This work is licensed under a Creative Commons Attribution 4.0 International License.

Identificação de avaliações falsas em português usando aprendizagem de máquina

Lucas Percisi [**Universidade Federal da Fronteira Sul** | ***lucas.percisi@estudante.uffs.edu.br***]
Guilherme Dal Bianco [**Universidade Federal do Rio Grande do Sul** | ***guilherme.dalbianco@uffs.edu.br***]

Universidade Federal da Fronteira Sul, SC-484, Km 02, s/n, Chapecó, SC, 89815-899, Brasil.

Received: 30 Novembro 2023 • **Accepted:** 11 Dezembro 2023 • **Published:** DD Month YYYY

Resumo Este estudo enfrenta o crescente desafio das avaliações falsas em plataformas online, um fenômeno que afeta significativamente a confiança do consumidor e o ambiente de negócios. Focando na detecção de avaliações falsas em língua portuguesa, uma área ainda pouco explorada, o trabalho concentra-se em análises de bares e restaurantes no Brasil e Portugal. Utilizando técnicas de aprendizado de máquina e processamento de linguagem natural, o estudo diferencia avaliações genuínas de fraudulentas. Para isso, foi criando um conjunto de dados via *web scraping* do site *yelp.com.br*, sobre o qual se realizou uma análise exploratória, enfatizando *features* textuais e comportamentais. Os resultados indicam que, com um conjunto de dados adequadamente balanceado, foi possível atingir um *F-Score* de 87.53%, um desempenho comparável aos estudos realizados em inglês. Este trabalho oferece *insights* valiosos para a identificação de avaliações falsas em português, contribuindo para a pesquisa e práticas nesse campo.

Keywords: Fake Reviews, Web Scraping, Machine learning, Natural Language Processing, Yelp

1 Introdução

O crescimento acelerado do comércio eletrônico, que disponibiliza uma vasta seleção de produtos e serviços online, tem atraído um número cada vez maior de consumidores para a esfera digital. Plataformas de negócios online como *Amazon*, *Yelp* e *TripAdvisor* oferecem aos clientes a oportunidade de compartilhar avaliações sobre suas experiências, refletindo opiniões sobre os produtos e serviços adquiridos por meio de pontuações e comentários. O conteúdo gerado por esses usuários desempenha um papel influente nas decisões de compra dos consumidores e é fundamental para manter a reputação e os lucros dos vendedores [Jindal and Liu, 2007].

A ascensão do comércio eletrônico, embora tenha trazido inúmeros benefícios, também trouxe consigo a possibilidade de práticas fraudulentas, como a criação de avaliações falsas. Essas avaliações podem distorcer a percepção dos consumidores sobre um produto ou serviço e afetar diretamente as vendas e a reputação das empresas [Luca and Zervas, 2016]. A complexidade na detecção dessas avaliações falsas em ambientes online sublinha a necessidade de desenvolver técnicas eficazes para combatê-las.

As abordagens existentes para identificar avaliações falsas geralmente se baseiam em características linguísticas e comportamentais. As características linguísticas referem-se ao conteúdo textual das avaliações. Já as características comportamentais dizem respeito aos padrões de comportamento dos usuários que publicam as avaliações, observáveis em elementos como a frequência de postagens, histórico de avaliações, data da avaliação, *rating* médio, entre outros indicativos de comportamento de spam [Sandulescu and Ester, 2015].

A tarefa de identificar avaliações falsas é desafiadora, especialmente devido à constante evolução das técnicas de criação dessas avaliações e ao envolvimento de indivíduos contratados especificamente para escrevê-las de forma enganosa. Essas avaliações, muitas vezes impulsionadas artificialmente para influenciar a percepção dos consumidores, são intencionalmente criadas para induzir ao erro, gerando impressões extremamente positivas ou negativas sobre produtos ou serviços. A análise das relações entre usuários e o comportamento coletivo de grupos de avaliadores pode ser um caminho eficaz para identificar essas atividades enganosas, onde padrões de interações e sentimentos expressos revelam comunidades que se engajam no impulsionamento artificial de avaliações [Choo *et al*., 2015].

Apesar de vários estudos realizados ao longo dos anos, ainda é desafiador estabelecer uma distinção clara entre avaliações legítimas e fraudulentas. Isso se deve à dificuldade em discernir objetivamente entre opiniões honestas e desonestas, e até mesmo avaliadores humanos têm dificuldades em identificar avaliações falsas [Crawford *et al*., 2015]. Portanto, é essencial analisar não apenas o conteúdo textual das avaliações, mas também as características dos perfis dos usuários que as realizam.

Este projeto tem como objetivo explorar técnicas de aprendizado de máquina supervisionado, extração de recursos e processamento de linguagem natural para identificar avaliações falsas em português na plataforma *yelp.com.br*. Essa área de pesquisa é menos explorada em comparação com as investigações em inglês, que são amplamente representadas nos estudos de [Birim *et al*., 2022], [Barbado *et al*., 2019], [Elmogy *et al*., 2021] e [Sihombing and Fong, 2019]. Embora a pesquisa em detecção de avaliações falsas em português seja menos comum, ela não é inexistente, como demonstram os trabalhos de [Lima, 2013], [Schuch, 2013] e [Cruz *et al*., 2020]. Esta lacuna é preocupante, pois a desinformação e as avaliações falsas são problemas globais, afetando consumi-

dores e empresas em todas as línguas e regiões. Além disso, técnicas desenvolvidas para o inglês podem não ser tão eficazes em português, devido a diferenças linguísticas e culturais.

Considerando a diversidade linguística e cultural, esta pesquisa busca preencher uma lacuna importante, aplicando técnicas de aprendizado de máquina e processamento de linguagem natural para identificar padrões de inautenticidade em avaliações online em português. O setor de bares e restaurantes no Brasil e Portugal foi escolhido devido à riqueza dos dados disponíveis, essenciais para uma análise detalhada das tendências em avaliações.

O objetivo deste trabalho é investigar o uso de aprendizado de máquina supervisionado e extração de características no contexto da plataforma social *yelp.com.br*, utilizando um conjunto de dados coletados através de *web scraping*, que inclui características textuais e comportamentais. Busca-se identificar atributos distintos entre avaliações verdadeiras e falsas, oferecendo implicações práticas para plataformas online.

A estrutura deste artigo se organiza da seguinte forma: a Seção 2 discorre sobre Trabalhos Relacionados, oferecendo um contexto teórico e explorando pesquisas anteriores. A Seção 3, Referencial Teórico, aprofunda-se nos conceitos e teorias subjacentes à pesquisa. A Seção 4 detalha o Conjunto de Dados, abordando a coleta, processamento e análise exploratória. A Seção 5 descreve a Metodologia aplicada na busca de um modelo eficiente de aprendizado de máquina para a identificação de avaliações fraudulentas. Finalmente, a Seção 6 é dedicada à Análise dos Resultados, discutindo *insights* e implicações práticas do estudo.

2 Referencial Teórico

Este estudo aprofunda-se na interseção entre a análise de texto e os métodos de aprendizado de máquina supervisionado, explorando diversas técnicas e ferramentas essenciais nesse domínio. Uma das técnicas adotadas é o *web scraping*, conforme detalhado no artigo [Ahmed *et al*., 2023]. A técnica de *web scraping* foi utilizada para a extração sistemática de dados da web, os quais foram posteriormente estruturados e salvos no formato CSV. Esta escolha se justifica pela facilidade de manipulação e pela compatibilidade do formato CSV com uma vasta gama de ferramentas analíticas. O processo de converter dados coletados da web, inicialmente não estruturados, em um formato CSV estruturado, possibilitou uma gestão eficiente e simplificada dos dados. Para o *web scraping*, empregamos um *plugin* de navegador denominado *WebScraper* [SIA, 2013]. Este plugin destaca-se por ser gratuito e simplificar significativamente o processo de extração de dados, eliminando a necessidade de desenvolver códigos específicos. Com uma configuração adequada, ele permite a coleta eficiente de dados do site alvo.

2.1 Metafeatures

No decorrer deste estudo, o conceito de *metafeatures* desempenha um papel fundamental. Em contextos de aprendizado de máquina e análise de dados, as *metafeatures* referem-se a atributos derivados ou calculados a partir do conjunto de dados original. Elas fornecem *insights* adicionais e são essenciais para a simplificação de modelos analíticos complexos. Ao contrário dos dados brutos, que são coletados diretamente, as *metafeatures* são fruto de processos analíticos, que agregam valor e compreensão mais profunda dos dados em análise [Shah *et al*., 2018].

2.2 Vetorizadores

No campo do Processamento de Linguagem Natural (PLN), a conversão de texto em dados numéricos é um passo crucial [Kao and Poteet, 2007]. Neste estudo, foi utilizado técnicas como *Term Frequency-Inverse Document Frequency* (*TF-IDF*), *Word2Vec* e *Bag of Words* (BoW) para este propósito.

O método *Term Frequency-Inverse Document Frequency (TF-IDF)* é uma técnica amplamente reconhecida por sua eficácia na ponderação de termos em documentos de texto. No *TF-IDF*, onde o *TF* representa a frequência de um termo em um documento, refletindo sua capacidade de descrever o conteúdo do documento. Por outro lado, o *IDF* mede a raridade de um termo em um conjunto de documentos. Essencialmente, a ideia principal do *TF-IDF* reside no fato de que, se um termo apresenta alta frequência em um documento específico (alto *TF*) e raramente aparece em outros documentos (alto *IDF*), ele é considerado mais relevante para aquele documento [Kuang and Xu, 2010]. Em um contexto de avaliações verdadeiras e falsas, isso pode significar identificar termos que são frequentemente usados em avaliações falsas mas raramente em avaliações verdadeiras, ou vice-versa.

No método *Bag of Words* (*BoW*), textos, como frases ou documentos, são representados pela agregação de suas palavras. Este modelo foca exclusivamente na frequência com que cada palavra ocorre no texto, desconsiderando qualquer aspecto gramatical ou a ordem em que as palavras aparecem. O *BoW* é amplamente utilizado em técnicas de classificação de documentos, onde as características usadas para treinar um classificador são baseadas na ocorrência ou frequência de cada palavra [Qader *et al*., 2019]. Ao contar a frequência das palavras, o *BoW* pode destacar termos que são anormalmente comuns em avaliações falsas. Por exemplo, avaliações falsas podem repetir certas palavras-chave para manipular a percepção de produtos ou serviços.

Por outro lado, o *Word2Vec* representa uma técnica sofisticada que converte palavras em vetores numéricos, empregando redes neurais pré-treinadas para analisar e aprender representações vetoriais de palavras a partir de conjuntos de dados textuais. O principal objetivo do *Word2Vec* é captar a essência semântica das palavras, de modo que palavras com significados semelhantes apresentem representações vetoriais próximas. Existem dois modelos fundamentais dentro do *Word2Vec*: o *Continuous Bag of Words* (CBOW), que prediz uma palavra com base em seu contexto, e o *Skip-gram*, que, inversamente, utiliza uma palavra para prever seu contexto circundante. Ambos os modelos são eficazes na captura de relações semânticas e sintáticas entre as palavras [Yue and Li, 2020]. Diferentemente de métodos como o *BoW*, o *Word2Vec* leva em conta o contexto semântico das palavras, ocasionando que palavras com significados semelhantes serão representadas por vetores semelhantes. Isso

pode ajudar a identificar avaliações falsas que usam certas palavras ou frases de maneira incomum ou fora de contexto.

Neste trabalho, o *TF-IDF*, *BoW* e *Word2Vec* serão referenciados como 'vetorizadores', para facilitar a compreensão, embora cada um tenha suas características distintas. Esses vetorizadores serão responsáveis por transformar o conteúdo textual em dados processáveis pelos algoritmos de aprendizado de máquina.

