s_i positivas.

A Tabela 2 mostra como a sentença "*Ótimo atendimento, comida, música e atendimento do garçom Jonathan.*" da Figura 1 seria anotada. Neste exemplo, o termo atendimento seria classificado com o tema serviço, o termo comida seria classificado com o tema de igual nome e o termo música seria classificado com o tema ambiente. Ressalta-se que apesar do termo "*serviço*" não aparecer explicitamente na sentença, a menção a "*atendimento*" refere-se à atitude do atendente, sendo assim classificado sob o tema "*serviço*".

Tabela 2. Exemplo de anotação.

comida	serviço	ambiente	geral	preço	bebida	localização	outros
1	1	1	0	0	0	0	0

Após a anotação do conjunto de dados, verificou-se que, das 4.000 sentenças anotadas, 54,25% possuíam múltiplos temas, enquanto 45,75% estavam associadas a um único tema. O conjunto de dados anotados foi disponibilizado para uso da comunidade científica[1].

3.4. Modelos Utilizados

Foram escolhidos três modelos clássicos de aprendizagem supervisionada: *Support Vector Machine* (SVM), *Random Forest* (RF) e *Logistic Regression* (LR). Adotou-se a técnica *One-vs-Rest*, que consiste em dividir a classificação multiclasse em vários problemas de classificação binária, distinguindo uma classe específica do conjunto das demais classes. Também foi utilizada uma ferramenta de AutoML. Neste estudo, optou-se pelo AutoGluon[2] porque é um dos métodos de AutoML bastante empregado recentemente [de Oliveira and de Melo 2021, Blohm et al. 2020] e que tem apresentado bons resultados. Para a construção destes modelos, foi utilizada a biblioteca de aprendizado de máquina Scikit-Learn[3] em Python.

3.5. Técnicas de pré-processamento

As técnicas de pré-processamento são fundamentais para a construção de modelos de classificação de texto, pois diversos estudos [Kadhim 2018, Srividhya and Anitha 2010] apontam que as aplicações dessas técnicas estão diretamente ligadas com a obtenção de melhores resultados. Neste estudo, aplicou-se a técnica de vetorização de texto TF-IDF nos experimentos com os modelos clássicos, visando a uma representação mais efetiva dos textso mais efetiva com vetores numéricos relacionados com a frequência das palavras nos comentários. Adicionalmente, foram experimentadas diversas técnicas de pré-processamento com o objetivo de investigar o impacto dessas técnicas no resultado final do método. Cada técnica foi aplicada individualmente e os seus respectivos resultados foram apresentados e discutidos na Seção 4.

As técnicas *lower* e *strip* são comumente aplicadas em problemas de classificação de texto. A primeira consiste em converter todas as letras em letras minúsculas, enquanto a segunda consiste em remover os espaços extras das sentenças. Além disso, outra técnica aplicada foi a remoção de todas as pontuações das sentenças, através do módulo *regular expressions (re)*[4] do Python. Adicionalmente, foi aplicada a técnica de *stopwords* da biblioteca *NLTK*[5], muito utilizada em estudos [Sarica and Luo 2021, Ghag and Shah 2015] de classificação de texto, que consiste em remover os termos que não possuem relevância semântica devido a sua alta frequência. Também da mesma biblioteca, foram aplicadas técnicas de manipulação de palavras: *lemmatization*, que consiste em converter palavras

[1] http://tiagodemelo.info/datasets.html
[2] https://auto.gluon.ai
[3] https://scikit-learn.org/stable
[4] https://docs.python.org/3/library/re.html
[5] https://www.nltk.org/

para a sua palavra de origem e *stemming*, que consiste em converter palavras para o seu radical.

Outra técnica aplicada foi o *POS tagging* da biblioteca *spaCy*[6], em que somente os termos classificados como substantivos, verbos e adjetivos foram mantidos. Após isso, foram selecionados apenas os substantivos, depois, apenas os verbos e por fim, apenas os adjetivos. Essa delimitação de classes gramaticais pode ser útil para a análise de sentimentos [de Oliveira and de Melo 2021].

Todos os experimentos com os modelos clássicos foram executados com o método de seleção de características *SelectPercentile* da biblioteca Scikit-Learn, que visa selecionar as características mais relevantes no conjunto de dados. Neste estudo, quatro percentuais das características totais foram considerados: 40%, 60%, 80% e 100%. Estes valores foram escolhidos com base em estudos anteriores e/ou considerações práticas, e foram otimizados usando o método de busca exaustiva (*grid search*) da biblioteca Scikit-Learn. Para o AutoGluon, essas técnicas não foram aplicadas, pois este pacote já executa uma busca exaustiva por modelos mais eficientes através do componente *TabularPredictor*, que automaticamente realiza algumas etapas de pré-processamento, incluindo a seleção de características.

