08, 09, 10 e 11 de novembro de 2022
ISSN 2177-3866

PÊGO NA MENTIRA! Desenvolvimento de Modelos Preditivos para detectar avaliações online fraudulentas de hotéis

RENATO CALHAU CODA
UNIVERSIDADE DE SÃO PAULO (USP)

ADÂMARA SANTOS GONÇALVES FELÍCIO
UNIVERSIDADE ESTADUAL DE CAMPINAS (UNICAMP)

JOSIVANIA SILVA FARIAS
UNIVERSIDADE DE BRASÍLIA (UNB)

RAFAEL DE FREITAS SOUZA
UNIVERSIDADE DE SÃO PAULO (USP)

PÊGO NA MENTIRA! Desenvolvimento de Modelos Preditivos para detectar avaliações *online* fraudulentas de hotéis

1. INTRODUÇÃO

O ambiente competitivo de negócios exige que as empresas façam melhores previsões para seus complexos cenários e reajam às mudanças celeremente. Para lidar com essa complexidade, as empresas passaram a incluir mais fontes de dados (internas / externas) em seus processos de tomada de decisão. A ampla difusão e a evolução das tecnologias digitais mudaram profundamente todos aspectos nas vidas das pessoas e tiveram um grande impacto em inúmeros setores econômicos.

Data Science e *Business Analytics* são duas tendências emergentes em nível global que, se combinadas, facilitam a identificação de oportunidades e desafios para o turismo. Isso pode ajudar a refletir como as habilidades analíticas podem sustentar um novo modelo de turismo, melhorando a experiência turística e, ao mesmo tempo, melhorando as capacidades das organizações que gerenciam turismo e hospitalidade com base em ferramentas mais sofisticadas para apoiar decisões. A *Data Science* permite o desenvolvimento de modelos de análise que promovam simultaneamente uma visitação de qualidade e permite um melhor monitoramento e gestão desses espaços, adotando métricas que, ao invés de apenas contar o número de visitas, avaliam objetivamente o valor (co)criado por cada turista que visita o destino (Rita & Rita, 2018).

Preços dinâmicos, reservas *online* e sistemas de recomendação baseados em conteúdo gerado pelo usuário (*User-Generated Content* ou UGC) dominaram o setor de turismo e, como resultado, transformaram a maneira como os indivíduos procuram a experiência de suas viagens (Raguseo, Neirotti & Paolucci, 2017; Sparks, So & Bradley, 2016). O UGC, em particular, é visto como espontâneo, perspicaz e com um *feedback* apaixonado "fornecido por clientes amplamente disponível, gratuito ou de baixo custo, e facilmente acessível em qualquer lugar, a qualquer hora" (Guo, Barnes & Jia, 2016, p. 468), e permite que aos consumidores descrever, reviver, reconstruir e compartilhar suas experiências.

Consequentemente, o UCG fornece uma oportunidade para experiências indiretas e, portanto, para desenvolver (ou encerrar) a longo prazo o relacionamento *online* e a lealdade do consumidor. Embora as avaliações dos clientes sejam pouco estruturadas, sendo mais ou menos focadas em uma única entidade ou aspecto da hospitalidade, ou são multilíngues, são uma importante fonte de informação para acadêmicos e gerentes, que ajudam a fornecer uma compreensão completa das preferências e demandas dos hóspedes, e serve como insumo para prever se eles voltarão ou não para o meio de hospedagem. Estudos consideraram a utilização da taxa de avaliação *online* dos consumidores sobre atrações turísticas como preditores de seu comportamento futuro (Pantano, Priporas & Stylos., 2017).

Do ponto de vista metodológico, os estudos existentes têm frequentemente adotado abordagens tradicionais (por exemplo, *surveys* e entrevistas) para investigar a percepção de turistas. Como o UGC continua a se popularizar, há uma tendência crescente de compartilhamento de experiências *online* (Yu & Sun, 2019). No entanto, desenhando a partir do marketing e da ciência de dados, os profissionais de marketing muitas vezes encontram dificuldades em desvendar os significados de texto de mídia social e outas plataformas (Bharadwaj, Ballings & Naik, 2020). Logo, os avanços recentes em processamento de linguagem natural (PLN) levaram a uma mudança radical da codificação manual convencional para automação da coleta e análise de dados. Derivado do campo da inteligência artificial (IA), linguística, ciência da computação e engenharia da informação, modelagem de tópicos e análise de sentimento emergiram como as técnicas de PLN mais amplamente adotadas em marketing e

gestão (Hannigan *et al*., 2019; Reisenbichler & Reutterer, 2019) para classificar a experiência do consumidor e para detectar a polaridade dos dados de texto.

As avaliações *online* funcionam como uma forma de boca-a-boca eletrônico (*electronic word of mouth*, ou *e-WOM*), que é uma variação tecnológica do boca-a-boca tradicional. A influência de comentários *online* pode afetar até 50% de todas as decisões de reserva de hotel (Duan *et al*., 2016). A reputação social é tão importante para os provedores de serviço que existem várias empresas especializadas de coleta, análise e gestão hoteleira revisões *online* (Antonio *et al*., 2018). Torres, Singh e Robertson-Ring (2015) revelaram que os clientes estão dispostos a pagar mais por um quarto conforme as avaliações aumentam, enquanto Viglia, Minazzi e Buhalis (2016) demonstraram que um aumento de um ponto na pontuação da revisão é associado a um aumento de 7,5% na taxa de ocupação.

A previsão de uma avaliação de revisão *online* com base em componentes quantitativos e qualitativos de avaliações permite identificar discrepâncias entre esses componentes e contribui para a detecção de análises fraudulentas. Também pode ser usado na criação de um sistema de classificação proprietário que se baseia na combinação de avaliações de diferentes fontes, que podem ser aplicadas em uma análise de conjunto competitivo ou para classificar hotéis sem um sistema de classificação de estrelas. Também pode ser usado para construir sistemas de recomendação para auxiliar os usuários na seleção de um hotel com base em avaliações e descrições textuais (Antonio *et al*., 2018).

Para preencher uma lacuna sobre previsões de classificação de avaliações *online*, este estudo se propõe a aplicar PLN e aprendizado de máquina (*machine learning*) para as análises *online* de meios de hospedagem publicados na plataforma Tripadvisor, para obter um modelo de previsão para classificações de revisão.

Como objetivos específicos, delimitaram-se: analisar avaliações de turistas (na plataforma Tripadvisor; aplicar técnicas de PLN para análise e agrupamentos das avaliações de turistas em fraudulenta e não-fraudulenta e aplicar um modelo de *machine learning* para previsão dos possíveis resultados de avaliação fraudulenta e avaliação não-fraudulenta

A indústria da hospitalidade e do turismo enfrenta novos problemas que exigem domínio de tecnologia moderna para transformar dados em conhecimento e oferecer suporte não apenas reativo, mas, acima de tudo, tomada de decisão gerencial proativa. A ciência de dados fornece os dados quantitativos necessários e abordagens analíticas consideradas necessárias para enfrentar os desafios contínuos apresentados pelo macroambiente e o próprio setor. Ciência de dados e análise de negócios são duas tendências emergentes em nível global que, se combinados, constituem um facilitador para a identificação de oportunidades e desafios para o turismo. Esta pesquisa contribui com a reflexão de como as habilidades analíticas podem sustentar um novo modelo de turismo, melhorando a experiência turística e, ao mesmo tempo, melhorando as capacidades das organizações que gerenciam turismo e hospitalidade com base na utilização de ferramentas mais sofisticadas para apoiar decisões (Rita, Rita & Oliveira, 2018).