2.3 Modelos de linguagem

No campo do *PLN*, os modelos de linguagem são essenciais para entender e gerar texto humano de maneira computacional. Entre os modelos mais básicos, além dos *N-grams*, estão os modelos baseados em regras e modelos estatísticos. Os modelos baseados em regras utilizam conjuntos de regras gramaticais e linguísticas predefinidas para analisar ou gerar texto, sendo eficazes em contextos com estrutura gramatical rigorosa. Já os modelos estatísticos, como os *N-grams*, trabalham com a probabilidade de sequências de palavras. Em modelos *N-gram*, um texto é tratado como uma sequência de N itens (geralmente palavras). Por exemplo, na sentença "Eu adoro ler livros", os *bi-grams* (N=2) seriam "Eu adoro", "adoro ler" e "ler livros". Neste contexto, um bigrama é uma sequência de duas palavras adjacentes, e a probabilidade de cada palavra é modelada com base na palavra anterior. Existem ainda modelos de linguagens mais complexos, como os modelos de linguagem neural, que utilizam redes neurais profundas para capturar contextos de longo alcance e características semânticas mais complexas [Speech, 2009]. Avaliações falsas, especialmente aquelas geradas por robôs ou escritores contratados [Crawford *et al*., 2015], podem exibir padrões de escrita distintos. Os *N-grams* podem ajudar a identificar esses padrões, como frases repetidas ou estilos de escrita que são atípicos para avaliações genuínas.

2.4 Aprendizado de Máquina

A aprendizagem de máquina, um ramo da inteligência artificial, utiliza métodos que permitem às máquinas aprimorar seu desempenho em uma tarefa específica através da experiência. Esta área se subdivide em várias categorias, como aprendizado supervisionado, não supervisionado e semi-supervisionado. No aprendizado supervisionado, os dados de treinamento são etiquetados em categorias definidas, e o algoritmo tem como objetivo aprender a atribuir rótulos a novos dados que não fazem parte do conjunto de treinamento original. Já no aprendizado não supervisionado, os dados de treinamento não são previamente rotulados. Aqui, a tarefa do algoritmo é identificar padrões e categorias intrínsecas nos dados para, posteriormente, categorizar novos dados [Mahesh, 2020].

Neste trabalho, foram utilizados algoritmos de aprendizado de máquina supervisionado para analisar e classificar dados das avaliações da plataforma *Yelp*. O objetivo principal é prever se uma avaliação é genuína ou falsa. Cada um desses algoritmos adota uma abordagem única e especializada para esta tarefa, processando tanto dados textuais vetorizados quanto informações numéricas.

O *Random Forest* é um método de aprendizado de máquina que consiste em uma coleção de árvores de regressão randomizadas. Cada árvore na floresta fornece uma previsão para um ponto de consulta, e a previsão final do *Random Forest* é uma combinação das previsões de todas as árvores individuais. Cada árvore é construída usando uma amostra aleatória dos dados e um subconjunto aleatório das característica para determinar as melhores divisões [Biau and Scornet, 2016]. No estudo de [Elmogy *et al*., 2021], o *Random Forest* foi o que obteve os melhores resultados para classificação de avaliações falsas.

Já o *Support Vector Machine (SVM)* têm como objetivo principal encontrar um hiperplano que melhor separe as classes de dados no espaço de características. O *SVM* busca maximizar a margem, que é a distância entre o hiperplano e os pontos de dados mais próximos de cada classe, conhecidos como vetores de suporte. Eficiente em dados linearmente separáveis, o SVM também pode lidar com dados não linearmente separáveis, que transforma os dados para uma dimensão superior onde a separação é possível. A escolha do *kernel* e seus parâmetros é crucial para o desempenho do modelo [Ben-Hur and Weston, 2010]. O *SVM* foi uma das opções de classificadores exploradas por [Elmogy *et al*., 2021]

O *K-Nearest Neighbors (KNN)*, que obteve o melhor desempenho no estudo de [Elmogy *et al*., 2021], é um método não-paramétrico e baseado em instâncias utilizado para tarefas de classificação e regressão. Ele funciona identificando os k vizinhos mais próximos de uma amostra de teste no espaço de características. Para classificação, o resultado é determinado pela maioria dos votos desses vizinhos, enquanto para regressão, pela média de seus valores. A proximidade é calculada usando métricas de distância, como a *Euclidiana*, por exemplo. O valor de k é crucial, influenciando diretamente a precisão do modelo: um 'k' pequeno pode ser sensível ao ruído, e um k grande pode suavizar demais as fronteiras de decisão. Diferente de outros algoritmos, o *KNN* não realiza um treinamento explícito, mas armazena os dados de treinamento e faz cálculos na fase de predição. Isso pode torná-lo computacionalmente intensivo em grandes conjuntos de dados. A seleção adequada de k e da métrica de distância, geralmente feita através de validação cruzada, é fundamental para o desempenho eficaz do *KNN* [Hastie *et al*., 2009].

A *Logistic Regression* (LR) é um método estatístico usado principalmente para problemas de classificação binária, onde a variável de saída possui dois estados possíveis (como "sim" ou "não"). Ela prevê a probabilidade de ocorrência de um evento aplicando a função logística ou *sigmóide*, que transforma valores contínuos em uma escala entre 0 e 1. Esse método foi explorado no estudo de [Sihombing and Fong, 2019] e de [Elmogy *et al*., 2021] com o proposito de identificar avaliações falsas. Diferente da regressão linear, a regressão logística lida bem com relações não lineares entre a variável dependente e as variáveis independentes. Ela estima coeficientes para cada variável independente, demonstrando o efeito de cada uma na probabilidade do evento [Daiv *et al*., 2020].

Por fim, o *XGBoost (eXtreme Gradient Boosting)* representa uma implementação otimizada do algoritmo de *boosting* de gradiente, um método de aprendizado de máquina

aplicado tanto em problemas de classificação quanto de regressão. Métodos de *boosting* foram explorados no trabalho de [Barbado *et al*., 2019], que utilizou o *AdaBoost*, alcançando o melhor desempenho em seu estudo. Similarmente, o *XGBoost* foi empregado no estudo de [Sihombing and Fong, 2019], onde também demonstrou um desempenho superior. O algoritmo *XGBoost* opera construindo árvores de decisão de maneira sequencial, com cada nova árvore sendo criada para corrigir os erros das árvores antecessoras. Diferentemente de outros métodos de *boosting*, o *XGBoost* integra várias otimizações que aprimoram tanto a eficiência quanto a eficácia. Uma dessas otimizações é o uso de regularização, que adiciona uma penalidade a modelos excessivamente complexos para prevenir o sobreajuste. Essa característica faz do *XGBoost* uma ferramenta particularmente eficaz em evitar que as árvores de decisão se tornem demasiadamente complexas e específicas aos dados de treinamento [Chen and Guestrin, 2016].

No desenvolvimento deste estudo, foi adotada uma abordagem meticulosa para otimizar os parâmetros dos classificadores utilizados. Através *GridSearchCV*, uma ferramenta do *scikit-learn* [Bird *et al*., 2009], para automatizar a busca dos melhores parâmetros e realizar a validação cruzada dos dados. A técnica de validação cruzada *stratificada* de 5 dobras, consistentemente empregada ao longo do trabalho, é um método que ajuda a mitigar a variabilidade dos resultados da validação cruzada e aumenta a confiabilidade na seleção do modelo, conforme sugerido por [Krstajic *et al*., 2014]. O uso do *GridSearchCV* em conjunto com a validação cruzada *stratificada* permite uma pesquisa exaustiva dos parâmetros e assegura uma seleção eficiente, levando em conta múltiplas divisões dos dados para evitar viés em um único conjunto de dados. A opção pela validação cruzada *stratificada* de 5 dobras também tem o objetivo de preservar a proporção de amostras em cada classe, assegurando assim a representatividade e robustez dos modelos de classificação desenvolvidos.

2.5 Métricas

No âmbito deste estudo, várias métricas foram empregadas para avaliar o desempenho dos classificadores, com o F-Score sendo a métrica central. Ao final do estudo, foram detalhadas todas as métricas comumente usadas para avaliar o desempenho dos modelos selecionados. A acurácia é uma medida geral de desempenho que indica a proporção de previsões corretas (tanto Verdadeiros Positivos - VP, quanto Verdadeiros Negativos - VN) em relação ao total de previsões feitas. Neste contexto, um Verdadeiro Positivo (*VP*) ocorre quando o modelo corretamente identifica uma avaliação falsa como falsa. Um Verdadeiro Negativo (*VN*), por outro lado, acontece quando uma avaliação verdadeira é corretamente identificada como verdadeira. Um Falso Positivo (*FP*) ocorre quando o modelo erroneamente classifica uma avaliação verdadeira como falsa, enquanto um Falso Negativo (*FN*) se dá quando uma avaliação falsa é erroneamente classificada como verdadeira.

A precisão, em seguida, é a proporção de *VP's* dentre todos os resultados classificados como positivos (*VP's + FP's*), indicando a confiança do modelo ao etiquetar uma avaliação como falsa. O *recall*, ou sensibilidade, refere-se à proporção de *VP's* identificados em relação a todos os exemplos que são efetivamente falsos (*VP's + FNs'*). O F-Score, nossa métrica de foco, harmoniza precisão e *recall* em uma única medida, calculando a média harmônica entre eles. Esta métrica é particularmente valiosa em cenários como a detecção de avaliações falsas, onde um equilíbrio entre a capacidade de identificar corretamente avaliações falsas (precisão) e a capacidade de capturar a maior quantidade possível dessas avaliações (*recall*) é crucial [Sokolova *et al*., 2006].

3 Trabalhos Relacionados

A literatura sobre a detecção de avaliações falsas ou *fake reviews* apresenta uma variedade de abordagens, com estudos como [Birim *et al*., 2022], [Barbado *et al*., 2019], [Elmogy *et al*., 2021] e [Sihombing and Fong, 2019] focando no uso de técnicas de aprendizado de máquina para identificar essas avaliações. Outro aspecto interessante é abordado por [Valdivia *et al*., 2019], que explora as incongruências nas avaliações, propondo um índice unificado para correlacionar melhor a nota do usuário e a análise de sentimento, contribuindo assim para um entendimento alternativo das avaliações de produtos e serviços em plataformas online.

O estudo de [Birim *et al*., 2022] foca na detecção de avaliações falsas na *Amazon*, utilizando um conjunto de dados balanceado em 10500 avaliações verdadeiras e 10500 falsas. Este trabalho analisa a eficiência de vários classificadores, incluindo *Decision Tree*, *Random Forest*, *Support Vector Machine (SVM)* e *Neural Networks*, testando diferentes combinações de características. Entre as características avaliadas estão pontuações de sentimento, distribuições de tópicos, distribuições de *clusters* e uma matriz esparsa de palavras, além de fatores centrados no revisor, como comprimento da avaliação e a característica de *Compra Verificada*. Os resultados destacam que características relacionadas ao comportamento, especialmente a *Compra Verificada*, são fundamentais na distinção entre avaliações autênticas e falsas, especialmente quando combinadas com aspectos textuais. Quando a pontuação de sentimentos foi removida do modelo de classificação RF, que combinava Distribuição de Tópicos e *Compra Verificada*, houve uma pequena melhora no modelo. Notavelmente, o modelo RF com Distribuição de Tópicos e *Compra Verificada* alcançou os melhores resultados, com um F-score de 80,08%. O estudo utilizou validação cruzada de 5 dobras para treinamento e teste.

Em [Barbado *et al*., 2019], foi realizado um estudo detalhado sobre a detecção de avaliações falsas, utilizando um conjunto de dados próprio e balanceado extraído do *Yelp*. Este conjunto de dados consiste em avaliações de quatro cidades dos EUA, focadas em empresas varejistas de eletrônicos, com um número igual de avaliações confiáveis e falsas para cada cidade, onde as avaliações filtradas pelo *Yelp* foram consideradas falsas. Ao analisar características textuais centradas na avaliação, a técnica de *TF-IDF* com bigramas foi a mais eficaz, mas alcançou um F-Score abaixo de 60%, indicando que as características textuais não são fortes indicadores de veracidade das avaliações no domínio de eletrônicos de consumo. No entanto, ao focar em características comportamentais dos usuários, o estudo, utilizando o classi-

ficador *AdaBoost*, alcançou um *F-Score* de 82% utilizando validação cruzada de 10 dobras. Este resultado ressalta a eficácia das características comportamentais na detecção de avaliações falsas e a importância de considerar o comportamento do usuário na plataforma, além do conteúdo textual, para identificar avaliações fraudulentas de maneira eficiente.