3.6. Métricas de Avaliação

Foram utilizadas as bastante conhecidas métricas de precisão (P), revocação (R) e F1-Score (F_1) para avaliar a identificação de temas em comentários sobre restaurantes. Seja *A* o conjunto de temas identificados corretamente, de acordo com um conjunto de referência, e seja *B* o conjunto de temas identificados pelo método classificador que está sendo avaliado. Precisão (P), revocação (R) e F1-Score (F_1) foram definidos como:

$$P = \frac{|A \cap B|}{|B|} \quad (2) \qquad R = \frac{|A \cap B|}{|A|} \quad (3) \qquad F_1 = \frac{2 \times (P \times R)}{(P + R)} \quad (4)$$

Micro F1-Score é calculada computando-se os valores globais de precisão e revocação para todas as classes e, em seguida, calculando-se a medida F1. Micro F1-Score considera igualmente relevante a classificação de cada sentença, independentemente da sua classe. Diferentemente, a métrica macro F1-Score considera igualmente importante a eficácia dos classificadores em cada classe, independentemente da quantidade de itens do conjunto. Desta forma, a análise dos classificadores através destas métricas fornecem avaliações complementares da efetividade de um classificador. Dado um conjunto de classes $|C|$, a macro F1-Score calcula a média de F1-Score de todas as classes, enquanto que a micro F1-Score considera a soma de precisões P e a soma das revocações R de todas as classes e calcula a média.

4. Resultados e Discussões

4.1. Análise dos Modelos

Para a construção de um modelo efetivo na classificação de texto multirrótulo com a análise de opinião, foram conduzidos experimentos utilizando-se dos seguintes modelos:

[6]https://spacy.io/

Support Vector Machine (SVM), *Logistic Regression* (LR), *Random Forest* (RF) e AutoGluon (AG). O SVM foi treinado com a seleção de 80% das características, enquanto o LR utilizou 100% das características e o RF, 40%. Estes valores foram obtidos através da aplicação do método de seleção de características. Além disso, uma segunda versão do AutoGluon foi utilizada, nomeada AG-BQ. Esta versão utiliza o parâmetro *best-quality*, que busca o melhor modelo possível independente do tempo de execução e do uso de disco. Adicionalmente, foram avaliadas as seguintes técnicas de pré-processamento:

- *LCase*: todos os caracteres das sentenças foram convertidos para minúsculo (*lowercase*).
- *RmSpac*: foram removidos todos os espaços extras das sentenças (*remove spaces*).
- *NoPunc*: foram removidas todas as pontuações das sentenças (*no punctuation*).
- *NoStp*: foram removidas todas as *stopwords* das sentenças (*no stopwords*).
- *Lemat*: todas as palavras das sentenças foram convertidas para a sua forma base (*lemmatization*).
- *Stem*: todas as palavras das sentenças foram reduzidas ao seu radical (*stemming*).
- *NSVFlt*: foram identificadas as classes gramaticais de cada palavra das sentenças e foram mantidas apenas as palavras classificadas como substantivos (N), verbos (V) e adjetivos (A) (*noun, verb, adjective filter*).
- *NounFlt*: foram identificadas as classes gramaticais de cada palavra das sentenças e foram mantidas apenas as palavras classificadas como substantivos (*noun filter*).
- *VerbFlt*: foram identificadas as classes gramaticais de cada palavra das sentenças e foram mantidas apenas as palavras classificadas como verbos (*verb filter*).
- *AdjFlt*: foram identificadas as classes gramaticais de cada palavra das sentenças e foram mantidas apenas as palavras classificadas como adjetivos (*adjective filter*).
- *StpStmSpa*: foram removidas todas as *stopwords* das sentenças, todas as palavras restantes foram reduzidas ao seu radical (*stemming*) e por fim, foram removidos todos os espaços extras das sentenças (*stopwords, stemming, space removal*).

A Tabela 3 apresenta o desempenho dos classificadores combinados com as diferentes técnicas de pré-processamento em termos de precisão (P), revocação (R) e F1-Score (F_1), considerando tanto as versões macro quanto micro de cada métrica. Os valores mais altos de cada técnica de pré-processamento para cada modelo estão marcados em itálico e o melhor resultado de cada métrica foi marcado em negrito. Neste experimento, cada resultado denota uma média de validação cruzada 5-*folds*.