Devido à natureza cíclica dos projetos de modelagem de previsão, a estrutura desta pesquisa é ligeiramente diferente de um estudo convencional. Uma breve revisão da literatura será seguida por uma descrição detalhada dos dados e da metodologia aplicada no modelo. Contudo, os resultados serão apresentados intercalados com a metodologia, pois precisam ser avaliados em cada ciclo para determinar a próxima etapa de execução do modelo.

2. REVISÃO BIBLIOGRÁFICA

2.1 *E-wom*

A comunicação *e-WOM*, que consiste na evolução da comunicação boca a boca em canal *online*, é reconhecida como ferramenta de marketing útil para construir relacionamentos com

consumidores, gerando consciência e interesse em determinados produtos e influenciando o comportamento de compra do consumidor (Vazquez, Dennis & Zhang., 2017). O *e-WOM* é baseado na geração, distribuição e recuperação de informação *online* de consumidores (UGC). Uma vez que esta informação é gerada por outros consumidores, ela é considerada mais confiável do que as informações geradas pelos prestadores de serviço (Tran, Strutton & Taylor, 2012). A mídia social aprimorou esse processo de compartilhamento de informações, dando aos consumidores a possibilidade de conversar em tempo real, por exemplo, por meio da criação de palavras de microblogging, que aumenta a velocidade com a qual os dados podem ser trocados (Hennig-Thurau, Wiertz & Feldhaus, 2015).

Os consumidores estão mais dispostos a avaliar voluntariamente os produtos *online* em serviços complexos, como turismo, hospitalidade, alimentação etc. Os indivíduos também analisam as avaliações, concentrando-se não apenas no resumo das avaliações com estrelas, mas também no conteúdo do texto de formato livre dos clientes, com base em recursos intangíveis experimentados subjetivamente (Serra-Cantallops & Salvi, 2014). Como as análises *online* têm grade impacto nos consumidores potenciais, as experiências de serviço podem ser co-criadas / co-destruídas pelo *e-WOM*, em que *e-WOM* positivo pode melhorar a lealdade do cliente e levar à promoção de marketing de produtos e serviços, enquanto o *e-WOM* negativo pode levar à desistência de um consumidor no momento da escolha de sua hospedagem. Porém, alguns *e-WOM* negativos podem ser por fruto de avaliações fraudulentas (Celuch, 2021).

Sanchez-Franco, Cepeda-Carrion e Roldán (2018) analisaram uma amostra de 47.172 avaliações de 33 hotéis urbanos localizados em Las Vegas (EUA) e registrado no Yelp, um agregador de avaliações de conteúdo relacionado a viagens, como TripAdvisor e Trivago. A pesquisa concentrou-se no pré-processamento do conjunto de dados para compreender a estrutura do corpus de avaliação do hotel, identificação dos tópicos relacionados à experiência do hóspede, examinando a estrutura semântica subjacente e reduzindo o número de tópicos em agrupamentos significativos e, subsequentemente, forneceu uma representação explícita de avaliações de hotéis que poderiam prever o desenvolvimento de longo prazo e relacionamentos mutuamente benéficos com os consumidores.

2.2 PLN, UGC avaliações *online*

O comportamento humano evoca uma reação funcional específica de pessoas ou grupos de indivíduos a um estímulo interno e externo. Isto inclui uma série de todas as ações físicas, bem como emoções observáveis relacionadas com os indivíduos. O objetivo da análise do comportamento humano é compreender o significado por trás das ações humanas, que às vezes são inconscientes, formulá-los e, eventualmente, fornecer suporte para seres humanos. O comportamento humano pode ser examinado em várias situações e em diferentes lugares na internet, como mídias sociais, marketing *online* etc. (Abdar & Yen, 2017).

Relativamente à UGC (User-Generated Content), sua análise pode revelar os sentimentos e experiências dos clientes (Buhalis & Sinarta, 2019), que podem ser esquecidos pelos provedores de serviço, promovendo, assim, a transformação de inteligência em turismo e hospitalidade (Shafqat & Byun, 2020). O avanço da ciência de dados provocou uma tendência crescente no uso de análise textual para revelar as percepções dos clientes (Celuch, 2021). Devido à natureza não estruturada das análises *online*, os estudiosos do turismo adotaram o PLN como uma ferramenta emergente para obter *insights* mais profundos sobre uma grande quantidade de dados de texto (Amado *et al*., 2018). Ou seja, as máquinas são capazes de ler, compreender e derivar significados das línguas humanas. Sob o pretexto de análise de texto, modelagem de tópicos e a análise de sentimentos são atualmente as abordagens dominantes usadas na vanguarda dos contextos de turismo. A primeira identifica padrões tópicos ocultos, enquanto a segunda extrai informações subjetivas (por exemplo, o tom de um frase) a partir de informações textuais (Yu & Zhang, 2020).

No caso de plataformas de avaliação *online*, Xiang *et al*. (2017) investigaram comentários postados em TripAdvisor, Expedia e Yelp, nos quais descobriram que os consumidores costumam discutir o serviço básico, valor, ponto de referência e atração, experiências gastronômicas e a atividade principal do hotel durante o encontro de serviço. Da mesma forma, outro estudo aproveitou o método de modelagem de tópicos para revelar fatores cruciais que influenciam as percepções dos passageiros em relação à qualidade do serviço no segmento de aviação (Korfiatis *et al*., 2019). No entanto, notavelmente, compreender os vários aspectos da experiências não são suficientes.

Os estudiosos perceberam a necessidade de avançar quantitavamente nos resultados ao incorporar os sentimentos dos consumidores por meio da análise do PLN. A sinergia de mineração de opinião e classificação de tópicos geram *insights* ainda mais abrangentes por meio da quantificação computacional de revisões *online* (Vu *et al*., 2019). Por exemplo, construída sobre os resultados da modelagem de tópicos, o PLN permite a pesquisadores a captura de uma visão geral sobre os aspectos positivos e negativos da experiência do consumidor em hotéis, restaurantes e qualquer outra entidade de serviço (Yu & Zhang, 2020).

3. MATERIAIS E MÉTODOS

Os meios de hospedagem são compreendidos como empreendimentos que prestam "[...] serviços de alojamento temporário, ofertados em unidades de frequência individual e de uso exclusivo do hóspede, bem como outros serviços necessários [...], mediante adoção de instrumento contratual, tácito ou expresso, e cobrança de diária" (Brasil, 2008). Partindo desta perspectiva, para obter uma visão mais geral, serão considerados comentários de quatro tipos de diferentes meios de hospedagem – hotel, pousada, *resort* e *flat* – de acordo com o Sistema Brasileiro de Classificação de Meios de Hospedagam (SBClass) do Ministério do Turismo (Mtur, 2018).

A proliferação do vírus SARS-COV-2, resultando na pandemia de Covid19, vem reafirmando, cada vez mais, a urgência com relação aos riscos crescentes à sobrevivência humana, o que coloca em xeque as engrenagens econômicas nos moldes vigentes centrados em falsas certezas, como a inexorabilidade dos recursos naturais e a convicção de que tudo e todos estão, permanentemente, sob controle. Ao contrário, a pandemia trouxe para reflexão um importante alerta: a realidade é permeada por incertezas que caracterizam a sociedade contemporânea. E, nessa perspectiva, o turismo talvez tenha sido o fenômeno contemporâneo que melhor tenha traduzido esse debate (Irving *et al*., 2020).