Outro estudo revisado foi o de [Elmogy *et al*., 2021], que conduziu uma análise detalhada sobre a detecção de avaliações falsas usando um conjunto de dados do *Yelp* focando no conteúdo textual. Este conjunto de dados incluiu 5.853 avaliações de 201 hotéis em Chicago, classificadas pelo *Yelp* em 4.709 avaliações reais e 1.144 falsas. Os autores não utilizaram validação cruzada em sua análise, mas sim um método de divisão de treino e teste com proporções de 70/30% respectivamente. Além de experimentar uma variedade de classificadores, incluindo *K-Nearest Neighbors (KNN)*, *Naive Bayes*, *Support Vector Machine (SVM)*, *Logistic Regression* e *Random Forest*, os autores também consideraram modelos de linguagem *N-gram*, especificamente *bigram* e *tri-gram*. Além disso, foram consideradas *metafeatures* como quantidade de letras maiúsculas, quantidade de pontuações e quantidade de *emojis* de cada avaliação, que são indicativos do comportamento dos usuários durante a escrita de suas avaliações. A pesquisa revelou que a incorporação dessas características melhorou o desempenho dos classificadores em 3,80%. O *KNN* (K=7) com *TF-IDF* e *tri-grams* apresentou o melhor F-Score entre os modelos testados, com um valor de 86,20%.

Em [Sihombing and Fong, 2019] , foi realizada uma investigação sobre a detecção de avaliações falsas em um conjunto de dados do *Yelp*. O conjunto de dados originalmente apresentava um desequilíbrio significativo, com uma proporção muito maior de avaliações não filtradas em comparação com as filtradas, totalizando 53400 avaliações verdadeiras e 8141 avaliações falsas. Para abordar este problema, os autores aplicaram técnicas de *oversampling*, criando cópias da classe minoritária. Eles exploraram diversas técnicas de classificação de aprendizado de máquina, incluindo *Logistic Regression*, *Gaussian Naive Bayes*, *Support Vector Machine* e *XGBoost*, utilizando características tanto textuais quanto comportamentais nas avaliações. O conjunto de dados treinado ficou com 53400 avaliações verdadeiras e 24421 avaliações falsas, sendo que, das 24421 avaliações falsas 16280 são provenientes do processo de *oversampling*. Os autores destacaram o desempenho superior do *XGBoost*, que alcançou um F-Score aproximado de 99%, e para os outros classificadores utilizados ficou no máximo em 78%. Os autores não descreveram o uso de validação cruzada no decorrer do estudo, nem forneceram informações adicionais sobre como o *XGBoost* alcançou tal eficácia. Esta falta de detalhamento pode sugerir a possibilidade de *overfitting* no modelo *XGBoost*. Por fim, a análise detalhada de [Sihombing and Fong, 2019] descreve problemas enfrentados no desempenho dos modelos testados devido o desbalanceamento do conjunto de dados.

No estudo de [Valdivia *et al*., 2019], a equipe de pesquisa concentrou-se em resolver as incongruências entre as avaliações numéricas dos usuários e as análises de sentimentos em textos de avaliações no *TripAdvisor*. Foram investigadas as discrepâncias frequentemente observadas onde os usuários, apesar de apresentarem classificações positivas, incluíam frases negativas ou neutras em suas avaliações textuais. Para isso, analisaram as opiniões sobre seis monumentos italianos e espanhóis, aplicando oito Métodos de Análise de Sentimento (*SAM's*) para analisar as correlações entre a polaridade numérica dada pelo usuário e a polaridade expressa nas sentenças das avaliações. Como solução, [Valdivia *et al*., 2019] propuseram o Modelo de Agregação de Polaridade, que integra a polaridade do usuário e a polaridade extraída pelos *SAM's*. Este modelo, calibrado através de um parâmetro β, mostrou-se eficaz em equilibrar as duas polaridades, fornecendo uma avaliação corrigida que fica entre a nota de avaliação do usuário com o sentimento empregado na avaliação detectado pelo SAM. O estudo revelou que esse novo modelo de agregação pode ser especialmente útil para reavaliar os sentimentos expressos nas avaliações de diversos monumentos, no entanto, o desafio de identificar um bom β em outros contextos pode dificultar a aplicação do modelo. O estudo de [Valdivia *et al*., 2019] destaca a importância de abordagens complementares para assegurar que as avaliações de produtos e serviços oferecidos online sejam fidedignas com a realidade, alinhando-se ao objetivo de identificar avaliações falsas.

Os resultados apresentados anteriormente oferecem uma visão ampla das abordagens para combater avaliações falsas, desde a análise de características textuais e comportamentais dos usuários até a integração de análise de sentimento. Os estudos destacam a importância da detecção de avaliações falsas, dada a influência significativa que as avaliações têm nas decisões dos consumidores [Jindal and Liu, 2007]. Além disso, os estudos revisados mostram uma tendência marcante nas pesquisas sobre avaliações falsas: a grande maioria foi realizada com conteúdo em inglês. Esta concentração no inglês reflete a ampla utilização da língua em plataformas globais e a maior facilidade de acesso a dados nesse idioma. No entanto, essa limitação linguística restringe a eficácia das técnicas de detecção de avaliações falsas, uma vez que não considera as características específicas de outros idiomas. Essa tendência para o inglês revela uma importante lacuna no campo de estudos de avaliações falsas, principalmente se considerarmos o alcance mundial do comércio eletrônico e das plataformas de avaliação online. Por isso, é crucial ampliar as pesquisas para incluir outros idiomas, como o português, para desenvolver métodos de detecção mais abrangentes e eficazes, adaptados às particularidades linguísticas e culturais de cada região.

4 Método de pesquisa

Este trabalho foi meticulosamente conduzido através de várias etapas, abrangendo desde a coleta inicial de dados até a análise exploratória detalhada. A metodologia adotada é ilustrada na Figura 1, que esquematiza o fluxo completo do processo de pesquisa. Este fluxo inclui:

1. **Coleta e Estruturação de Dados:** Inicialmente, foi realizado o *web scraping* para a aquisição de dados. Essa etapa envolveu a coleta e a organização dos dados brutos, estabelecendo uma base sólida para as análises subsequentes.

2. **Pré-processamento dos dados:** Aqui, os dados coletados passaram por um processo de limpeza e padronização, garantindo a qualidade e a uniformidade dos dados.
3. **Extração de *features*:** Esta fase foi dedicada à identificação e extração de características relevantes dos dados, preparando-os para a modelagem.
4. **Classificação das *features*:** As *features* extraídas foram então categorizadas em comportamentais ou textuais, facilitando a seleção de variáveis de acordo com cada fase da pesquisa.
5. **Busca dos melhores parâmetros:** Realizou-se a seleção das melhores *features* e hiperparâmetros para o treinamento, conforme as necessidades de cada fase da pesquisa, aplicando técnicas de otimização para melhorar o desempenho do modelo.
6. **Treinamento dos modelos:** Com as *features* e hiperparâmetros otimizados, os modelos resultantes foram treinados.
7. **Análise dos resultados:** Finalmente, os resultados obtidos foram analisados, permitindo avaliações e interpretações detalhadas do desempenho dos modelos.

A Figura 1 abaixo representa graficamente essa metodologia, proporcionando uma visão geral e integrada do processo de pesquisa empregado neste trabalho.

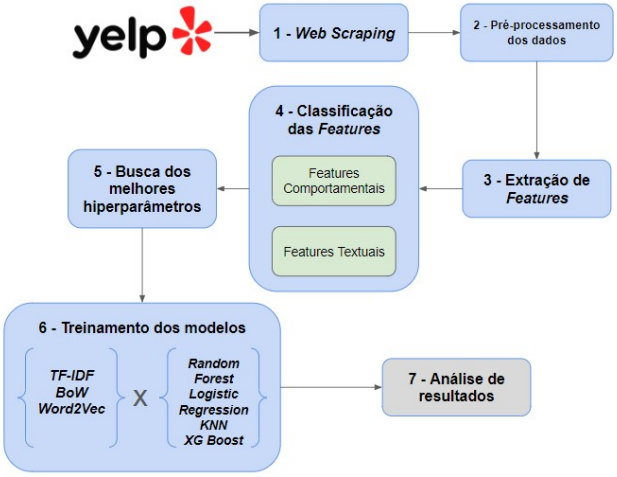

Figure 1. Estrutura geral do trabalho.

4.1 Coleta dos dados

O conjunto de dados foi coletado por meio de técnicas de *web scraping* utilizando a ferramenta *WebScraper*. O foco da coleta de dados esteve em avaliações relacionadas a restaurantes e bares, com particular atenção às avaliações escritas em português. Este direcionamento para o setor de alimentação deve-se à abundância de dados disponíveis e à relevância desses estabelecimentos para análises de avaliações online.

A escolha da plataforma *Yelp* como fonte primária de dados é fundamentada em suas características únicas e aplicabilidade em algoritmos de aprendizado de máquina. O *Yelp* se destaca por possuir uma seção específica em cada página de estabelecimento denominada 'avaliações não recomendadas'. Essa área, visível ao público, contém avaliações que são filtradas pela própria plataforma, e embora visíveis ao público, não são recomendadas para os usuários. A disponibilidade desses dados permitiu considerá-los como 'falsos' para o propósito deste estudo, viabilizando a definição de um conjunto de dados com rótulos claros de verdadeiro e falso. A relevância dessa escolha é corroborada pelo estudo de [Mukherjee *et al*., 2013], que afirma a consistência do filtro do *Yelp* em categorizar uma avaliação como falsa quando de fato ela é filtrada pela plataforma. No entanto, é importante notar que a rotulagem como *falsa* não necessariamente indica uma avaliação que foi feita com a intenção de enganar. Em alguns casos, a classificação *falsa* pode representar uma discrepância entre a avaliação e a experiência relatada do produto. Portanto, o termo *falsa* usado aqui é uma simplificação, sendo mais apropriado considerar essas avaliações como potencialmente enganosas ou inconsistentes.

Apesar de uma maior variedade de informações estar disponível nas avaliações verdadeiras, a seleção se concentrou nas características comuns a ambos os tipos, assegurando uma abordagem justa e consistente na formação do conjunto de dados para o treinamento dos modelos. As informações coletadas abrangem tanto aspectos textuais quanto comportamentais dos usuários. As características textuais extraídas incluem o conteúdo das avaliações, que oferecem percepções valiosas sobre o estilo de escrita e a linguagem utilizada pelos usuários. Quanto às características comportamentais, foram coletadas informações como a quantidade de amigos dos usuários na plataforma, refletindo seu nível de interação social; o número total de avaliações feitas por cada usuário, indicando seu nível de atividade; e a quantidade de fotos postadas, que pode sinalizar o engajamento visual do usuário com os estabelecimentos. Além disso, foram capturadas a classificação dada pelo usuário e a presença de foto no perfil do usuário, elementos que podem influenciar a percepção de credibilidade das avaliações. A data da postagem de cada avaliação também foi registrada, porém não utilizamos essa *feature* nesse estudo.

A Figura 2 apresenta um exemplo ilustrativo de uma avaliação genuína coletada de um restaurante em São Paulo. Por outro lado, a Figura 3 mostra um caso de avaliação falsa do mesmo estabelecimento. Embora ambas as avaliações tenham recebido nota cinco, elas apresentam diferenças significativas. Adicionalmente, a Tabela 1 oferece uma visão comparativa da estrutura do conjunto de dados inicial, exibindo as *features* coletadas de cada exemplo conforme Figura 2 e 3.