Os resultados da Tabela 3 revelaram as seguintes indicações que contribuem para responder à segunda pergunta de pesquisa (PP2). A técnica de pré-processamento de texto *Stem* foi a que se apresentou como mais eficaz. A eficácia dessa técnica pode ser atribuída à melhoria da consistência, pois o *stemming* trata várias formas de uma palavra com uma única representação. Por exemplo, as palavras "gostar", "gosto" e "gostei" seriam reduzidas ao mesmo radical "gost". Isso ajuda a melhorar a eficiência do classificador ao reduzir o ruído e a variabilidade nos dados. Além disso, o *stemming* ajuda a aumentar a cobertura dos modelos de classificação ao permitir que eles tratem palavras semelhantes de maneira equivalente e isto pode ser útil em cenários onde os dados de treinamento são limitados ou desbalanceados, como é o caso do conjunto de dados utilizado nos experimentos.

Em relação aos classificadores, o classificador LR_{Stem} alcançou o melhor resultado nas métricas macro e micro de F1-Score. Isto responde à primeira pergunta de pesquisa

Tabela 3. Tabela de resultados.

Modelos	Macro			Micro		
	P	R	F_1	P	R	F_1
SVM	0,846	0,788	0,807	0,897	0,869	0,883
SVM_{LCase}	0,846	0,788	0,807	0,897	0,869	0,883
SVM_{RmSpac}	0,846	0,788	0,807	0,897	0,869	0,883
SVM_{NoPunc}	0,838	0,780	0,800	0,894	0,862	0,878
SVM_{NoStp}	*0,853*	0,793	*0,812*	*0,899*	0,869	0,884
SVM_{Lemat}	0,846	0,788	0,806	0,897	0,869	0,883
SVM_{Stem}	0,845	0,796	0,810	0,897	*0,877*	*0,887*
SVM_{NSVFlt}	0,800	0,779	0,785	0,869	0,846	0,857
$SVM_{NounFlt}$	0,733	0,733	0,728	0,835	0,797	0,815
$SVM_{VerbFlt}$	0,381	0,505	0,421	0,453	0,555	0,499
SVM_{AdjFlt}	0,462	0,551	0,481	0,531	0,615	0,570
$SVM_{StpStmSpa}$	0,850	*0,797*	*0,812*	0,897	0,876	0,886
LR	0,846	0,805	0,812	0,892	0,889	0,890
LR_{LCase}	0,842	0,805	0,811	0,892	0,888	0,890
LR_{RmSpac}	0,842	0,805	0,811	0,892	0,888	0,890
LR_{NoPunc}	0,844	0,796	0,805	0,889	0,880	0,884
LR_{NoStp}	0,845	0,806	0,809	0,887	0,892	0,889
LR_{Lemat}	0,847	0,804	0,813	0,892	0,886	0,889
LR_{Stem}	*0,855*	0,808	***0,815***	*0,894*	0,892	***0,893***
LR_{NSVFlt}	0,838	0,778	0,789	0,874	0,858	0,866
$LR_{NounFlt}$	0,812	0,699	0,742	0,869	0,778	0,821
$LR_{VerbFlt}$	0,502	0,353	0,406	0,586	0,464	0,518
LR_{AdjFlt}	0,603	0,386	0,461	0,689	0,485	0,569
$LR_{StpStmSpa}$	0,839	***0,811***	0,812	0,887	***0,895***	0,891
RF	0,901	0,749	0,790	0,907	0,831	0,867
RF_{LCase}	0,904	0,750	0,792	0,908	0,831	0,868
RF_{RmSpac}	0,902	0,749	0,790	0,908	0,833	0,869
RF_{NoPunc}	0,902	0,742	0,786	0,904	0,825	0,863
RF_{NoStp}	0,867	0,763	0,798	0,903	0,842	0,871
RF_{Lemat}	0,896	0,750	0,793	0,909	0,832	0,869
RF_{Stem}	0,894	0,759	0,795	***0,912***	0,845	0,877
RF_{NSVFlt}	***0,905***	0,752	0,791	***0,912***	0,840	0,875
$RF_{NounFlt}$	0,824	0,681	0,736	0,879	0,761	0,816
$RF_{VerbFlt}$	0,508	0,324	0,380	0,592	0,438	0,503
RF_{AdjFlt}	0,614	0,379	0,454	0,682	0,478	0,562
$RF_{StpStmSpa}$	0,883	*0,768*	*0,803*	0,908	*0,849*	*0,878*
AG-BQ	0,812	0,738	0,769	0,876	0,812	0,843
AG-BQ_{LCase}	0,804	0,721	0,756	0,874	0,819	0,845
AG-BQ_{RmSpac}	0,812	0,738	0,769	0,876	0,812	0,843
AG-BQ_{NoPunc}	0,813	0,735	0,768	0,874	0,802	0,836
AG-BQ_{NoStp}	0,811	0,740	0,771	0,874	0,810	0,841
AG-BQ_{Lemat}	0,813	0,744	0,772	0,877	0,812	0,843
AG-BQ_{Stem}	*0,820*	*0,765*	*0,788*	*0,882*	*0,829*	*0,855*
AG-BQ_{NSVFlt}	0,802	0,724	0,759	0,866	0,791	0,827
AG-$BQ_{NounFlt}$	0,785	0,639	0,700	0,849	0,723	0,781
AG-$BQ_{VerbFlt}$	0,417	0,357	0,254	0,423	0,579	0,476
AG-BQ_{AdjFlt}	0,619	0,317	0,313	0,526	0,486	0,492
AG-$BQ_{StpStmSpa}$	0,813	0,755	0,781	0,877	0,826	0,850
AG	0,808	0,723	0,760	0,874	0,810	0,840
AG_{LCase}	0,806	0,722	0,758	0,871	0,809	0,839
AG_{RmSpac}	0,808	0,723	0,760	0,874	0,810	0,840
AG_{NoPunc}	0,798	0,711	0,747	0,864	0,789	0,825
AG_{NoStp}	0,806	0,729	0,762	0,873	0,809	0,840
AG_{Lemat}	0,809	0,730	0,763	0,870	0,804	0,836
AG_{Stem}	0,814	0,744	0,773	0,876	0,823	0,849
AG_{NSVFlt}	0,799	0,713	0,750	0,862	0,786	0,822
$AG_{NounFlt}$	0,783	0,638	0,699	0,846	0,725	0,781
$AG_{VerbFlt}$	0,331	0,383	0,249	0,397	0,583	0,453
AG_{AdjFlt}	0,505	0,308	0,291	0,516	0,484	0,486
$AG_{StpStmSpa}$	*0,816*	*0,751*	*0,779*	*0,877*	*0,826*	*0,850*