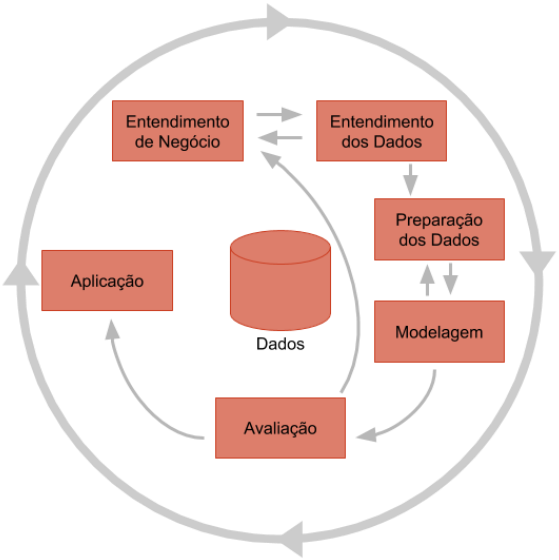
Neste contexto, foram utilizados comentários dos 25 meios de hospedagens brasileiros melhores ranqueados no Tripadvisor. Como delimitação temporal, optou-se pelo marco do advento da pandemia, período entre Janeiro de 2020 a Junho de 2022.

Machine learning (ML) contempla implementação de técnicas para definir os métodos e algoritmos utilizados para mineração de dados, com o objetivo de extrair padrões, correlações e conhecimento de fontes de dados aparentemente não estruturadas (Mariani *et al*., 2018). O ML consiste em aprender algoritmos que descobrem regras gerais ou padrões em grandes volumes de conjunto de dados, filtrando com base em várias variáveis, agrupando uma grande coleção de objetos em um pequeno número de classes. Sejam supervisionadas (os resultados surgem do treino de algoritmos com dados pré-rotulados) ou não-supervisionadas (os algoritmos derivam dos resultados dos próprios dados), estas técnicas visam dar ao computador a capacidade de realizar uma tarefa usando abordagens generalizadas sem ser explicitamente programado para tarefas únicas (Mariani *et al*., 2018).

Para esta pesquisa, foi adotado o modelo sugerido por Chapman *et al*. (2000) para mineração de dados (*Cross Industry Standard Process for Data Mining* ou CRISP-DM) na construção modelos de previsão a partir da coleta de dados e seleção de recursos para o conjunto de dados, criação de modelo de desenvolvimento e avaliação. Devido à sua praticidade e

simplicidade, o CRISP-DM é usado em muitos campos diferentes, incluindo turismo e hospitalidade (Antonio *et al*., 2018). O CRISP-DM define as cinco etapas necessárias para construir um modelo de previsão: compreensão do negócio e dos dados, preparação de dados, modelagem, avaliação e desdobramento, e desenvolvimento. O modelo pode ser conferido na Figura 1.

Figura 1 – Fases do modelo CRISP-DM



Fonte: Adaptado de Chapman *et al*. (2000).

Para a fase de compreensão do negócio e dos dados, as classificações requerem a definição de critérios de sucesso e um método para medi-los. Portanto, esta pesquisa adotou o*cutoff* - um ponto de corte que o pesquisador escolhe, definido para que sejam classificadas as observações em função das suas probabilidades calculadas e, desta forma, é utilizado quando há o intuito de se elaborarem previsões de ocorrência do evento para observações não presentes na amostra com base nas probabilidades das observações presentes na amostra.

Esta pesquisa focalizou-se em 25 hotéis localizados no Brasil, vendedores do "Traveller's Choice – Os melhores dos melhores de 2022", em votação realizada pelos próprios viajantes, que experimentam e compartilham as suas experiências. A distribuição, por classificação e frequência de notas (de um a cinco), é apresentada na Tabela 1.

Tabela 1 – Resumo de dados dos hotéis

Avaliação (qtd de estrelas)	**N**	**(%)**
Uma estrela	60	0,7
Duas estrelas	75	0,8
Três estrelas	196	1,5
Quatro estrelas	728	6
Cinco estrelas	11590	91
Total	12.649	100

Fonte: Elaboração própria.

Para realizar a “raspagem” dos dados, o extrator da *web* (*crawler*) teve de simular o comportamento humano, para evitar revisões traduzidas por máquina e configurado para extrair apenas avaliações em língua portuguesa. Portanto, depois de entrar na página do *site* de reservas do hotel, o extrator usou um procedimento totalmente automatizado para clicar no *link* das avaliações de visitantes, e, em seguida, clicou na caixa de seleção "Mostrar avaliações” Isso então selecionou a lista suspensa "Classificar por" e clicou em "Data (mais recente para mais antiga)".Os extratores coletaram avaliações entre janeiro de 2020 e junho de 2022. Durante este período, foram consideradas todas as avaliações realizadas nos 25 hotéis, com um total de 12.649 avaliações. Notas quatro e cinco foram consideradas avaliações positivas e de três a um foram consideradas avaliações negativas.

Quando o conjunto de dados é desbalanceado, isto é, as classes não estão representadas de maneira equivalente ou similar, o classificador pode ignorar a importância das classes minoritárias e tender a fazer mais predições para as classes majoritárias, desta forma ocasionando um baixo desempenho de classificação de exemplos que pertençam às classes minoritárias.Métodos baseados em amostragem (*sampling*, ou *resampling*) são os métodos mais simples e mais antigos utilizados para o balanceamento de conjuntos de dados. Eles consistem na modificação da estrutura do conjunto de dados desbalanceado, de maneira a deixá-lo com quantidades equivalentes de amostras para as classes presentes, seja através da remoção (*undersampling*) ou adição (*oversampling*) de novas amostras (Lunardon *et al*., 2014). Esta pesquisa utilizou-se do método *Random Over-Sampling Examples* (ROSE) para balanceamento da amostra.

O desenvolvimento de um modelo de previsão envolve selecionar, mesclar, limpar, codificar e construir *features* a partir do conjunto de dados original. Esse processo é conhecido como “engenharia de recursos” e pode ter um impacto significativo no desempenho do modelo (Kuhn & Johnson, 2013). Os recursos resultantes originam um conjunto de dados de modelagem que é usado para construir e avaliar o modelo de previsão. Desenvolver e encontrar o melhor modelo de previsão é um processo iterativo que muitas vezes requer voltar no *pipeline*. Por exemplo, o conjunto de recursos mais adequado geralmente é alcançado por meio da realização de experimentos parciais, às vezes exigindo que voltemos ao estágio de entendimento do negócio para derivar novos recursos.

Para a segunda fase, preparação dos dados, o desenvolvimento de um modelo de previsão envolve a seleção, fusão, limpeza, codificação e construção de *features* do conjunto de dados original. A modelagem de previsão requer um conjunto de dados bidimensional, que é composto de linhas e colunas. Cada linha representa a unidade de análise, enquanto as colunas representam atributos, descritores ou variáveis (Abbott, 2014). Estas *features* foram criadas aplicando a normalização mín-máx, que é um dos métodos de normalização mais comuns utilizados para dimensionar variáveis (Abbott, 2014).

Após o pré-processamento do texto, os dados foram transformados em uma representação de saco de palavras (*bag of words*), que é um dos métodos mais populares usados na mineração de texto (Antonio et al., 2018). Este método resultou na criação de duas matrizes documento-termo, uma por documento (revisão) e outra por frases (frases em comentários). Cada matriz consiste em linhas, cada uma das quais representa um documento e colunas, sendo uma representando uma frequência de palavra (ou seja, todas as palavras apresentadas em um documento). Essas matrizes produziram as informações para criar dois novos conjuntos de características de dados, a saber: número de palavras por revisão e número de frases por revisão.

Existem várias abordagens para a aplicação da análise de sentimento. Nesta pesquisa, optou-se por utilizar uma abordagem baseada na polaridade léxica. Esta abordagem se baseia em dicionários de palavras de opinião com classificação de polaridade (Ravi e Ravi, 2015). Portanto, a seleção do dicionário é uma consideração metodológica importante. Um aspecto relevante na escolha do dicionário é sua adequação ao domínio do texto. Com base neste

critério, o léxico de sentimento SentiLex-PT 02 (Silva et al., 2012) e e Oplexicon_v3.0 (Souza & Vieira, 2012) foram selecionados.