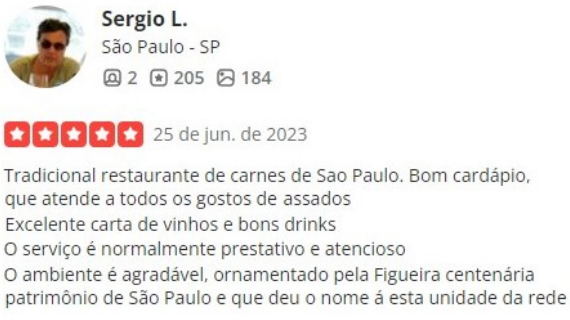

Figure 2. Informações extraídas de uma avaliação verdadeira do *Yelp*.

Figure 3. Informações extraídas de uma avaliação falsa do *Yelp*.

Table 1. Comparação entre uma Avaliação Verdadeira e uma Falsa extraídas da plataforma *Yelp*.

Avaliação Verdadeira	**Avaliação Falsa**
Conteúdo: Tradicional restaurante de carnes de Sao Paulo. Bom cardápio, que atende a todos os gostos de assados Excelente carta de vinhos e bons drinks O serviço é normalmente prestativo e atencioso O ambiente é agradável, ornamentado pela Figueira centenária patrimônio de São Paulo e que deu o nome á esta unidade da rede	**Conteúdo:** otima comida e atendimento, recomendo muitooo
Quantidade de Amigos: 2	**Quantidade de Amigos:** 0
Quantidade de Avaliações: 205	**Quantidade de Avaliações:** 2
Quantidade de Fotos: 184	**Quantidade de Fotos:** 0
Classificação: 5	**Classificação:** 5
Usuário com Foto: Sim	**Usuário com Foto:** Sim
Data da postagem: 25/06/2023	**Data da postagem:** 20/08/2023
Avaliação Falsa: Não	**Avaliação Falsa:** Sim

As diferenças observadas entre as avaliações genuínas e falsas são marcantes e podem ser utilizadas como indicadores na detecção de avaliações inautênticas. Primeiramente, a avaliação verdadeira apresenta um conteúdo textual detalhado e específico, mencionando aspectos concretos do restaurante, como o cardápio variado, a qualidade dos vinhos e bebidas, além da decoração e do serviço. Em contrapartida, a avaliação falsa exibe um conteúdo textual superficial e genérico, sem detalhar experiências ou elementos específicos do estabelecimento.

Outro aspecto relevante é a interação do usuário com a plataforma. Na avaliação verdadeira, o usuário possui um histórico robusto, evidenciado pela quantidade significativa de amigos, avaliações e fotos postadas, sugerindo um envolvimento consistente e autêntico com a plataforma. Por outro lado, na avaliação falsa, a ausência de amigos e fotos, juntamente com um número extremamente baixo de avaliações, indica um perfil pouco ativo ou recém-criado, o que pode ser um sinal de inautenticidade.

Essas diferenças, somadas ao contexto geral de cada avaliação, fornecem indícios valiosos para a identificação de avaliações falsas, contribuindo significativamente para a precisão de modelos de detecção baseados em análise de dados e características do usuário.

Além do conjunto de dados principal em português, um segundo conjunto em inglês foi coletado. Este conjunto em inglês compreendeu 26.692 avaliações verdadeiras e 14.567 avaliações falsas de bares e restaurantes dos EUA. Com a mesma estrutura do conjunto em português, ele foi submetido aos mesmos procedimentos de pré-processamento, exceto pelo *POS-Tagging*, conforme detalhado na Seção 4.2. O objetivo de incluir este conjunto de dados em inglês é possibilitar a validação dos modelos finais desenvolvidos para o português.

4.2 Pré-processamento

O pré-processamento de dados textuais é uma etapa crucial na análise de dados linguísticos, particularmente no contexto de aprendizado de máquina. Esta fase começa com a padronização dos dados, que visa homogeneizar os formatos e corrigir eventuais inconsistências. A padronização é fundamental para garantir que os dados estejam em um formato uniforme e apto para análise, prevenindo erros nas etapas posteriores [Speech, 2009].

Durante o pré-processamento do conjunto de dados coletado, foram implementadas várias ações para assegurar sua qualidade e exatidão. Inicialmente, registros sem comentários foram removidos e valores nulos foram tratados, além disso, foi realizada uma padronização das variáveis para uniformizar os dados. Um passo crítico na preparação dos conjuntos de dados foi a implementação de um filtro linguístico com a biblioteca *langdetect* [Shuyo, 2010]. Este filtro selecionou exclusivamente avaliações em português para o conjunto de dados em português e, similarmente, apenas avaliações em inglês para o conjunto em inglês. Esse procedimento se mostrou vital, pois a análise preliminar indicou a presença de avaliações em várias línguas, indicando a presença de avaliações realizadas por estrangeiros em estabelecimentos brasileiros e portugueses. De maneira similar, o conjunto em inglês também passou por esse filtro, assegurando a consistência linguística. No conjunto final de dados em português, incluíram-se 13.957 avaliações, com 3.387 classificadas como falsas e 10.570 como verdadeiras.

4.3 Extração de *Features*

Na busca por compreender e identificar avaliações falsas, este estudo se embasa na análise aprofundada das interações dos usuários com a plataforma e do conteúdo que compartilham. Cada informação desempenha um papel específico na análise do conjunto de dados, conforme detalhado na Tabela 2 e descrito a seguir.

As características comportamentais (C), descritas pela quantidade de amigos (*qtd_friends*), o número total de avaliações feitas pelo usuário (*qtd_reviews*), a quantidade de fotos postadas (*qtd_photos*), a classificação dada pelo usuário (*rating*), e a presença de foto no perfil do usuário (*user_has_photo*) oferecem uma visão quantitativa da atividade do usuário na plataforma.

As *features* textuais (T), representadas pelo conteúdo das avaliações (*content*), são analisadas para extrair padrões de linguagem e expressões que podem ser indicativos de autenticidade ou falsidade. Para complementar esta análise, o estudo incorpora *metafeatures* comportamentais (MC), como a contagem de pontuações (*punctuation_count*), a contagem de letras maiúsculas (*capital_count*), e o total de palavras na avaliação (*word_count*), que possibilitam um exame mais detalhado das nuances no modo de escrever dos usuários.

Table 2. Descrição dos tipos e características do conjunto de dados

Nome da Característica	Tipo de Característica	Descrição
qtd_friends	C	Número de amigos do usuário na plataforma.
qtd_reviews	C	Total de avaliações feitas pelo usuário.
qtd_photos	C	Número de fotos postadas pelo usuário.
rating	C	Classificação dada pelo usuário ao estabelecimento.
user_has_photo	C	Presença de foto no perfil do usuário.
punctuation_count	MC	Contagem de pontuação no conteúdo da avaliação.
capital_count	MC	Número de letras maiúsculas no conteúdo da avaliação.
word_count	MC	Total de palavras na avaliação.
content	T	Texto da avaliação.
content_tagged	MT	Avaliação com *POS-tagging* aplicado.

Além disso, foi realizada a eliminação de *stop words*, que são palavras comuns que geralmente não contribuem para o significado de um texto, como preposições e artigos. A remoção dessas palavras ajuda a reduzir o ruído nos dados, permitindo que os algoritmos de aprendizado de máquina se concentrem nas palavras que são verdadeiramente significativas e informativas. Além disso, foi realizada a eliminação de dados duplicados, incompletos ou que não agregam valor à análise proposta, para garantir que o conjunto de dados mantenha apenas as informações mais pertinentes e úteis [Kondrak, 2005].

Por fim, após a remoção das *stop-words* foi adicionado ao conjunto de dados uma *metafeature* textual (MT) (*content_tagged*) que incorpora a classificação gramatical de cada palavra da avaliação, possibilitando observar diferenças de significados de palavras iguais em contextos diferentes. Este *POS-tagger*, desenvolvido pelo [Inoue, 2019], foi treinado com a biblioteca *NLTK* e o corpus *Mac-Morpho* [Fonseca and Rosa, 2013], alcançando uma acurácia de 92.19%. As *features* são categorizadas em diferentes tipos para facilitar a análise. Neste processo, a abordagem adotada foi concatenar a classificação gramatical com a palavra correspondente, enriquecendo o contexto textual. As *features* comportamentais (C) referem-se a características relacionadas ao comportamento do usuário na plataforma. As *metafeatures* comportamentais (MC) são atributos derivados das *features* comportamentais, fornecendo *insights* adicionais. As *features* textuais (T) estão ligadas ao conteúdo escrito das avaliações. Por fim, as *metafeatures* textuais (MT) representam informações extraídas ou derivadas do texto das avaliações.

4.4 Análise Exploratória de Dados

A Análise Exploratória de Dados (*AED*) realizada neste estudo abordou a comparação de características entre avaliações verdadeiras e falsas no conjunto de dados em português. Para uma análise mais equitativa da distribuição de características entre avaliações verdadeiras e falsas, aplicamos gráficos de densidade, conhecidos como *Kernel Density Estimates (KDE)*. Estes gráficos proporcionam uma visão contínua e suave da distribuição dos dados, sendo úteis em conjunto de dados com desequilíbrio nos rótulos, conforme descrito por [Silverman, 1986]. Além disso, os gráficos *KDE* permitem uma observação clara de onde existe uma maior concentração de informação por classe. Por exemplo, suponha que ao observarmos que o pico das avaliações verdadeiras em relação à quantidade de amigos teria sido de 0.02 no eixo y para 0 amigos, enquanto que para as avaliações falsas, o pico teria sido de 0.04 também para 0 amigos. Isso sugere uma maior densidade de avaliações falsas para usuários sem amigos listados. Importante ressaltar que os valores de densidade em um gráfico KDE representam estimativas de probabilidade e não quantidades diretas. Assim, um valor mais alto no eixo y para as avaliações falsas indica uma tendência de maior probabilidade de encontrar avaliações falsas entre usuários com 0 amigos, mas não necessariamente o dobro do número de avaliações verdadeiras, conforme exemplo.

4.4.1 Análise das *features*

A análise exploratória realizada neste estudo oferece um entendimento mais profundo sobre as características comportamentais dos usuários na plataforma *Yelp* para o conjunto de dados desse trabalho. No primeiro gráfico, Figura 4, observamos a distribuição da quantidade de amigos dos usuários. Aqui, é notável uma maior prevalência de avaliações falsas entre usuários com um número menor de amigos, sugerindo que a escassez de conexões sociais na plataforma pode estar relacionada a comportamentos fraudulentos. Na Figura 5 foca na quantidade de avaliações postadas pelos usuários. Este gráfico indica que usuários com poucas avaliações prévias tendem mais a publicar avaliações inautênticas, destacando um padrão de comportamento em que usuários menos ativos na plataforma são propensos a participar em atividades suspeitas. Finalmente, o gráfico da Figura 6 analisa a relação entre a quantidade de fotos postadas pelos usuários e a autenticidade das avaliações. É revelado que usuários que compartilham poucas ou nenhuma foto estão mais inclinados a postar avaliações suspeitas, enfatizando que um baixo engajamento visual na plataforma pode ser um indicativo de avaliações falsas. Cada um desses aspectos, analisado individualmente, contribui para um entendimento mais abrangente das tendências e características associadas às avaliações falsas.

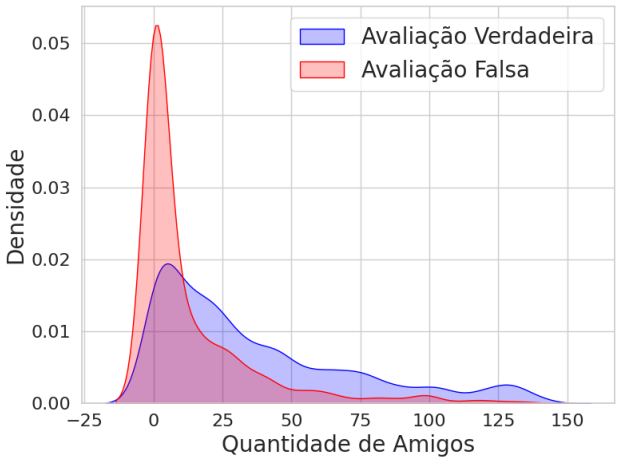

Figure 4. Densidade da quantidade de amigos.