(PP1). Nota-se que o classificador LR_{Stem} não possui os melhores resultados nas métricas de precisão e revocação, mas ainda sim possui o melhor macro e micro F1-Score. Isso se dá pelo fato de que este classificador obteve o melhor equilíbrio entre as métricas, assim sendo o modelo mais eficaz. Observa-se ainda que os classificadores alcançaram resultados de micro F1-Score superiores aos resultados alcançados pelos mesmos classificadores em termos de macro F1-Score. A métrica macro F1-Score dá o mesmo peso para cada sentença, independentemente do número de ocorrências. Já a métrica micro F1-Score calcula o resultado como média de todas as classes. Portanto, essa métrica é mais útil quando existe um desequilíbrio de classes, pois dá mais peso para as classes mais frequentes. No domínio do problema, entende-se que as classes mais frequentes são as mais relevantes, pois correspondem aos temas mais comentados pelos clientes dos restaurantes. De toda maneira, optou-se por também apresentar os resultados de macro F1-Score para avaliação mais completa do estudo.

Nota-se ainda que o método de AutoML (AG) apresentou resultados inferiores aos comparados com os modelos tradicionais de aprendizagem de máquina. Dentre as razões para esse resultado está o fato dos modelos de AutoML terem dificuldades com conjuntos de dados desbalanceados, pois a maioria dos algoritmos de AutoML são projetados para maximizar a precisão geral, o que pode levar a um desempenho mais baixo na classificação de classes com menor suporte. Um outro fator é que os sistemas de AutoML frequentemente fazem uma busca exaustiva ou parcial sobre diferentes modelos e hiperparâmetros. Esta busca pode ser limitada em termos de tempo de execução, especialmente em problemas de classificação de texto, o que pode resultar em um modelo subótimo.

Finalmente, nota-se também que, embora os classificadores utilizados nos experimentos fossem consideravelmente diferentes, há mínima diferença nos resultados dos classificadores quanto a diferentes técnicas de pré-processamento.

4.2. Análise de Erros

Para aprimorar o entendimento dos resultados exibidos na Tabela 3, os valores obtidos por meio da validação cruzada de 5-folds para o melhor modelo - o LR_{Stem} - são apresentados na Figura 2. Os gráficos representados em azul claro denotam o desempenho de acordo com a métrica macro, enquanto as caixas em amarelo claro indicam o desempenho com base na métrica micro.

Verifica-se que o modelo LR_{Stem} exibe uma variabilidade mais expressiva na métrica macro em comparação com a métrica micro. A variação mais acentuada da métrica macro em relação à métrica micro está tipicamente associada à distribuição de classes dentro dos conjuntos de dados. Assim, uma maior variação na métrica macro pode ser um indicativo de que o desempenho do modelo é mais instável em classes menos representadas. Este é um aspecto importante a ser considerado durante a avaliação e a otimização do modelo.