Por fim, cada *feature* foi determinada por meio de duas técnicas utilizadas na criação de características para modelagem de previsão de dados, que são frequência de termo (TF) e frequência de documento de frequência inversa de termo (TF-IDF). TF é uma estatística numérica que representa a frequência com que cada termo é utilizado em um documento. TF-IDF também é uma estatística numérica, mas representa o peso composto de cada termo em um documento. Os termos podem ser uma única palavra (também conhecido como unigrama ou n-grama) ou podem ser uma sequência contígua de n palavras em um texto (Liu & Zhang, 2012). Selecionar o algoritmo apropriado para predição é uma das etapas mais importantes para avaliação da predição. De acordo com a literatura (Antonio *et al*., 2018), os modelos mais utilizados para este tipo de teste são: Regressão logística, árvore de decisão impulsionada (*boosted decision tree*), floresta de decisão (*decision forest* ou *random forest*) e redes neurais. A árvore de decisão é uma das técnicas de classificação mais populares e comuns, e uma das razões mais importantes por trás de sua popularidade é o fato de ser flexível e de fácil compreensão. (Khodabandehlou & Rahman, 2017).

As árvores de decisão são classificadores baseados em regras que classificam amostras por respondendo consecutivamente perguntas das *features* de entrada. A classificação final de uma nova amostra é concluída assim que todas as questões forem divididas forem respondidas, começando do nó raiz da árvore até um nó de folha. Depois disso, a nova amostra é atribuída à classe que parecia ser mais frequente nas amostras de treinamento que terminaram neste desdobramento de folha.

Random forests são executadas com os seguintes procedimentos (Shmueli *et al*., 2017): desenho de várias amostras aleatórias, com reposição dos dados pela abordagem de amostragem *bootstrap*, utilização de um subconjunto aleatório de preditores em cada estágio, ajuste de uma classificação (ou regressão) para cada amostra (e assim obter uma “floresta”), e combinação de classificações das árvores individuais para obter melhores previsões.

A regressão logística é uma ferramenta estatística que pode descrever e testar hipóteses sobre as relações entre um resultado categórico e variáveis preditoras contínuas. O método tem como objetivo principal estudar a probabilidade de ocorrência de um evento definido por Y que se apresenta na forma qualitativa dicotômica (Y = 1 para descrever a ocorrência do evento de interesse e Y = O para descrever a ocorrência do não evento), com base no comportamento de variáveis explicativas (Peng *et al*., 2002).

Para a análise da eficiência global do modelo (EGM), que corresponde ao percentual de acerto da classificação para um determinado *cutoff*, foram utilizados os parâmetros de sensitividade e especificidade. A sensitividade diz respeito ao percentual de acerto, para um determinado *cutoff*, considerando-se apenas as observações que, de fato, são evento. A especificidade, por outro lado, refere-se ao percentual de acerto, para um dado cutoff, considerando-se apenas as observações que não são evento.

Estas análises de sensibilidade podem ser feitas com qualquer valor de *cutoff* entre 0 e 1, o que permite que o pesquisador possa tomar uma decisão no sentido de definir um *cutoff* que atenda aos seus objetivos de previsão. Se, por exemplo, o objetivo for o de maximizar a eficiência global do modelo, pode ser utilizado um determinado *cutoff* que poderá gerar valores de sensitividade ou de especificidade não maximizados. Se, por outro lado, o objetivo for o de maximizar a sensitividade, ou seja, a taxa de acerto para aqueles que são evento, poderá ser definido outro *cutoff* que não necessariamente aquele que maximizará a eficiência global do modelo (Fávero & Belfiore, 2017).

Matrizes de confusão são utilizadas para determinar a diferença entre os valores previsto e o verdadeiro. Essa matriz corresponde a uma tabulação cruzada entre a classificação de acordo

com o algoritmo e a classificação real observada na amostra. Precisão, Recall e F1-score são usados como métricas de avaliaçãopara o desempenho do modelo.

Tabela 2 – Matriz de confusão

Classe	Positiva	Negativa
Positiva	Verdadeiro positivo (VP)	Falso negativo (FN)
Negativa	Falso positivo (FP)	Verdadeiro negativo (VN)

Fonte: Matlatipov *et al*. (2022).

A acurácia corresponde ao percentual de casos que são corretamente classificados. Precisão corresponde ao percentual de observações que foram corretamente classificadas. Portanto, há um valor de precisão para cada classe (0 ou 1), sendo a de casos verdadeiros a sensitividade e especificidade para os casos falsos. Recall, para a classe C: corresponde ao percentual de previsões da classe C que foram corretamente classificadas e F-1 corresponde a uma combinação entre precisão e recall (Matlatipov *et al*., 2022).

Para representação gráfica da análise de sensibilidade para fins comparativos entre os modelos utilizados, foi utilizada a curva ROC (*Receiver Operating Characteristic*), um gráfico que apresenta a variação da sensitividade em função de (1 - especificidade). Por meio da curva de sensibilidade, verifica-se que é possível definir o *cutoff* que iguala a sensitividade com a especificidade, ou seja, o *cutoff* que faz com que a taxa de acerto de previsão para aqueles que serão evento seja igual à taxa de acerto para aqueles que não serão evento.

Esta avaliação do melhor modelo preditivo foi realizada por meio da validação cruzada *k-fold*, que é uma técnica bem conhecida e amplamente utilizada para avaliação de modelos preditivos, pois permite o desenvolvimento de modelos que acabaram não ajustados e pode ser generalizado para conjuntos de dados independentes (Fávero & Belfiore, 2017). A *k-fold* funciona por meio do particionamento aleatório dos dados de amostra em subamostras de tamanho k. Nesta pesquisa, o conjunto de dados foi dividido em 10 dobras, número sugerido por Smola e Vishwanathan (2008). Em seguida, cada uma das 10 dobras será utilizada como um conjunto de teste e as nove restantes serão utilizadas como dados de treinamento. Optou-se pela ferramenta R para criação do conjunto de dados.

4. ANÁLISE DOS RESULTADOS

A modelagem de previsão requer um conjunto de dados bidimensional composto de linhas e colunas. Cada linha representa a unidade de análise, enquanto as colunas representam atributos, descritores ou variáveis (Abbott, 2014). Optou-se pela ferramenta R para criação do conjunto de dados. Os campos coletados foram: *Ranking* do Hotel (de acordo com o Traveller's Choice), nome do hotel, nota da avaliação, título da avaliação, texto da avaliação e data da avaliação (mês e ano).

Antes de aplicar a análise de sentimentos às descrições das avaliações para extrair possíveis *features* adicionais, o passo mais crítico na transformação do texto de uma forma não estruturada para uma forma estruturada é o pré-processamento (Han *et al*., 2016). Este processo permite a retenção de informações relevantes e a remoção de informações irrelevantes. O procedimento foi realizado por meio da aplicação das seguintes etapas de pré-processamento, algumas das quais são dependentes do idioma: (a) Transformação de todos os textos das avaliações em minúsculas; (b) *Stemming* (redução) de palavras comuns de hospitalidade, como "quartos", "restaurantes" e outras que possam ser significativas para a interpretação dos dados; (c) Normalização dos termos utilizados para escrever algumas palavras ou expressões que poderiam ser escritos de forma diferente ou com erros ortográficos; por exemplo, "vc" e "você" como o mesmo *token*; (d) Normalização dos termos usados para escrever algumas palavras ou expressões que poderiam ser escritas de forma diferente ou com alteração de gênero; por

exemplo, “ótimo”, “ótimos”, “ótima” e “ótimas” e (e) Remoção de pontuação, números e *stop words*.