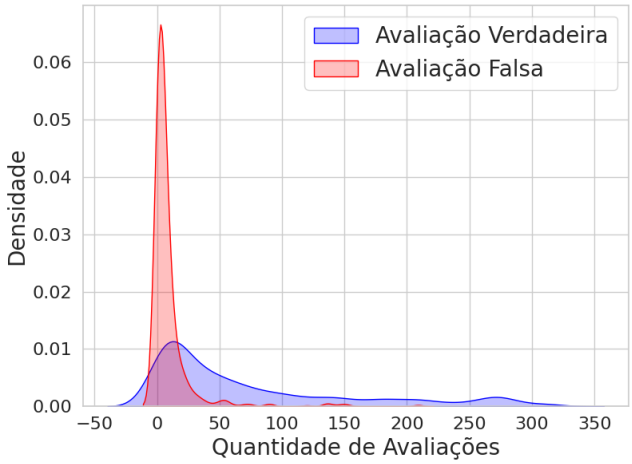

Figure 5. Densidade da quantidade de avaliações.

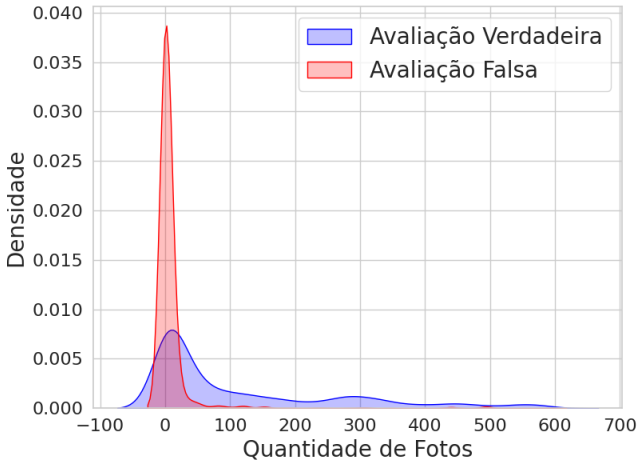

Figure 6. Densidade da quantidade de fotos.

A Figura 7 exibe um histograma de densidade que compara a presença de fotos de perfil em avaliações verdadeiras e falsas. Avaliações falsas possuem maior propensão de estar associadas a usuários sem foto de perfil. Esse padrão sugere que perfis sem fotos podem ser mais propensos a produzir avaliações inautênticas. Em contraste, avaliações verdadeiras tendem a ser de usuários com fotos de perfil, indicando um nível maior de autenticidade.

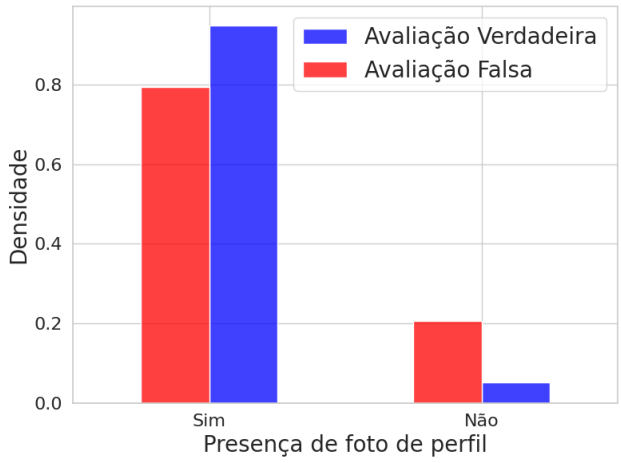

Figure 7. Densidade dos usuários que possuem foto de perfil

O histograma de densidade na Figura 8 destaca um padrão relevante na distribuição das classificações das avaliações. A análise mostra que as avaliações, especialmente as falsas, tendem a concentrar-se nas extremidades da escala, com notas significativamente mais altas nos valores 1 e 5. Este comportamento indica que usuários que publicam avaliações falsas são mais propensos a atribuir classificações extremas - seja a pontuação mínima ou máxima - em comparação com avaliações verdadeiras. Tal tendência para notas polarizadas nas avaliações falsas pode refletir uma tentativa de influenciar fortemente a percepção geral de um estabelecimento, seja positiva ou negativamente.

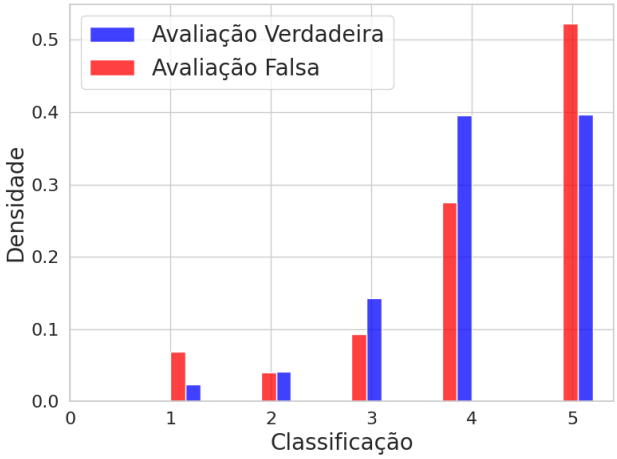

Figure 8. Densidade da classificação das avaliações

Na matriz de correlação apresentada na Figura 9, as *metafeatures* textuais e as *features* comportamentais são visualmente distintas em termos de correlação. As *metafeatures* textuais, localizadas no canto inferior direito da matriz, exibem uma forte correlação positiva, aproximando-se de 1, o que reflete sua inter-relação direta, visto que todas estão intrinsecamente ligadas à natureza do conteúdo textual.

Em contraste, as *features* comportamentais, situadas no canto superior esquerdo da matriz, também mostram uma forte correlação positiva dentro deste grupo, indicando valores próximos de 1. Por outro lado, observa-se baixa ou nenhuma correlação entre as *features* comportamentais e as *metafeatures* textuais, com valores próximos de 0, indicando que estas variáveis tendem a operar independentemente umas das outras. Isso reforça a falta de relação significativa entre esses dois conjuntos de variáveis. Notavelmente, nenhuma das relações na matriz apresenta uma correlação fortemente negativa, que se aproximaria de -1, destacando que as variáveis, em sua maior parte, não se influenciam inversamente de maneira significativa.

4.4.2 Análise de *metafeatures*

As diferenças nos gráficos de densidade para as *metafeatures* textuais oferecem outras perspectivas sobre as discrepâncias entre avaliações verdadeiras e falsas. Essa análise foca na quantidade de pontuação, palavras e letras maiúsculas em cada avaliação, *metafeatures* relacionadas diretamente ao conteúdo textual.

A Figura 10 agrupa vários gráficos de densidade, cada um

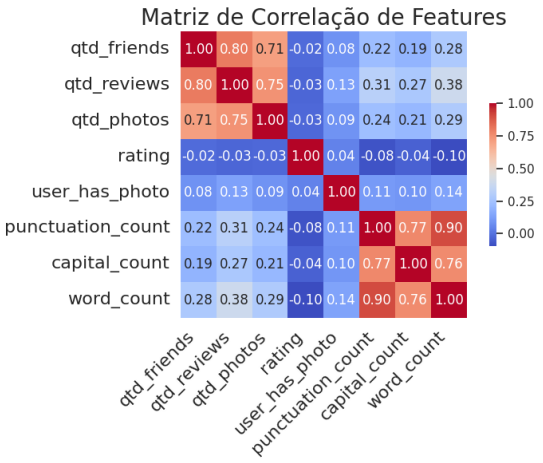

Figure 9. Matriz de Correlação entre as *Features*.

explorando uma *metafeature* textual diferente. Na Figura 10 (a), observamos a densidade da quantidade de pontuação usada nas avaliações. Este gráfico mostra que, embora existam diferenças entre avaliações verdadeiras e falsas, a discrepância não é tão acentuada quanto em outras características comportamentais. Avaliações falsas tendem a usar menos pontuação, o que pode ser um indicativo sutil, mas não um diferenciador claro. Já na Figura 10 (b), a análise da quantidade de palavras nas avaliações revela que as falsas geralmente são mais curtas que as verdadeiras. Essa tendência, embora notável, não fornece um padrão distintivo suficiente para ser usada como único critério de identificação de avaliações falsas. Por fim, na Figura 10 (c), a distribuição do uso de letras maiúsculas é examinada. Aqui, vemos uma tendência de uso mais baixo de letras maiúsculas em avaliações falsas. Embora esta seja uma diferença observável, ela não é tão pronunciada a ponto de atuar como um indicador conclusivo por si só.

Essa análise conjunta das *metafeatures* textuais ilustra que, apesar de existirem variações sutis entre avaliações verdadeiras e falsas, elas não são tão pronunciadas quanto as diferenças encontradas nas características comportamentais. Isso sugere que para uma detecção eficaz de avaliações falsas, é importante combinar *features* textuais e comportamentais. Contudo, é necessário um cuidado na seleção de *metafeatures* textuais para o modelo, evitando a inclusão de variáveis com relevância distintiva limitada, para manter a eficácia do modelo de detecção.

5 Abordagem e seleção do modelo

Neste estudo, focado em identificar avaliações falsas em português, três abordagens principais foram exploradas: (i) baseada exclusivamente nas características comportamentais dos usuários; (ii) explora somente o conteúdo textual das avaliações; e (iii) combina ambas as abordagens.

A Figura 11 ilustra o diagrama de seleção de modelo adotado. Nas etapas indicadas pelas caixas azuis, o *GridSearchCV* [Bird *et al*., 2009] foi utilizado para determinar as características comportamentais mais relevantes, os *n-grams* mais eficazes para cada combinação de vetorizador e classificador, e os hiperparâmetros ideais para os classificadores com foco no conteúdo textual. Nesta fase da pesquisa, foi procedido com o balanceamento do conjunto de dados em português, alinhando-o pela classe minoritária. No estudo de [Sihombing and Fong, 2019] é relatado uma situação semelhante, em que os resultados obtidos com dados desbalanceados eram consideravelmente inferiores em comparação com um conjunto de dados balanceado. Como resultado, o conjunto de dados desse estudo agora consiste em 3387 avaliações verdadeiras e 3387 avaliações falsas. Para alcançar esse equilíbrio, foi excluído um número de avaliações verdadeiras excedentes. A seleção dos dados a serem removidos foi feita fixando o parâmetro de estado aleatório em 42 no método *sample* do *dataframe Pandas* [Pedregosa *et al*., 2011], garantindo assim a reprodutibilidade dos testes. Em todas as etapas de busca de melhores parâmetros, foi utilizada a técnica de validação cruzada de 5 dobras.

A escolha dos algoritmos *K-Nearest Neighbors (KNN)*, *Random Forest*, *XGBoost*, *Logistic Regression* e *Support Vector Machine (SVM)* para identificar avaliações falsas foi baseada na eficácia descritas na revisão da literatura deste estudo. Na revisão, os autores demonstraram o sucesso de algoritmos como *Random Forest* [Birim *et al*., 2022], *AdaBoost* [Barbado *et al*., 2019], *KNN* [Elmogy *et al*., 2021] e *XGBoost* [Sihombing and Fong, 2019], além de terem testado também o *SVM* [Birim *et al*., 2022] e a *Logistic Regression* [Elmogy *et al*., 2021] na classificação de avaliações falsas.

5.1 Seleção das melhores *features* comportamentais

O processo de busca das melhores *features* comportamentais por classificador envolveu duas etapas. Na etapa inicial, os classificadores *Random Forest*, *Logistic Regression*, *SVM*, *KNN* e *XGBoost* foram otimizados utilizando o *GridSearchCV* [Bird *et al*., 2009], focando nas *features* comportamentais (C + MC). Este processo buscou as melhores configurações para cada classificador, baseando-se em um conjunto específico de opções predefinidas. As configurações otimizadas para cada classificador estão detalhadas na Tabela 3.

Table 3. Melhores Configurações de hiperparâmetros por Classificador.