A Figura 3 apresenta os resultados da precisão macro para cada tema nos 5 *folds* do modelo LR_{Stem}. Observa-se que outros foi o tema onde ocorreu a maior variação. A instabilidade e variabilidade dos resultados deste tema é devido ao fato que todas as opiniões que não foram identificadas como algum outro tema, foram assinaladas como outros. Por exemplo, as sentenças "*é necessário muita atenção ao pagamento de conta em grupo maior*" (*sic*) e "*recomendo no dia meno movimentado*" foram assinaladas com

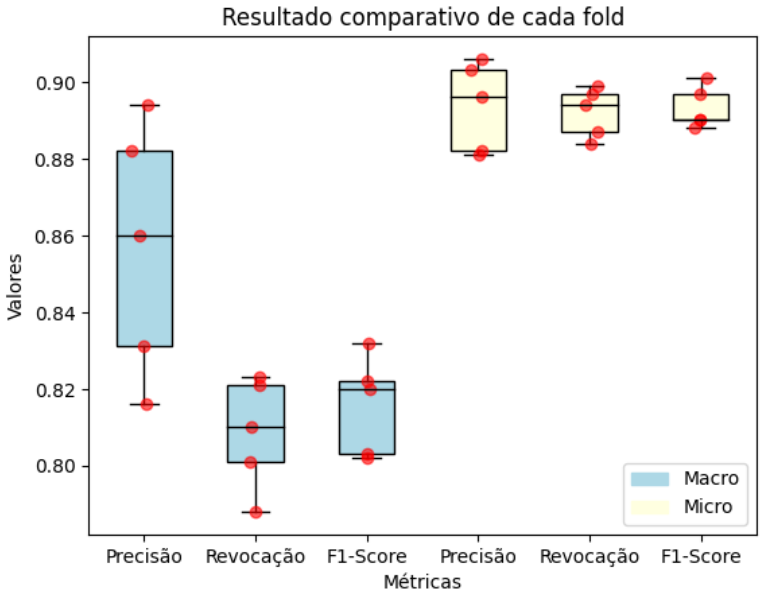

Figura 2. Resultado de cada *fold* da validação cruzada do método LR_{Stem} que apresentou o melhor resultado.

o tema outros. Consequentemente, a alta variabilidade de textos dificulta o aprendizado por parte dos classificadores. Com base nisso, é válido destacar que o tema outros é responsável por tamanha variação, visto que este possui 5 valores de precisão macro bem diferentes.

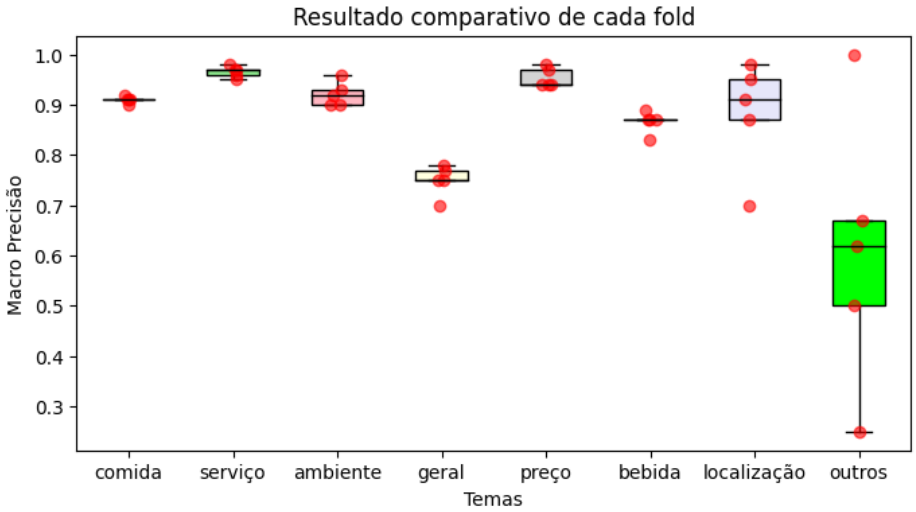

Figura 3. Resultado de cada *fold* da validação cruzada do método LR_{Stem} que apresentou o melhor resultado para a métrica precisão.

4.3. Estudo de Caso

Esta seção avalia a viabilidade da aplicação do método LR_{Stem} em um grande volume de dados, visando responder à pergunta de pesquisa PP3. Para tal, foram coletados comentários das unidades de restaurantes da rede Coco Bambu. A escolha pelo Coco Bambu

deve-se ao fato de a empresa possuir 47 unidades distribuídas por todas as regiões do Brasil, o que pode representar uma maior diversidade de tipos de comentários. Utilizou-se o Google Reviews como fonte de dados, de onde foram coletados 279.005 comentários publicados no período de 2010 a 2023.