Após o pré-processamento do texto, os dados foram transformados em uma representação de saco de palavras (*bag of words*), que é um dos métodos mais populares usados na mineração de texto. Este método resultou na criação de duas matrizes documento-termo, uma por documento (avaliação) e outra por frases (frases de avaliações). Cada matriz consiste em linhas, cada uma das quais representa um documento e colunas, cada uma representando uma frequência de palavras (ou seja, todas as palavras apresentadas em um documento). Essas matrizes produziram as informações para criar dois novos conjuntos de dados características, a saber, número de palavras por avaliação e número de frases por avaliação.

Então, a análise de sentimento foi aplicada ao componente de texto pré-processado das avaliações, para obter recursos adicionais do conjunto de dados. A polaridade da força do sentimento foi tomada sob duas perspectivas distintas: por documento (texto completo da descrição da avaliação), de forma semelhante o relatado por Han *et al*. (2016) e Bjørkelund *et al*. (2012); e por frase, seguindo o trabalho de Duan *et al*. (2016). Enquanto a primeira permite a compreensão da opinião global da avaliação, a segunda relaciona a opinião a aspectos particulares.

Na literatura, existem diversas abordagens para a aplicação da análise de sentimentos. Neste estudo, optou-se por utilizar uma abordagem baseada no léxico de polaridade. Essa abordagem se baseia em dicionários de palavras de opinião com classificação de polaridade (Ravi & Ravi, 2015). Portanto, a escolha do dicionário a ser utilizado é aspecto metodológico importante (Han *et al*., 2016). Com base neste critério, foi selecionado o léxico de sentimentos para a língua portuguesa SentiLex-PT 02 (Silva et al., 2012) e Oplexicon_v3.0 (Souza & Vieira, 2012). A força do sentimento de cada sentença foi calculada contando-se os pontos positivos e palavras negativas de cada frase e, em seguida, aplicando-se a fórmula usada por Bjørkelund *et al*. (2012):

$$força\ do\ sentimento = \frac{\Sigma\ palavras\ positivas}{\Sigma palavras\ positivas + \ \Sigma palavras\ negativas}$$

Esta fórmula resulta em um valor entre -1 e 1, onde -1 é perfeitamente negativo e 1 é perfeitamente positivo. A força do sentimento global de cada revisão foi então calculada como a média força do sentimento de todas as frases de revisão. Com isso, foi possível detectar a avalição com o sentimento mais positivo:

> “Como sempre espetacular. Funcionários de alto nível profissional. Hotel aconchegante. Bonito, charmoso, atual e extremamente mais uma vez e como sempre perfeito. Café da manhã excepcional e o chá da tarde maravilhoso com uma música tranquila ambiente com um pianista, tudo com muito super tranquilo e muito familiar. Nível máximo na avaliação. Sempre que vou à Gramado é minha opção de tranquilidade e paz. Já é minha 6a. estadia. Vou voltar sempre que for à Gramado. MARAVILHOSO em todos os quesitos de um hotel 5 estrelas. PARABÉNS SEMPRE !!”

Da mesma forma, também foi possível detectar a avaliação com o sentimento mais negativo dentre todas as avaliações presentes na coleta de dados:

> “Nossa quarta visita ao (hotel) foi extremamente frustrante e decepcionante. A começar pelos valores diferenciados praticados para quem é daqui do Sudeste e as facilidades para os “regionais”, nos sentimos explorados! A qualidade dos serviços prestados é de total decadência comparados aos anos anteriores (fomos em 2018, 2019 e 2021). A desculpa usada da pandemia não se emprega, já que fomos em fevereiro do ano passado, dentro da pandemia e ainda tinha um pouco de qualidade.... Esse negócio de usar um porta-talheres de couro e não de papel descartável como era antigamente é de extremo mau gosto e sem higiene! Aliás cadê o guardanapo timbrado de antigamente? Esse guardanapo oferecido atualmente é péssimo e sem qualidade nenhuma!”

Após retirada dos comentários neutros (sentimento=0), conversão de numéricos para categóricos (1=positivo e -1=negativo) e novo agrupamento dos dados, foi possível classificar as avaliações como fraudulentas e não-fraudelentas, em que a contradição entre avaliação positiva e sentimento negativo (falso-positivo) e seu oposto, avaliação negativa e sentimento positivo (falso-negativo) foram classificadas como avaliações fraudulentas. Os outros resultados (verdadeiro positivo e verdadeiro negativo) foram classificados como avaliações não-fraudulentas, conforme esquematizado na Tabela 2.

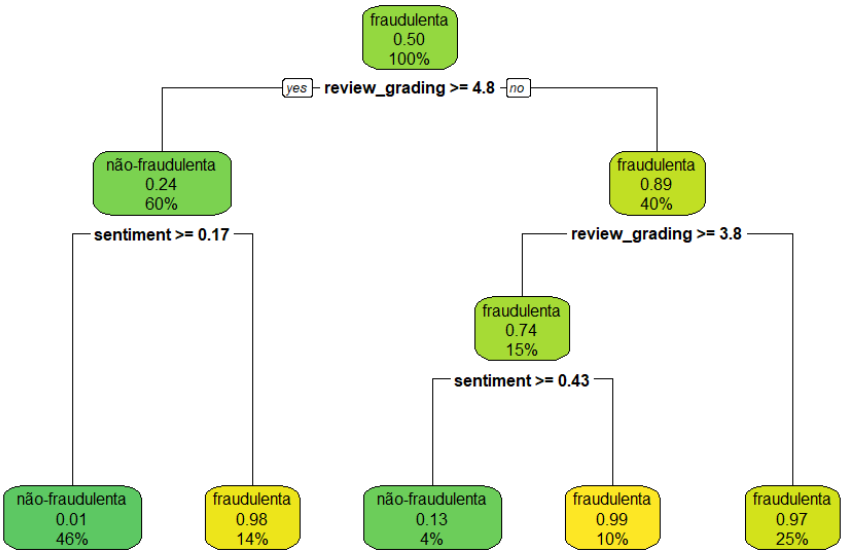
Tabela 3 – Classificação das avaliações em fraudulentas e não-fraudulentas

Avaliação (Nota)	Sentimento	Resultado	Classificação
Positiva (4 e 5)	1	Positivo verdadeiro	Não-fraudulenta
Negativa (1 a 3)	1	Falso-negativo	Fraudulenta
Positiva (4 e 5)	-1	Falso-positivo	Fraudulenta
Negativa (1 a 3)	-1	Negativo verdadeiro	Não-fraudulenta

Fonte: Elaboração própria.

Neste ponto, as *features* apresentadas no conjunto de dados usado para construir o modelo preditivo foram: Avaliação do meio de hospedagem (numérico), Número de palavras por avaliação (numérico), Número de sentenças por avaliação (numérico), Número de palavras positivas (númerico), Número de palavras negativas (numérico) e Sentimento (numérico). Os modelos escolhidos para predição foram árvore de decisão (*decision tree*), *random forest* e regressão logística. A Figura 2 demonstra o resultado alcançado do primeiro modelo.

Figura 2 – Árvore de decisão como modelo preditivo das avaliações do Tripadvisor



Fonte: Elaboração própria.

A primeira pergunta exibida no nó raiz da árvore é se *feature* de entrada "Avaliação do meio de hospedagem" (*review_rating*) é maior ou igual a 4,8. Se uma amostra preencher esta condição, então podemos continuar para o ramo esquerdo da árvore com a condição de entrada *sentiment*, sendo maior ou iguam a 0,17. Este procedimento continua até chegarmos a um nó folha. Cada nó folha está associado à classe final, de fraudulenta ou não-fraudulenta.

Para a análise de desempenho dos modelos, foram utilizados critérios empíricos de desempenho do em termos de sua precisão preditiva, ou seja, taxas de erro, estatísticas AUROC (*area under the receiver operating characteristic curve*) e elevações cumulativas (Baesens *et al*., 2002). A precisão geral da previsão, como a taxa de erro simples ou, inversamente, a porcentagem de classificação correta (PCC), é uma medida convencional de desempenho para classificadores. A taxa de erro simples de um classificador geralmente é explicada em uma matriz de confusão. Com base na tabela, taxa de erro simples = (falsos negativos (FN) + falsos positivos (FP))/total de registros. Em outras palavras, o PCC é: 1 – taxa de erro simples.