Classificador	Hiperparâmetros
Random Forest	max_depth: None, min_samples_leaf: 1, min_samples_split: 2, n_estimators: 1000
Logistic Regression	C: 250, penalty: l2, solver: lbfgs
KNN	n_neighbors: 13, metric: manhattan, weights: uniform, p: 1
SVM	C: 100, gamma: scale, kernel: rbf, max_iter: 5000
XGBoost	learning_rate: 0.01, max_depth: 11, n_estimators: 500

No estudo de [Elmogy *et al*., 2021], os autores observaram que a pontuação de sentimento das avaliações influenciava negativamente o desempenho do modelo. Portanto, com o objetivo de identificar se a remoção de alguma *feature* do conjunto de dados deste estudo influenciaria negativamente o modelo, foi realizado um processo iterativo na segunda etapa de seleção das melhores *features* comportamentais. Este processo consistiu em verificar a importância de cada *feature* em relação a cada classificador.

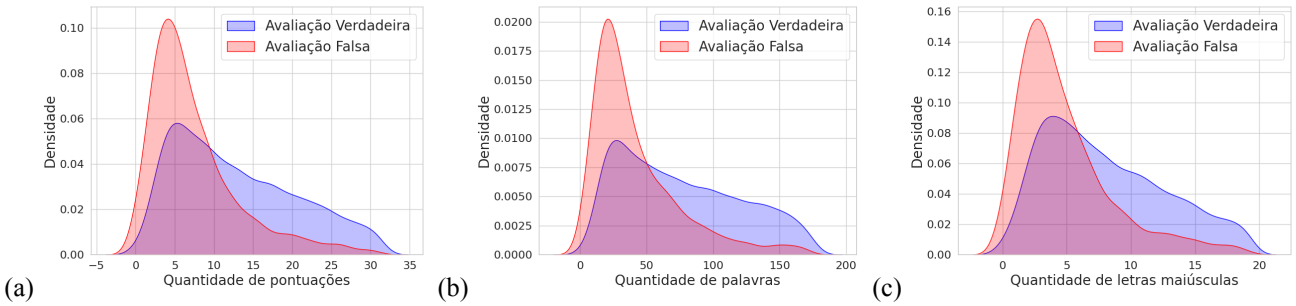

Figure 10. Densidade das *metafeatures* comportamentais.

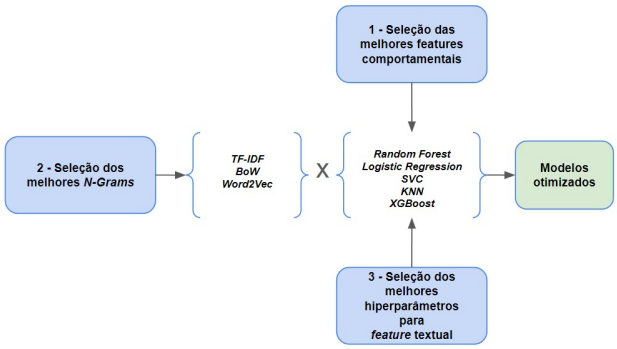

Figure 11. Diagrama da seleção do modelo.

Após estabelecer os melhores hiperparâmetros para as *features* comportamentais (C) e as *metafeatures* comportamentais (MC), procedeu-se com um ciclo de treinamento e teste iterativo, utilizando apenas essas *features*. Em cada ciclo, o F1-score foi avaliado e a *feature* de menor importância, identificada por técnicas de permutação [Breiman, 2001], foi removida. Este processo continuou até que restassem somente duas *features*, permitindo assim identificar as combinações de *features* mais eficientes para cada classificador. Os conjuntos de *features* que maximizaram a eficácia de cada classificador podem ser encontrados na Tabela 4.

Table 4. Melhores *Features* comportamentais por Classificador

Classificador	F1-Score médio	Melhores Features
Random Forest	**88.79%**	qtd_friends, qtd_reviews, qtd_photos
Logistic Regression	81.50%	qtd_friends, qtd_reviews, user_has_photo, word_count
KNN	84.94%	qtd_friends, qtd_reviews, qtd_photos
SVM	82.60%	qtd_friends, qtd_reviews, qtd_photos, punctuation_count, word_count
XGBoost	88.68%	qtd_friends, qtd_reviews, qtd_photos

Ao examinar as *features* de maior importância em cada classificador, conforme ilustrado na Tabela 4, observamos uma predominância de *qtd_friends*, *qtd_reviews* e *qtd_photos*. Esta tendência sugere que essas *features* são fundamentais para otimizar o desempenho do F1-score no modelo. Tal observação está em consonância com a análise exploratória, que já apontava para diferenças significativas no comportamento dessas *features*. Além disso, a Tabela 4 destaca a habilidade dos modelos selecionados em classificar avaliações com o *F1-Score* médio acima de 80% baseando-se unicamente em *features* comportamentais. O *Random Forest* e o *XGBoost* pela melhor eficácia, com um F1-score médio de 88.79%, e 88.68% respectivamente. Esses resultados são consistentes com aqueles encontrados nos estudos de [Birim *et al.*, 2022] e [Barbado *et al.*, 2019], que chegam na casa do 80% com *features* comportamentais.

5.2 Seleção dos *N-Grams*

Após a definição das melhores *features* para cada classificador, a próxima fase do estudo concentrou-se exclusivamente no conteúdo textual. Conforme adotado trabalho do [Elmogy *et al.*, 2021], onde os autores empregaram o uso de *N-Grams* e obtiveram melhorias na eficácia do modelo, nesta etapa, o objetivo principal é identificar o *N-gram* com maior eficácia para cada combinação de classificador e vetorizador. Para alcançar isso, empregou-se o método *GridSearchCV*, para estabelecer os *N-grams* que proporcionaram o melhor desempenho em termos de F1-score médio.

A Tabela 5 resume os resultados desta fase, apresentando as configurações ótimas de *N-grams* para cada cenário de vetorização utilizando o *TF-IDF* e o *BoW*. Com exceção do vetorizador *Word2Vec*, que se limita a *uni-grams* devido à sua natureza intrínseca. O *Word2Vec* é especialmente projetado para aprender representações vetoriais de palavras individuais em seus contextos específicos, ao invés de focar em sequências de palavras, característica predominante nos métodos de *N-grams*.

Por outro lado, para os outros vetorizadores, como *TF-IDF* e *BoW*, foram avaliadas diversas combinações de *N-grams*, incluindo seis variações diferentes: (1, 1), (1, 2), (1, 3), (2, 2), (2, 3) e (3, 3). A escolha desses *N-grams* foi feita com o intuito de explorar a eficácia de diferentes tamanhos de sequências de palavras na captura de padrões no conteúdo textual das avaliações.

A relevância desta etapa do estudo está na percepção de que a eficácia na identificação de avaliações falsas é influenciada não somente pela seleção de *features*, mas igualmente pelo modo como o conteúdo textual é processado e examinado. As configurações de *N-grams* escolhidas visam não apenas aprimorar o desempenho de cada modelo, mas também a otimizar o processo de treinamento, evitando opções menos eficientes.

5.3 Seleção dos hiperparâmetros para a *feature* textual

Nesta fase do estudo, o objetivo central foi identificar os hiperparâmetros mais adequados dos classificadores para o

Table 5. Melhores *N-grams* por classificador e vetorizador.

Vetorizador	Classificador	Melhor n-gram
TF-IDF	Random Forest	(1, 3)
	Logistic Regression	(1, 1)
	KNN	(1, 1)
	SVM	(1, 1)
	XGBoost	(1, 2)
BoW	Random Forest	(1, 2)
	Logistic Regression	(1, 3)
	KNN	(1, 1)
	SVM	(1, 2)
	XGBoost	(1, 3)

processamento do conteúdo textual. Para alcançar a configuração ideal dos classificadores, utilizou-se o método *GridSearchCV* com os vetorizadores parametrizados com os melhores *N-Grams* selecionados anteriormente.

Este processo minucioso resultou em ajustes precisos para cada classificador em diferentes contextos de vetorização, refletindo a complexidade e a especificidade do desafio de identificar avaliações falsas. Os resultados detalhados deste processo, incluindo os hiperparâmetros finais ajustados para cada classificador, são apresentados na Tabela 6. Importante ressaltar que um desempenho mais baixo era esperado nesta fase, principalmente ao considerar as evidências na literatura que indicam que o uso exclusivo de *features* textuais não é a abordagem mais promissora na detecção de avaliações falsas [Barbado *et al*., 2019].

5.4 Modelos Selecionados

Por fim, uma seleção de 15 modelos distintos foi realizada, com base em sua eficácia comprovada em cada etapa do processo de definição de parâmetros. Esse processo envolveu uma análise detalhada das combinações possíveis de características comportamentais e textuais, além da seleção criteriosa de *N-grams* apropriados para cada vetorizador. A escolha desses modelos foi orientada pelo objetivo de maximizar a eficácia na identificação de avaliações falsas, levando em consideração as particularidades do idioma português e as nuances comportamentais dos usuários.

Os detalhes das melhores configurações para cada combinação de vetorizador e classificador foram compilados na Tabela 7, que resume as escolhas estratégicas feitas. Esta tabela apresenta uma visão clara das combinações mais eficientes entre vetorizadores como *TF-IDF*, *BoW* e *Word2Vec*, e classificadores, incluindo *Random Forest*, *Logistic Regression*, *KNN*, *SVM* e *XGBoost*. Esse processo meticuloso de seleção dos melhores hiperparâmetros e características foi fundamental para avaliar os modelos mais promissores, dada a vasta gama de combinações possíveis.

6 Análise dos Resultados

Os testes foram realizados no Anaconda versão 2.5.1, utilizando a IDE *Spyder* 5.4.3 e *Python* versão 3.11.5. Um aspecto crucial do estudo foi a análise do impacto de conjuntos de dados balanceados e desbalanceados [Shah *et al*., 2018],

Table 6. Desempenho de cada Classificador e Vetorizador com os melhores hiperparâmetros e *N-Grams*.

Vetorizador	Classificador	F1-Score Médio	Hiperparâmetros
TF-IDF	Random Forest	73.58%	max_depth: 1000, min_samples_leaf: 1, min_samples_split: 3, n_estimators: 500
	Logistic Regression	69,60%	C: 1, penalty: l2, solver: newton-cg
	KNN	69.07%	metric: euclidean, n_neighbors: 17, weights: uniform
	SVM	71.50%	C: 100, gamma: auto, kernel: rbf, max_iter: 2000
	XGBoost	72.38%	learning_rate: 0.01, max_depth: 15, n_estimators: 1000, min_child_weight: 10
BoW	Random Forest	73.63%	max_depth: 1000, min_samples_leaf: 1, min_samples_split: 2, n_estimators: 1000
	Logistic Regression	72.01%	C: 1, penalty: l2, solver: newton-cg, max_iter: 2000
	KNN	67.08%	metric: euclidean, n_neighbors: 3, weights: uniform
	SVM	73.29%	C: 100, gamma: auto, kernel: rbf, max_iter: 5000
	XGBoost	66.86%	learning_rate: 0.01, max_depth: None, n_estimators: 500, min_child_weight: 1
Word2Vec	Random Forest	65.27%	max_depth: None, min_samples_leaf: 1, min_samples_split: 3, n_estimators: 1000
	Logistic Regression	61.67%	C: 500, solver: newton-cg, penalty: l2, max_iter: 2000
	KNN	59.11%	metric: euclidean, n_neighbors: 17, weights: uniform
	SVM	66.61%	C: 50, gamma: auto, kernel: poly, max_iter: 1000
	XGBoost	64.48%	learning_rate: 0.01, max_depth: 9, n_estimators: 500, min_child_weight: 10

considerando o desequilíbrio inerente ao conjunto de dados em português. O foco dos modelos finais foi integração de características comportamentais eficazes com conteúdo textual otimizado para cada classificador. Foi utilizado o conjunto de dados balanceado em português como base para essa integração, resultando em combinações otimizadas de *features* comportamentais, *N-grams* e hiperparâmetros.