A Figura 4 exibe a distribuição dos temas por unidade do Coco Bambu. Cada ponto no gráfico representa o valor percentual do tema em relação ao total de temas assinalados para o restaurante. Portanto, por existirem 47 unidades, há 47 pontos para cada tema no gráfico. A forma de cor cinza que envolve os pontos representa a densidade desses percentuais. O tema serviço é o mais comentado e também possui a maior amplitude, enquanto o tema outros é o menos representativo. Assim, fica evidenciado que, em geral, o tema mais comentado é o de serviço. No entanto, em algumas unidades do Coco Bambu, o tema comida torna-se o assunto mais relevante. É possível observar que unidades distintas possuem relevância maior para diferentes temas, como é o caso do tema ambiente, que atinge mais de 20% de ocorrências na unidade "Coco Bambu Coffee Break" de Fortaleza.

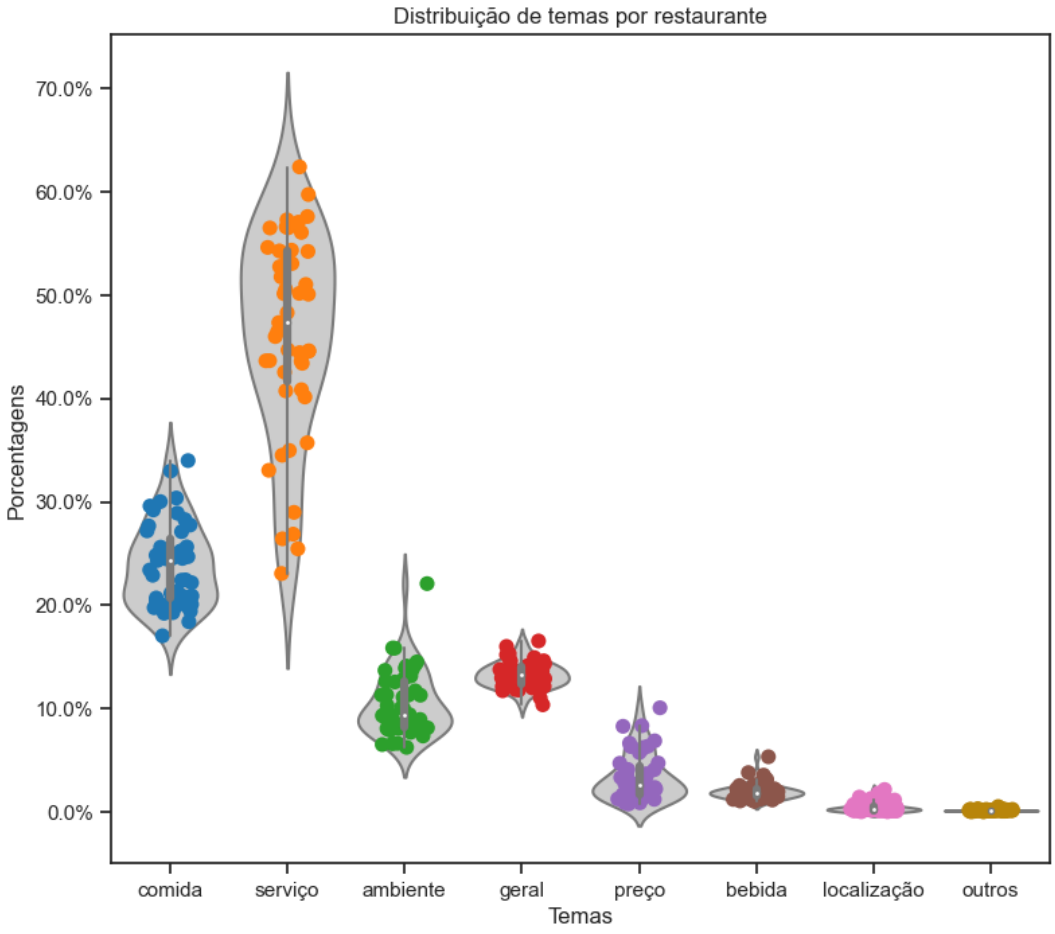

Figura 4. Distribuição dos temas por unidade de restaurante do Coco Bambu.

A Figura 5 representa a distribuição dos temas de outra forma. Cada coluna é uma unidade do Coco Bambu, e o percentual de cada tema em relação ao seu respectivo restaurante está marcado conforme a legenda apresentada. Isso possibilita a observação

do caso específico de cada restaurante e suas peculiaridades. Um exemplo é a unidade "Coco Bambu Beira Mar"de Fortaleza, que possui uma distribuição distinta de temas em relação às outras unidades. Nesta unidade, o tema comida é mais relevante que o de serviço, o que é algo incomum, e seus temas geral e ambiente possuem altos valores em relação aos demais.