Além da taxa de erro simples, outras medidas de precisão de classificação oferecem mais percepção do problema. A sensibilidade é a precisão entre as os casos positivos e especificidade entre os negativos. Sensibilidade e especificidade superam os lados negativos da precisão da classificação. Sensibilidade e especificidade são, respectivamente, definidas como

$$sensitividade = \frac{VP}{(VP + FN)} = taxa\ de\ positivos\ verdadeiros$$

$$especificidade = \frac{VN}{(VN + FP} = taxa\ de\ negativos\ verdadeiros$$

Fonte: Cui *et al*., (2008).

O *cutoff* serve para que o pesquisador avalie a real incidência do evento para cada observação e a compare com a expectativa de que cada observação incida, de fato, no evento. Com isto feito, será possível avaliar a taxa de acerto do modelo com base nas próprias observações presentes na amostra e, por inferência, assumir que tal taxa de acerto se mantenha quando houver o intuito de avaliar a incidência do evento para outras observações não presentes na amostra, que é o valor previsto (Fávero & Belfiore, 2017). Deifniu-se um *cutoff* de 0,6 para definição da EGM, ponto de encontro no cruzamento entre as linhas de sensibilidade e especificidade.

A maioria dos classificadores simplesmente minimiza o número total de classificações incorretas. No entanto, os custos desiguais de classificação incorreta têm um impacto significativo na precisão da previsão, especialmente para a classe menor. Logo, pesquisadores precisam considerar os custos de erros de classificação na seleção do modelo.

Na validação do *k-Fold* (*cross-validation*), as seguintes etapas foram executadas: divisão do conjunto de treinamento em N partições, treinamento do modelo em N-1 partições e teste em relação à partição restante, repetição da etapa anterior iterativamente e, a cada iteração, alteração do valor restante da partição. Uma vez que o processo foi repetido N vezes (10 no caso desta pesquisa), os resultados obtidos de cada iteração são posteriormente calculados para obtenção da pontuação final.

Tabela 3 – Resultados dos modelos preditivos testados – Análise de Robustez

Modelo	Espec.	Sens.	Rec.	F1	Acc	AUC
Árvore de decisão	97%	97%	97%	49%	97%	98%
Random Forest	99%	99.5%	100%	33%	97%	99%
Regressão Logística	95%	97%	96%	96%	96%	98%

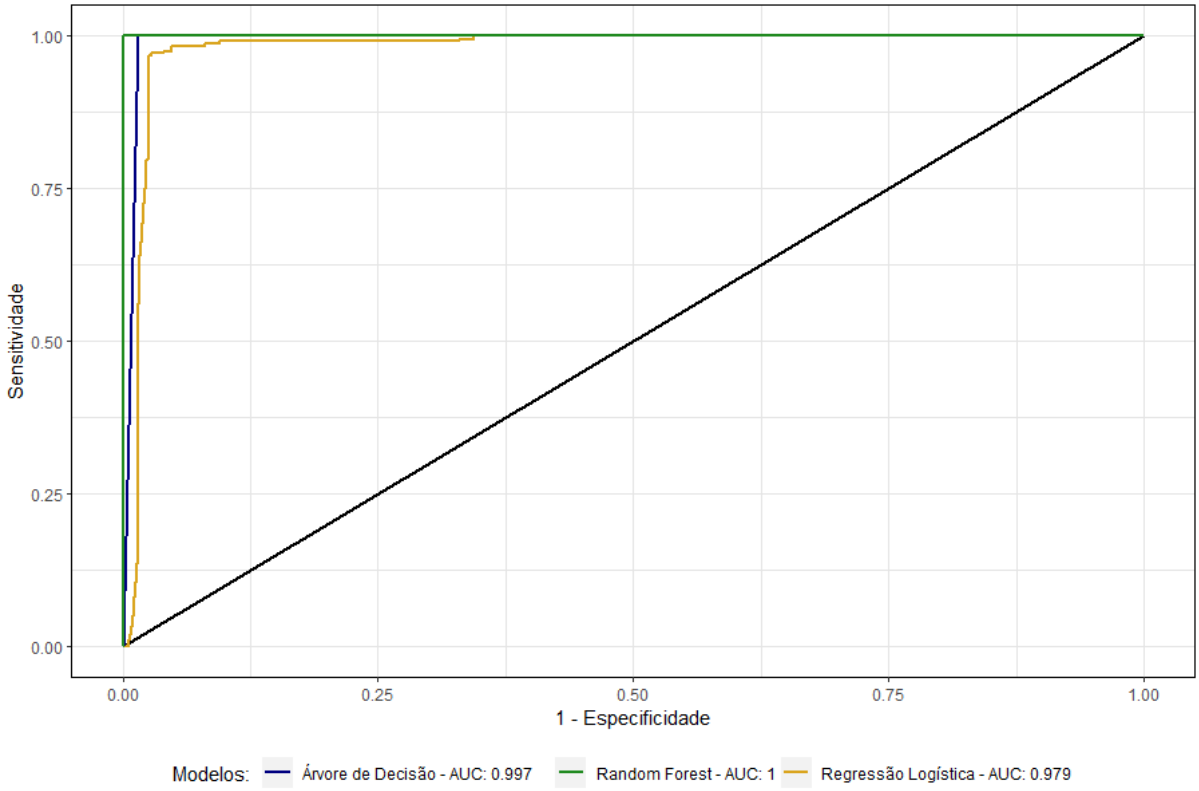
Fonte: Elaboração Própria.

Duas das principais vantagens do uso de validações cruzadas são, primeiro, que o *overfitting* dos dados de treinamento são geralmente evitados e, segundo, a melhor combinação

de hiperparâmetros é selecionado para fornecer resultados duradouros relevantes, uma vez que o modelo for implantado (Ippolito, 2022).

A curva ROC mostra o comportamento propriamente dito do *trade off* entre sensitividade e especificidade e, ao trazer no eixo das abscissas os valores de (1 - especificidade), apresenta formato convexo em relação ao ponto (0, 1). Desta forma, um determinado modelo com maior área abaixo da curva ROC (*Area Under Curve* - AUC) apresenta maior eficiência global de previsão. Em outras palavras, é possível, por exemplo, incluir novas variáveis explicativas na modelagem. A comparação do desempenho global dos modelos poderá ser elaborada com base na área abaixo da curva ROC, já que, quanto maior a sua convexidade em relação ao ponto (0,1), maior a sua área e, consequentemente, melhor performance do modelo para fins preditivos (Fávero & Belfiore, 2017).

Figura 3 – Curva ROC comparativa entre os modelos preditivos



Fonte: Elaboração própria

A análise ROC é geralmente resumida graficamente traçando uma curva em um plano bidimensional onde a taxa de verdadeiros positivos (TP) (sensibilidade) é plotado ao longo do eixo y e a taxa de FP (1 – especificidade) ao longo do eixo x, enquanto o ponto de operação (corte) é variado. Na prática, um ponto na curva ROC reflete a taxa deTP com a taxa de FP (alarme falso) em um determinado ponto de operação (valor limite), assim pode ser considerado como uma razão dessas duas medidas.

Outra estatística útil que resume o desempenho de um determinado modelo é o AUROC, ou a área acima da linha diagonal (Figura 3). Um modelo que perfeitamente discrimina entre o TP e FP terá um índice de área igual a 1,0 e um modelo sem poder discriminatório resultará em um índice de área de 0,50. Assim, o modelo de melhor performance é aquele que engoloba a maior área AUC. No caso dos modelos testados, a *random forest* obteve o melhor AUC, com 100 %, enquanto os modelos árvore de decisão e regressão lógística obtiveram 99% e 97%, respectivamente.