Paralelamente, durante a coleta de dados em português, foi coletado um conjunto de dados correspondente em inglês para bares e restaurantes, processado de forma idêntica ao conjunto em português, a fim de avaliar o impacto da mudança de linguagem quando aplicada aos modelos. Os conjuntos de dados avaliados nos modelos selecionados seguem conforme a lista:

1. **Conjunto de dados balanceado em português (i):** Composto por 3.387 avaliações verdadeiras e 3.387 avaliações falsas.
2. **Conjunto de dados balanceado em inglês (ii):** Este conjunto consiste de 3.387 avaliações verdadeiras e 3.387 avaliações falsas.

Table 7. Modelos resultantes por combinação de Vetorizador e Classificador

Vetorizador	Classificador	Melhor N-gram	Melhores *Features* Comportamentais
TF-IDF	Random Forest	(1, 3)	qtd_friends, qtd_reviews, qtd_photos
	Logistic Regression	(1, 1)	qtd_friends, qtd_reviews, user_has_photo, word_count
	KNN	(1, 1)	qtd_friends, qtd_reviews, qtd_photos
	SVM	(1, 1)	qtd_friends, qtd_reviews, qtd_photos, punctuation_count, word_count
	XGBoost	(1, 2)	qtd_friends, qtd_reviews, qtd_photos
BoW	Random Forest	(1, 2)	qtd_friends, qtd_reviews, qtd_photos
	Logistic Regression	(1, 3)	qtd_friends, qtd_reviews, user_has_photo, word_count
	KNN	(1, 1)	qtd_friends, qtd_reviews, qtd_photos
	SVM	(1, 2)	qtd_friends, qtd_reviews, qtd_photos, punctuation_count, word_count
	XGBoost	(1, 3)	qtd_friends, qtd_reviews, qtd_photos
Word2Vec	Random Forest	(1, 1)	qtd_friends, qtd_reviews, qtd_photos
	Logistic Regression	(1, 1)	qtd_friends, qtd_reviews, user_has_photo, word_count
	KNN	(1, 1)	qtd_friends, qtd_reviews, qtd_photos
	SVM	(1, 1)	qtd_friends, qtd_reviews, qtd_photos, punctuation_count, word_count
	XGBoost	(1, 1)	qtd_friends, qtd_reviews, qtd_photos

3. **Conjunto de dados balanceado em inglês (iii):** Composto por 14.597 avaliações verdadeiras e 14.597 avaliações falsas.
4. **Conjunto de dados desbalanceado em português (iv):** Este conjunto é formado por 10.570 avaliações verdadeiras e 3.387 avaliações falsas.
5. **Conjunto de dados desbalanceado em inglês (v):** Este conjunto inclui 10.570 avaliações verdadeiras e 3.387 avaliações falsas.

Após a seleção dos modelos finais, as melhores *features* numéricas para cada classificador foram padronizadas e transformadas em matrizes esparsas para assegurar a compatibilidade com as matrizes textuais vetorizadas. Dessa forma, as *features* textuais e comportamentais foram unificadas. Essas matrizes numéricas foram então concatenadas horizontalmente às matrizes de texto vetorizado. Esta combinação resultou em matrizes finais nas quais as colunas iniciais representam características do texto, enquanto as últimas colunas refletem as *features* comportamentais dos usuários. Este procedimento foi consistentemente aplicado em todos os conjuntos de dados, permitindo que os modelos de classificação processassem simultaneamente informações textuais e comportamentais, o que foi essencial para a análise final dos modelos.

Para os cinco conjuntos de dados, os modelos foram treinados e testados individualmente, mantendo exclusivamente os hiperparâmetros e *features* selecionados na fase de otimização dos modelos. Esta fase de otimização empregou o conjunto de dados base (i) na busca e definição destes parâmetros.

A Figura 12 exibe os três melhores desempenhos de F-Score para cada conjunto de dados, organizados por vetorizador e classificador. Cada barra no gráfico é acompanhada por um indicador no topo (barra de erro), representando a variação mínima e máxima de desempenho daquele modelo durante a validação cruzada.

Para o conjunto de dados que embasou a busca dos melhores parâmetros (i), podemos observar resultados mais estáveis e homogêneos, destacando o *KNN* e o *XGBoost* com todas as opções de vetorizadores, passando da marca dos 85% de *F-Score*. Aos demais modelos, apesar de terem ficado abaixo dos 85%, tiveram um bom desempenho comparados aos outros conjuntos de dados, provavelmente porque cada modelo foi otimizado para esse mesmo conjunto de dados. Além disso, a variação de *F-Score* nos modelos ficou baixa, com exceção do *SVM*, que testado junto com o *Word2Vec* apresentou oscilações fora do normal.

Para o conjunto de dados (ii), que é em inglês e está balanceado com a mesma quantidade do conjunto de dados (i), os resultados destacaram-se para a *Logistic Regression* utilizando *TF-IDF* e *BoW*, alcançando cerca de 83% de *F-Score*. Neste cenário, mesmo com a diminuição do desempenho dos modelos em geral, os melhores resultados aproximaram-se dos observados no conjunto de dados (i), mantendo a consistência na variação da validação cruzada. De forma similar, o *SVM* apresentou instabilidade ao ser aplicado com *Word2Vec* e *TF-IDF*.

Para o conjunto de dados (iii), que além de ser em inglês e balanceado, é mais de quatro vezes maior que o conjunto de dados (i), os resultados ganharam destaque para a *Logistic Regression* com *BoW* e *TF-IDF*, inclusive *XGBoost* com *TF-IDF*, que estão em torno da marca de 85% de *F-Score*. Nesse conjunto também é possível observar uma variação de desempenho maior entre os classificadores no conjunto de dados (i), isso novamente devido ao conjunto de dados (i) ter sido a base de otimização dos parâmetros. Além disso, a variação da validação cruzada nesse conjunto de dados foi menor, devido maior quantidade de dados. É possível observar que o *SVC*, apesar de ter ficado estável com os 3 vetorizadores, teve o pior desempenho nesse conjunto de dados.

Agora, observando o comportamento dos modelos no conjunto de dados desbalanceados (iv), em português, nota-se uma piora considerável em comparação com o conjunto (i). A diferença do melhor *F-Score* do conjunto (iii) com o melhor *F-Score* do conjunto (i) chega a 15.5%. Além disso, para os melhores do conjunto (iii), é notável também uma estabilidade da validação cruzada tanto quanto no conjunto (i). Neste conjunto, o *SVM* também se comportou de forma anormal quando executado junto com o *Word2Vec*, zerando alguns F-Scores durante a validação cruzada.

Por fim, o conjunto de dados (v), que tem o mesmo dimensionamento do conjunto de dados (iv), apresenta uma diferença de desempenho mais intensa entre os classificadores. Porém, como no cenário (iii) em inglês, o *Logistic Regression* junto com o *BoW* se destacou, obtendo um melhor desempenho do que o *KNN* do conjunto de dados (iii). Com exceção desse modelo, os outros que se destacaram foram o

Logistic Regression e *XGBoost* com *TF-IDF*, aproximando-se dos 70%.

É notável o impacto que o desbalanceamento de dados tem no desempenho dos modelos no contexto desse trabalho, além de que, observando a Figura 12, a mudança de língua traz diferenças de forma geral nos modelos. Entretanto, ao selecionar um modelo específico, observa-se um desempenho consistente, reforçando a confiabilidade dos indicadores de avaliações falsas em ambos os idiomas. Nos conjuntos de dados em inglês, a *Logistic Regression* apresentou os melhores resultados. Já nos conjuntos em português, os modelos *KNN* e *XGBoost* se destacaram em eficácia. Analisando todos os cenários e a tabela 8, percebe-se a presença do *XGBoost* como o terceiro melhor em todos os cenários, evidenciando sua robustez no contexto testado. Por outro lado, o *SVM*, que apresentou várias instabilidades nos testes, parece ser sensível à quantidade de dados disponíveis para treinamento. Especificamente no conjunto de dados (iii), que possui o maior volume de dados, o *SVM* exibiu resultados mais estáveis. Isso sugere que uma quantidade maior de dados é necessária para otimizar o desempenho deste modelo.

O balanceamento dos conjuntos de dados influenciou positivamente o desempenho dos modelos, e a maior dimensão dos conjuntos de dados contribuiu para a redução na variação das métricas, enfatizando a importância de um volume substancial de dados para a eficiência na detecção de avaliações falsas online. Além disso, revisitando a Tabela 4, que mostra o desempenho dos modelos considerando apenas *features* comportamentais, constatou-se que os melhores foram o *Random Forest* e o *XGBoost*, com 88,79% e 88,68% respectivamente com o conjunto de dados base (i), e nos modelos que consideram *features* textuais, nenhum superou essa marca, alinhando-se com os achados do estudo de [Birim *et al*., 2022], onde os autores observaram que uma *feature* baseada em conteúdo textual reduziu o desempenho do modelo. Outra conclusão interessante é que na Tabela 8, que mostra os três melhores modelos para cada conjunto de dados, nos que são desbalanceados, apesar de terem casos com alta acurácia e alta precisão, o *recall* é bem reduzido, indicando que muitas das avaliações que são realmente falsas não estão sendo corretamente identificadas pelo modelo.

De forma geral, estes resultados revelaram modelos capazes de alcançar desempenhos comparáveis na língua portuguesa a outros modelos na literatura em língua inglesa com a mesma proposta: identificar avaliações falsas através do uso de aprendizado de máquina e processamento de linguagem natural. Os resultados estão em consonância com os resultados dos estudos de [Elmogy *et al*., 2021], [Birim *et al*., 2022], [Barbado *et al*., 2019] e [Sihombing and Fong, 2019], e que também destacam a importância das *features* comportamentais em relação às textuais, aspecto crucial para obter os resultados apresentados. Além disso, o trabalho de [Sihombing and Fong, 2019] relatou os problemas causados pelo desbalanceamento de dados, um fenômeno que também foi observado neste estudo.

7 Conclusão

O presente estudo abordou a identificação de avaliações falsas em português provenientes da plataforma *Yelp*, empregando uma combinação de técnicas de aprendizado de máquina e análise de características textuais e comportamentais. Foi demonstrada a eficácia dos modelos desenvolvidos, evidenciando a capacidade de distinguir entre avaliações autênticas e inautênticas utilizando aprendizado de máquina e técnicas de processamento de linguagem natural em um contexto linguístico pouco explorado, a língua portuguesa. Notavelmente, os resultados obtidos foram iguais ou melhores que aqueles revisados na literatura utilizando dados na língua inglesa.

As descobertas deste estudo têm implicações relevantes para a comunidade científica e para os profissionais envolvidos na detecção de fraudes online. Elas oferecem novas perspectivas sobre a combinação de características textuais e comportamentais e ressaltam a necessidade de utilizar conjuntos de dados balanceados e representativos para o treinamento. Além disso, os resultados alcançados com a validação dos métodos utilizando dados em inglês estão em consonância com os estudos revisados neste trabalho, contribuindo também para a confiabilidade dos resultados obtidos com os dados em português e assegurando a robustez das conclusões deste trabalho.

Este estudo indica caminhos para futuras investigações, recomendando a exploração da análise de sentimentos e o aperfeiçoamento da lematização em português. Sugere-se também a pesquisa de outras abordagens de utilização do *POS-Tagging* e a pesquisa sobre outros métodos de extração de características do conteúdo textual, afim de melhorar o desempenho dos modelos validados. Estas direções têm o objetivo de melhorar a precisão na identificação de avaliações falsas e ampliar a aplicabilidade dos modelos em diversos contextos e idiomas.

Como proposta prática, considerando que as características comportamentais foram cruciais para alcançar bons resultados, recomenda-se que as plataformas de comércio eletrônico aprimorem continuamente seus mecanismos para registrar dados comportamentais dos usuários, além de outras informações que facilitem a identificação de fraudes.

Por fim, destaca-se a importância da colaboração interdisciplinar e da continuação das pesquisas nessa área, dada sua crescente relevância no mundo digital contemporâneo. Espera-se que este estudo inspire futuras investigações e contribua de maneira significativa para o campo do processamento de linguagem natural e a detecção de fraudes online.