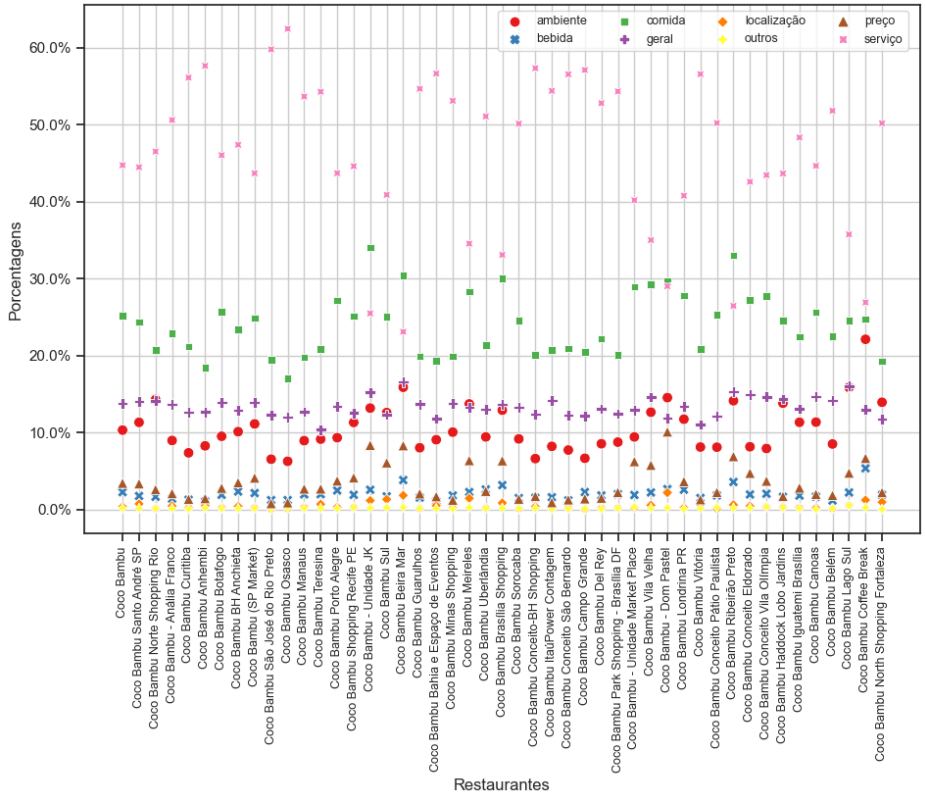

Figura 5. Outra distribuição dos temas por unidade de restaurante do Coco Bambu.

5. Conclusões e Trabalhos Futuros

Neste estudo, realizou-se uma investigação exploratória centrada em métodos de classificação de comentários sobre restaurantes brasileiros. Os resultados obtidos forneceram *insights* valiosos, permitindo a conclusão de que a combinação do modelo de *Logistic Regression* (LR) com a técnica de pré-processamento denominada *stemming* demonstrou ser o classificador mais eficaz.

Essa abordagem destacou-se, apresentando uma performance superior na tarefa de classificação, o que revela a sua potencialidade na análise de sentimentos em textos escritos em língua portuguesa. Adicionalmente, a implementação do modelo LR_{Stem} em um conjunto de dados reais proporcionou a identificação das variações nos comentários de diferentes unidades de restaurantes pertencentes a uma mesma franquia. Esse fato atesta a eficiência do modelo no reconhecimento de particularidades na expressão de opiniões dos consumidores, mesmo dentro de um contexto aparentemente homogêneo.

Para trabalhos futuros, considera-se fundamental aprofundar as investigações em duas direções principais. Primeiramente, propõe-se o estudo de estratégias de tratamento

do desbalanceamento dos dados. Esta é uma questão relevante, pois pode influenciar significativamente o desempenho dos modelos de aprendizado de máquina, levando a classificações enviesadas ou a um desempenho insatisfatório quando confrontado com classes minoritárias.

Adicionalmente, considera-se relevante estender a aplicação da metodologia proposta para outros domínios, avaliando a sua eficácia em contextos diversos e expandindo, assim, o seu potencial de aplicação. Esta abordagem permitirá uma validação mais robusta da metodologia e contribuirá para a identificação de possíveis ajustes necessários para melhorar ainda mais a eficiência do modelo na análise de sentimentos.

Referências

AL-Bakri, N. F., Al-zubidi, A. F., Alnajjar, A. B., and Qahtan, E. (2021). Multi label restaurant classification using support vector machine. *Periodicals of Engineering and Natural Sciences*, 9(2):774–783.