CONSIDERAÇÕES FINAIS

Dada a importância das avaliações *online* para o segmento de turismo, este estudo contribui para a pesquisa sobre o impacto da reputação social do setor. No caso, a combinação de técnicas de aprendizado de máquina e processamento de linguagem natural permitiu a previsão das classificações de avaliações *online*. Em primeiro lugar, este trabalho mostrou que os recursos quantitativos das avaliações podem ser complementados com recursos de análise de sentimentos, extraído do componente textual das avaliações, o que leva a um modelo aprimorado de predição de avaliações. Em segundo lugar, este trabalho mostrou que recursos adicionais baseados em TF-IDF também contribuem para modelos de previsão aprimorados. Em suma, este estudo foi capaz de mostrar que a previsão de avaliações *online* são possíveis.

Estes resultados na previsão de notas de avaliações abrem caminho para muitas aplicações diferentes para esses modelos. Uma aplicação óbvia é prever como um consumidor avaliaria quantitativamente uma estada em um hotel com base em seus comentários de texto. Por exemplo, isso pode ser aplicado às pesquisas de *check-out* padrão que os hotéis fazem aos seus hóspedes. No entanto, a previsão de nota tem potencial para ser usada como medida e *proxy* sozinho.

A utilização de modelos preditivos não só tem potencial para ser usado em pesquisas científicas, como uma medida para substituir ou complementar as classificações de revisão, mas também podem ser usados em termos das seguintes aplicações de negócios, como na detecção de fraude ou avaliação de discrepâncias, que englobam diferenças substanciais entre os resultados previstos e as classificações reais da avaliação podem ser usado para identificar comentários em que as classificações reais estão fora dos padrões de classificação de avaliações semelhantes. Estas diferenças podem surgir por discrepâncias entre as variáveis qualitativas e quantitativas das avaliações, que podem indicar que as avaliações são fraudulentas ou focam sobre tópicos e aspectos que revisões semelhantes não fazem.

Por exemplo, em vez de exigir que o hotel determine uma equipe para analisar cada avaliação e identificar possíveis avaliações fraudulentas, um hotel pode executar este modelo periodicamente para selecionar automaticamente quais avaliações podem ser fraudulentas com base na diferença entre a classificação de avaliação prevista e a classificação real.

Em suas ferramentas de reservas *online*, as redes hoteleiras, agências de viagens ou sites de metabuscadores geralmente permitem que os consumidores filtrem suas pesquisas pelas características do hotel, preço, localização e classificações de reputação social. No entanto, essa classificação de reputação social geralmente é a classificação média de um determinado portal externo na internet. Se, em vez de usar essa classificação de reputação social, o *site* usasse uma classificação baseada em um método preditivo próprio, então recomendações de resenhas a serem lidas ou serviços/produtos a serem comprados levariam também em conta este componente qualitativo. Por exemplo, um *site* de uma rede hoteleira pode recomendar hotéis a um usuário com base no índice de dados agregados de seus próprios hotéis, em vez de usar a classificação de um determinado *site* ou outro critério externo. Este poderia ser um método mais transparente e confiável para hóspedes que desejassem selecionar hotéis por reputação social.

Embora se espere que os resultados deste estudo sejam promissores, algumas limitações precisam ser destacadas. O banco de dados continha avaliações apenas em língua portuguesa e com única fonte, que foi o Tripadvisor. Estudos futuros poderiam analiasar avaliações em diferentes idiomas e com fontes de dados de outras plataformas, como o Booking, Yelp, Airbnb e Expedia. Ademais, outros critérios de medição poderiam ser utilizados, como MAE e RMSE para comprovar a robustez do modelo preditivo, assim como comparar seus resultados com outros modelos, como regressão logística e redes neurais.

Embora este estudo não realize uma caracterização semântica dos termos, técnicas como a desambiguação do sentido da palavra permitem uma melhor compreensão de uma palavra de acordo com a frase e o contexto em que está sendo usado. O TF-IDF não é capaz de identificar gírias e ironias, por exemplo. Aplicando essas técnicas podem levar a pontuações melhores e

certamente vale a pena considerar como um possível próximo passo no avanço dos estudos de PLN. Por fim, mesmo após o balanceamento das amostras, os modelos testados tendem ao overfitting. Sugere- se que futuras pesquisas trabalhem com uma gama maior de dados, a fim de que se evidencie a perfomance de diferentes modelos preditivos.

REFERÊNCIAS

Abdar, M. and Yen, N.Y. (2017), Understanding regional characteristics through crowd preference and confidence mining in P2P accommodation rental service, *Library Hi Tech*, 35 *(4)*, 521-541.

Abbott, D.(2014) Applied Predictive Analytics: Principles and Techniques for the Professional Data Analyst, *John Wiley and Sons*, Indianapolis, IN.

Amado, A., Cortez, P., Rita, P. and Moro, S. (2018), Research trends on big data in marketing: a text mining and topic modeling based literature analysis, *European Research on Management and Business Economics*, 24 *(1)*, 1-7.

Antonio, N., de Almeida, A.M., Nunes, L., Batista, F.; Ribeiro, R. (2018), Hotel online reviews: creating a multi-source aggregated index, *International Journal of Contemporary Hospitality Management*, 30 (12), 3574-3591.

Bharadwaj, N., Ballings, M.; Naik, P.A. (2020), Cross-media consumption: insights from super bowl advertising, *Journal of Interactive Marketing*, 50 *(1)*, 17-31.

Bjørkelund, E., Burnett, T.H.; Nørvag, K. (2012) "A study of opinion mining and visualization of hotel reviews", Proceedings of the 14th International Conference on Information Integration and Web-Based Applications and Services, *ACM*, 229-238.

Brasil (2008). Lei n. 11.771, de 17 de setembro de 2008. Dispõe sobre a Política Nacional de Turismo, define as atribuições do Governo Federal no planejamento, desenvolvimento e estímulo ao setor turístico. Disponível em: <http://www.planalto.gov.br/ccivil_03/_ato2007-2010/2008/lei/l11771.htm> Acesso em 14 nov. 2021.

Buhalis, D. and Sinarta, Y. (2019), Real-time co-creation and nowness service: lessons from tourism and hospitality, *Journal of Travel and Tourism Marketing*, 36 *(5)*, 563-582.

Celuch, K. (2021), Customers' experience of purchasing event tickets: mining online reviews based on topic modeling and sentiment analysis, *International Journal of Event and Festival Management*, 12(1), 36-50.

Chapman, P. Clinton, J. Kerber, R. Khabaza, T. Reinartz, T. Shearer, C. and Wirth, R. (2000), "CRISP-DM 1.0: Step-by-step data mining guide", The Modeling Agency, available at: https://the-modelingagency.com/crisp-dm.pdf Acesso em: 14. Nove.2021.

Cui, G., Leung Wong, M., Zhang, G. Li, L. (2008), Model selection for direct marketing: performance criteria and validation methods, *Marketing Intelligence & Planning*, 26 *(3)*, 275-292.

Duan, W., Yu, Y., Cao, Q.; Levy, S. (2016), Exploring the impact of social media on hotel Service performance: a sentimental analysis approach, *Cornell Hospitality Quarterly*, 57 *(3)*, 282-296.

Fávero, L., Belfiore, P.(2017), *Manual de Análise de Dados*. Rio de Janeiro: Elsevier.

Guo, Y., Barnes, S.J. and Jia, Q. (2016), Mining meaning from online ratings and reviews: tourist satisfaction analysis using Latent Dirichlet Allocation, *Tourism Management*, 59*(4)*, 467-483.