Disponibilidade de dados e materiais

Todo o projeto, incluindo o *dataset*, os resultados e os *scripts* de *web scraping* utilizados, estão disponíveis publicamente no *GitHub* como contribuição para a comunidade. Os interessados podem acessar o repositório para obter mais detalhes e explorar os recursos do projeto:

- **Repositório GitHub:** `https://github.com/lucaspercisi/yelp-fake-reviews-ptbr`

O repositório inclui:

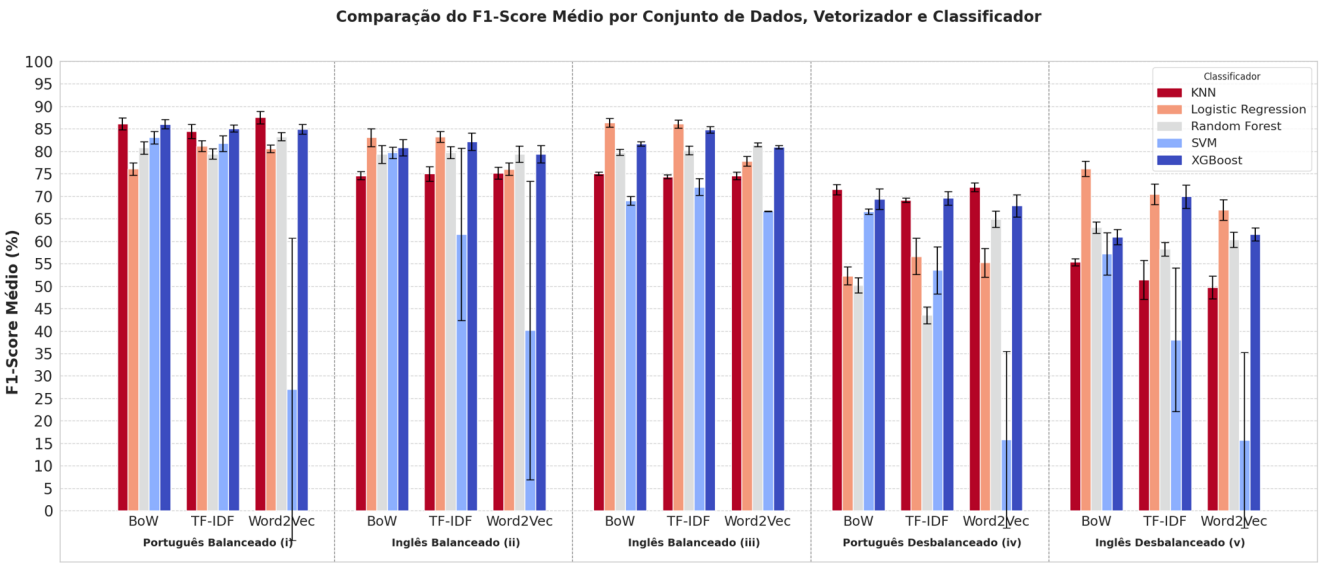

Figure 12. F1-Score de cada Cenário, Vetorizador e Classificador

Table 8. Comparação de Desempenho por Cenário, Vetorizador e Classificador

Conjunto de dados	Tamanho	Vetorizador	Classificador	Accuracy	Precision	Recall	F1-Score
Português Balanceado (i)	V: 3387 | F: 3387	Word2Vec	KNN	86.54%	81.54%	**94.48%**	**87.53%**
Português Balanceado (i)	V: 3387 | F: 3387	BoW	KNN	85.09%	80.46%	92.71%	86.15%
Português Balanceado (i)	V: 3387 | F: 3387	BoW	XGBoost	85.27%	81.70%	90.94%	86.06%
Inglês Balanceado (ii)	V: 3387 | F: 3387	TF-IDF	Logistic Regression	84.31%	89.56%	77.68%	83.19%
Inglês Balanceado (ii)	V: 3387 | F: 3387	BoW	Logistic Regression	82.95%	82.45%	83.73%	83.07%
Inglês Balanceado (ii)	V: 3387 | F: 3387	TF-IDF	XGBoost	81.74%	80.36%	84.06%	82.16%
Inglês Balanceado (iii)	V: 14597 | F: 14597	BoW	Logistic Regression	86.26%	85.96%	86.69%	86.32%
Inglês Balanceado (iii)	V: 14597 | F: 14597	TF-IDF	Logistic Regression	86.64%	89.71%	82.77%	86.10%
Inglês Balanceado (iii)	V: 14597 | F: 14597	TF-IDF	XGBoost	84.47%	82.79%	87.04%	84.86%
Português Desbalanceado (iv)	V: 10570 | F: 3387	Word2Vec	KNN	86.29%	71.31%	72.78%	72.03%
Português Desbalanceado (iv)	V: 10570 | F: 3387	BoW	KNN	84.08%	63.21%	82.40%	71.53%
Português Desbalanceado (iv)	V: 10570 | F: 3387	TF-IDF	XGBoost	85.58%	71.30%	67.96%	69.58%
Inglês Desbalanceado (v)	V: 10570 | F: 3387	BoW	Logistic Regression	**89.75%**	87.70%	67.20%	76.09%
Inglês Desbalanceado (v)	V: 10570 | F: 3387	TF-IDF	Logistic Regression	88.80%	**97.83%**	55.06%	70.45%
Inglês Desbalanceado (v)	V: 10570 | F: 3387	TF-IDF	XGBoost	86.56%	76.49%	64.39%	69.91%

- Conjunto de dados em português e inglês utilizadas nesse trabalho
- *Scripts* para *web scraping* específico para o site da *Yelp* utilizado com a ferramenta *WebScrapper*.
- Código fonte utilizado no desenvolvimento desse trabalho.
- Arquivos com os resultados dos cenários de treinamento.

References

Ahmed, A., Khan, M. A., and Ishtiaq, A. (2023). Web scraping for scientific discovery: Strategies for secure data retrieval, structured transformation, and relevant content selection.

Barbado, R., Araque, O., and Iglesias, C. A. (2019). A framework for fake review detection in online consumer electronics retailers. *Information Processing & Management*, 56(4):1234–1244.

Ben-Hur, A. and Weston, J. (2010). A user's guide to support vector machines. *Data mining techniques for the life sciences*, pages 223–239.

Biau, G. and Scornet, E. (2016). A random forest guided tour. *Test*, 25:197–227.

Bird, S., Klein, E., and Loper, E. (2009). *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

Birim, Ş. Ö., Kazancoglu, I., Mangla, S. K., Kahraman, A., Kumar, S., and Kazancoglu, Y. (2022). Detecting fake reviews through topic modelling. *Journal of Business Research*, 149:884–900.

Breiman, L. (2001). Random forests. *Machine learning*, 45:5–32.

Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794.

Choo, E., Yu, T., and Chi, M. (2015). Detecting opinion spammer groups through community discovery and sentiment analysis. In *Data and Applications Security and Privacy XXIX: 29th Annual IFIP WG 11.3 Working Conference, DBSec 2015, Fairfax, VA, USA, July 13-15, 2015, Proceedings 29*, pages 170–187. Springer.

Crawford, M., Khoshgoftaar, T. M., Prusa, J. D., Richter, A. N., and Najada, H. A. (2015). Survey of review spam detection using machine learning techniques. *Journal of*

Big Data, 2(1):1–24.

Cruz, B. d. P. A., Pimenta, S., Dutton, A., and Ross, S. D. (2020). Fake on-line reviews em restaurantes: intenção de boicote ou intenção de buycott de telespectadores do programa pesadelo na cozinha?

Daiv, K., Lachake, M., Jagtap, P., Dhariwal, S., and Gutte, V. (2020). An approach to detect fake reviews based on logistic regression using review-centric features. *Int. Res. J. Eng. Technol.(IRJET)*, 7(06):2107–2112.

Elmogy, A. M., Tariq, U., Ammar, M., and Ibrahim, A. (2021). Fake reviews detection using supervised machine learning. *International Journal of Advanced Computer Science and Applications*, 12(1).

Fonseca, E. and Rosa, J. L. G. (2013). Mac-morpho revisited: Towards robust part-of-speech tagging. In *Proceedings of the 9th Brazilian symposium in information and human language technology*.

Hastie, T., Tibshirani, R., Friedman, J. H., and Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer.

Inoue, M. (2019). Pos-tagger-portuguese-nltk. `https://github.com/inoueMashuu/POS-tagger-portuguese-nltk`. Acessado em: 11 de novembro de 2023.

Jindal, N. and Liu, B. (2007). Analyzing and detecting review spam. In *Seventh IEEE international conference on data mining (ICDM 2007)*, pages 547–552. IEEE.

Kao, A. and Poteet, S. R. (2007). *Natural language processing and text mining*. Springer Science & Business Media.

Kondrak, G. (2005). N-gram similarity and distance. In *International symposium on string processing and information retrieval*, pages 115–126. Springer.

Krstajic, D., Buturovic, L. J., Leahy, D. E., and Thomas, S. (2014). Cross-validation pitfalls when selecting and assessing regression and classification models. *Journal of cheminformatics*, 6(1):1–15.

Kuang, Q. and Xu, X. (2010). Improvement and application of tf• idf method based on text classification. In *2010 International Conference on Internet Technology and Applications*, pages 1–4. IEEE.

Lima, H. d. C. S. d. C. (2013). Detectando avaliações spam em uma rede social baseada em localização.

Luca, M. and Zervas, G. (2016). Fake it till you make it: Reputation, competition, and yelp review fraud. *Management Science*, 62(12):3412–3427.

Mahesh, B. (2020). Machine learning algorithms-a review. *International Journal of Science and Research (IJSR).[Internet]*, 9(1):381–386.

Mukherjee, A., Venkataraman, V., Liu, B., and Glance, N. (2013). What yelp fake review filter might be doing? In *Proceedings of the international AAAI conference on web and social media*, volume 7, pages 409–418.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., *et al*. (2011). Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830.

Qader, W. A., Ameen, M. M., and Ahmed, B. I. (2019). An overview of bag of words; importance, implementation, applications, and challenges. In *2019 international engineering conference (IEC)*, pages 200–204. IEEE.

Sandulescu, V. and Ester, M. (2015). Detecting singleton review spammers using semantic similarity. In *Proceedings of the 24th international conference on World Wide Web*, pages 971–976.

Schuch, F. A. (2013). Detecção automática de spams de opinião em avaliações de produtos na língua portuguesa.

Shah, R., Khemani, V., Azarian, M., Pecht, M., and Su, Y. (2018). Analyzing data complexity using metafeatures for classification algorithm selection. In *2018 Prognostics and System Health Management Conference (PHM-Chongqing)*, pages 1280–1284. IEEE.

Shuyo, N. (2010). Language detection library for java.

SIA, W. G. (2013). Webscraper.

Sihombing, A. and Fong, A. C. M. (2019). Fake review detection on yelp dataset using classification techniques in machine learning. In *2019 international conference on contemporary computing and informatics (IC3I)*, pages 64–68. IEEE.

Silverman, B. W. (1986). *Density estimation for statistics and data analysis*, volume 26. CRC press.

Sokolova, M., Japkowicz, N., and Szpakowicz, S. (2006). Beyond accuracy, f-score and roc: a family of discriminant measures for performance evaluation. In *Australasian joint conference on artificial intelligence*, pages 1015–1021. Springer.

Speech, J. D. M. J. (2009). Language processing: An introduction to natural language processing. *Computational Linguistics, and Speech Recognition*, 2.

Valdivia, A., Hrabova, E., Chaturvedi, I., Luzón, V. M., Troiano, L., Cambria, E., and Herrera, F. (2019). Inconsistencies on tripadvisor reviews: A unified index between users and sentiment analysis methods. *Neurocomputing*, 353:3–16.

Yue, W. and Li, L. (2020). Sentiment analysis using word2vec-cnn-bilstm classification. In *2020 seventh international conference on social networks analysis, management and security (SNAMS)*, pages 1–5. IEEE.