Blohm, M., Hanussek, M., and Kintz, M. (2020). Leveraging automated machine learning for text classification: Evaluation of automl tools and comparison with human performance. *arXiv preprint arXiv:2012.03575*.

Catelli, R., Pelosi, S., and Esposito, M. (2022). Lexicon-based vs. bert-based sentiment analysis: A comparative study in italian. *Electronics*, 11(3):374.

Cysneiros Aragão, M. V., Guimarães Afonso, A., Ferraz, R. C., Gonçalves Ferreira, R., and Gomes Leite, S. (2023). A practical evaluation of automl tools for binary, multiclass, and multilabel classification.

Das, B. and Chakraborty, S. (2018). An improved text sentiment classification model using tf-idf and next word negation. *arXiv preprint arXiv:1806.06407*.

de Melo, T. (2021). Análise de comentários das plataformas online de restaurante michelin no brasil. In *A produção do conhecimento nas ciências da comunicação*, pages 226–238.

de Oliveira, M. and de Melo, T. (2021). An empirical study of text features for identifying subjective sentences in portuguese. In *Intelligent Systems: 10th Brazilian Conference, BRACIS 2021, Virtual Event, November 29–December 3, 2021, Proceedings, Part II 10*, pages 374–388. Springer.

Domeniconi, G., Moro, G., Pasolini, R., and Sartori, C. (2016). A comparison of term weighting schemes for text classification and sentiment analysis with a supervised variant of tf. idf. In *Data Management Technologies and Applications: 4th International Conference, DATA 2015, Colmar, France, July 20-22, 2015, Revised Selected Papers 4*, pages 39–58. Springer.

Gandhi, U. D., Malarvizhi Kumar, P., Chandra Babu, G., and Karthick, G. (2021). Sentiment analysis on twitter data by using convolutional neural network (cnn) and long short term memory (lstm). *Wireless Personal Communications*, pages 1–10.

Ghag, K. V. and Shah, K. (2015). Comparative analysis of effect of stopwords removal on sentiment classification. In *2015 international conference on computer, communication and control (IC4)*, pages 1–6. IEEE.

He, J., Wang, C., Wu, H., Yan, L., and Lu, C. (2019). Multi-label chinese comments categorization: comparison of multi-label learning algorithms. *Journal of New Media*, 1(2):51.

Kadhim, A. I. (2018). An evaluation of preprocessing techniques for text classification. *International Journal of Computer Science and Information Security (IJCSIS)*, 16(6):22–32.

Kumar, J., Konar, R., and Balasubramanian, K. (2020). The impact of social media on consumers' purchasing behaviour in malaysian restaurants. *Journal of Spatial and Organizational Dynamics*, 8(3):197–216.

Li, J., Kim, W. G., and Choi, H. M. (2021). Effectiveness of social media marketing on enhancing performance: Evidence from a casual-dining restaurant setting. *Tourism Economics*, 27(1):3–22.

Priyadarshini, I. and Cotton, C. (2021). A novel lstm–cnn–grid search-based deep neural network for sentiment analysis. *The Journal of Supercomputing*, 77(12):13911–13932.

Sarica, S. and Luo, J. (2021). Stopwords in technical language processing. *Plos one*, 16(8):e0254937.

Singh, J. and Tripathi, P. (2021). Sentiment analysis of twitter data by making use of svm, random forest and decision tree algorithm. In *2021 10th IEEE International Conference on Communication Systems and Network Technologies (CSNT)*, pages 193–198. IEEE.

Spolaôr, N., Lee, H. D., and Monard, M. C. (2014). Seleção de atributos para aprendizagem multirrótulo. Master's thesis, Universidade de São Paulo.

Srividhya, V. and Anitha, R. (2010). Evaluating preprocessing techniques in text categorization. *International journal of computer science and application*, 47(11):49–51.

Wang, Y., Kim, J., and Kim, J. (2021). The financial impact of online customer reviews in the restaurant industry: A moderating effect of brand equity. *International Journal of Hospitality Management*, 95:102895.

Wever, M., Tornede, A., Mohr, F., and Hüllermeier, E. (2021). Automl for multi-label classification: Overview and empirical evaluation. *IEEE transactions on pattern analysis and machine intelligence*, 43(9):3037–3054.

Yu, C.-E. and Zhang, X. (2020). The embedded feelings in local gastronomy: a sentiment analysis of online reviews. *Journal of Hospitality and Tourism Technology*, 11(3):461–478.

Zhang, M.-L. and Zhou, Z.-H. (2013). A review on multi-label learning algorithms. *IEEE transactions on knowledge and data engineering*, 26(8):1819–1837.