Hannigan, T.R., Haans, R.F.J., Vakili, K., Tchalian, H., Glaser, V.L., Wang, M.S., Kaplan, S.; Jennings, P.D. (2019), Topic modeling in management research: rendering new theory from textual data, *Academy of Management Annals*, 13 *(2)*, 586-632.

Hennig-Thurau, T., Wiertz, C.; Feldhaus, F. (2015), Does Twitter matter? The impact of

microblogging word of mouth on consumers' adoption of new movies, *Journal of the Academy of Marketing Science*, 43 (3), 375-394.

Hu, M. and Liu, B. (2004), Mining and summarizing customer reviews", in Kim, W. And Kohavi, R. (Eds), Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, *ACM*, New York, NY, 168-177.

Ippolito, P.P. (2022). Hyperparameter Tuning. In: Egger, R. (eds) *Applied Data Science in Tourism. Tourism on the Verge*. Springer, Cham. https://doi.org/10.1007/978-3-030-88389-8_12.

Irving, M., Coelho, A., Arruda, T. (2020) Turismos, sustentabilidades e pandemias: Incertezas e caminhos possíveis para planejamento turístico no horizonte da Agenda 2030, *Observatório de Inovação do Turismo*, 19 *(especial)*, 73-105.

Khodabandehlou, S.; Zivari Rahman, M. (2017), Comparison of supervised machine learning techniques for customer churn prediction based on analysis of customer behavior, *Journal of Systems and Information Technology*, 19 *(1/2)*, 65-93.

Korfiatis, N., Stamolampros, P., Kourouthanassis, P. and Sagiadinos, V. (2019), Measuring Service quality from unstructured data: a topic modeling application on airline passengers' online reviews, *Expert Systems with Applications*, 116 *(1)*, 472-486.

Kuhn, M. ; Johnson, K. (2013) Applied Predictive Modeling, *Springer New York*, NY.

Liu, B.; Zhang, L. (2012), A survey of opinion mining and sentiment analysis, in Aggarwal, C.C. and Zhai, C.X. (Eds), *Mining Text Data*, Springer, 415-463.

Lunardon, N., Menardi, G., and Torelli, N. (2014). ROSE: a Package for Binary Imbalanced Learning. *R Jorunal*, 6 *(1)* , 82–92.

Matlatipov, S., Rahimboeva, H., Rajabov, J., & Kuriyozov, E. (2022). Uzbek Sentiment Analysis based on local Restaurant Reviews. arXiv preprint arXiv:2205.15930.

Mariani, M., Baggio, R., Fuchs, M.; Höepken, W. (2018), Business intelligence and big data in hospitality and tourism: a systematic literature review, *International Journal of Contemporary Hospitality Management*, 30 *(12)*, 3514-3554.

Mtur (2018). Sistema Brasileiro de Classificação de Meios de Hospedagem. Disponível em: http://www.classificacao.turismo.gov.br/MTURclassificacao/mtur-site/index.jsp. Acesso em 18 Out. 21.

Pantano, E., Priporas, C.V. and Stylos, N. (2017), 'You will like it!' Using open data to predict tourists'response to a tourist attraction, *Tourism Management*, 60*(1)*, 430-438.

Peng, C.-Y.J., Lee, K.L., Ingersoll, G.M., (2002). An introduction to logistic regression analysis and reporting. *J. Educ. Res*. 96 *(1)*, 3–14.

Raguseo, E., Neirotti, P. and Paolucci, E. (2017), How small hotels can drive value their way in infomediation. The case of 'Italian hotels vs OTAs and TripAdvisor', *Information & Management*, 54*(6)*, 745-756.

Ravi, K.; Ravi, V. (2015), A survey on opinion mining and sentiment analysis: Tasks, Approaches and applications, *Knowledge-Based Systems*, 89 *(1)*, 14-46.

Reisenbichler, M. and Reutterer, T. (2019), Topic modeling in marketing: recent advances and research opportunities, *Journal of Business Economics*, 89 *(3)*, 327-356.

Rita, P., Rita, N. and Oliveira, C. (2018), Data science for hospitality and tourism, *Worldwide Hospitality and Tourism Themes*, 10 *(6)*, 717-725.

Sanchez-Franco, M.J., Cepeda-Carrion, G.; Roldán, J.L. (2019), Understanding relationship quality in hospitality services: A study based on text analytics and partial least squares, *Internet Research*, 29 *(3)*, 478-503.

Saralegi, X. and San Vincente, I. (2013), Elhuyar at TASS 2013", in Esteban, A.D., Loinaz, I.A. and Román, J.V. (Eds), Proceedings of "XXIX Congreso de La Sociedad Española de Procesamiento de Lenguaje Natural, presented at the XXIX Congreso de la Sociedad

Española de Procesamiento de Lenguaje Natural, Madrid, *El Congreso Español de Informática*, Spain, 143-150.

Serra-Cantallops, A.; Salvi, F. (2014), New consumer behavior: a review of research on eWOM and hotels, *International Journal of Hospitality Management*, 36 *(January)*, 41-51.

Shafqat, W.; Byun, Y.C. (2020), A recommendation mechanism for under-emphasized Tourist spots using topic modeling and sentiment analysis, *Sustainability*, 12 *(1)*, 320.

Shmueli, G., Bruce, P. C., Yahav, I., Patel, N. R., & Lichtendahl Jr, K. C. (2017). Data mining for business analytics: concepts, techniques, and applications in R. *John Wiley & Sons*.

Silva, M.J., Carvalho, P.; Sarmento, L. (2012), "Building a sentiment lexicon for social judgement mining, in Caseli, H., Villavicencio, A., Teixeira, A. and Perdigão, F. (Eds), *Computational Processing of the Portuguese Language*, Springer Berlin Heidelberg, New York, NY, 218-228.

Smola, A.; Vishwanathan, S.V.N. (2008), Introduction to Machine Learning, Cambridge University Press, Cambridge, UK.

Souza, M.; Vieira, R. (2012) Sentiment Analysis on Twitter Data for Portuguese Language. *10th International Conference Computational Processing of the Portuguese Language*.

Sparks, B.A., Kam Fung So, K. and Bradley, G.L. (2016), Responding to negative online reviews: the effects of hotel responses on customer inferences of trust and concern, *Tourism Management*, 53 *(4)*, 74-85.

Torres, E.N., Singh, D. and Robertson-Ring, A. (2015), Consumer reviews and the creation of bookingtransaction value: Lessons from the hotel industry, *International Journal of Hospitality Management*, 50 *(1)*, 77-83.

Tran, G.A., Strutton, D. and Taylor, D.G. (2012), Do microblog postings influence consumer perceptions of retailers'e-servicescapes?, *Management Research Review*, 35 (9), 818-836.

Vazquez, D., Dennis, C.; Zhang, Y. (2017), Understanding the effect of smart retail brand-consumer communications via mobile instant messaging (MIM) – an empirical study in the Chinese context, *Computers in Human Behavior*, 77 (1), 425-436.

Viglia, G., Minazzi, R. and Buhalis, D. (2016), The influence of e-word-of-mouth on hotel occupancy rate, *International Journal of Contemporary Hospitality Management*, 28 (9), 2035-2051.

Vu, H.Q., Li, G., Law, R. and Zhang, Y. (2019), Exploring tourist dining preferences based On restaurant reviews, *Journal of Travel Research*, 58 *(1)*, 149-167.

Xiang, Z., Du, Q., Ma, Y. and Fan, W. (2017), A comparative analysis of major online review platforms: implications for social media analytics in hospitality and tourism, *Tourism Management*, 58 *(1)*, 51-65.

Yu, C.E.; Sun, R. (2019), The role of Instagram in the UNESCO's creative city of gastronomy: a case study of Macau, *Tourism Management*, 75 *(1)*, 257-268.

Yu, C.-E.; 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.