**UNIVERSIDADE PRESBITERIANA MACKENZIE**

**Faculdade de Computação e Informática**

**Ciência De Dados**



**Projeto Aplicado II – IMBD**

**Gabriel Chaves Gonçalves**

**Italo Aparecido Lopes**

**São Paulo**

**2025**

**Sumário**

[**1.** **NOMENCLATURA** 6](#_Toc194009050)

[**2.** **INTRODUÇÃO** 6](#_Toc194009051)

[**3.** **TIPO DE ORGANIZAÇÃO** 7](#_Toc194009052)

[**4.** **OBJETIVO DO ESTUDO** 8](#_Toc194009053)

[**5.** **FUNDAMENTAÇÃO TEÓRICA** 9](#_Toc194009054)

[**5.1.** **ANÁLISE EXPLORATÓRIA DE DADOS** 9](#_Toc194009055)

[**5.1.1.** **METADADOS** 9](#_Toc194009056)

[**5.1.2.** **OUTLIERS** 10](#_Toc194009057)

[**5.1.3.** **ANÁLISE DE DISTRIBUIÇÃO DOS DADOS** 10](#_Toc194009058)

[**5.1.4.** **IQR** 10](#_Toc194009059)

[**5.2.** **PRÉ-PROCESSAMENTO DOS DADOS** 11](#_Toc194009060)

[**5.2.1.** **VETORIZAÇÃO DE TEXTO** 11](#_Toc194009061)

[**5.2.1.1.** **TF-IDF** 12](#_Toc194009062)

[**5.2.2.** **CODIFICAÇÃO DE VARIÁVEIS CATEGÓRICAS** 12](#_Toc194009063)

[**5.2.2.1.** **Label Encoding** 12](#_Toc194009064)

[**5.2.3.** **PARTICIONAMENTO DE CONJUNTOS DE DADOS** 13](#_Toc194009065)

[**5.2.3.1.** **Train\_test\_split** 13](#_Toc194009066)

[**5.3.** **PROCESSAMENTO DOS DADOS** 14](#_Toc194009067)

[**5.3.1.** **KNN** 14](#_Toc194009068)

[**5.3.2.** **REGRESSÃO LOGÍSTICA** 15](#_Toc194009069)

[**6.** **CRONOGRAMA** 17](#_Toc194009070)

[**7.** **ANÁLISE EXPLORATÓRIA DE DADOS** 17](#_Toc194009071)

[**7.1.** **RESULTADOS DA ANÁLISE E CONCLUSÃO DE CASO** 18](#_Toc194009072)

[**8.** **PRÉ-PROCESSAMENTO** 27](#_Toc194009073)

[**8.1.** **VETORIZAÇÃO** 27](#_Toc194009074)

[**8.2.** **CODIFICAÇÃO DOS RÓTULOS** 28](#_Toc194009075)

[**8.3.** **PARTICIONAMENTO DO *DATASET*** 29](#_Toc194009076)

[**8.4.** **MÉTRICAS** 29](#_Toc194009077)

[**9.** **LINK GITHUB** 30](#_Toc194009078)

[**10.** **REFERÊNCIAS BIBLIOGRÁFICAS** 30](#_Toc194009079)

**Lista de Figuras**

[Figura 1 - Logo da Organização 7](#_Toc194087815)

[Figura 2 – Informações do dataset via df.info() 18](#_Toc194087816)

[Figura 3 - Vetorização da coluna 'review' 28](#_Toc194087817)

[Figura 4 - Codificação da coluna "sentiment" 28](#_Toc194087818)

[Figura 5 - Train\_Test\_Split 29](#_Toc194087819)

**Lista de Equações**

[Equação 1 – Equação do método K-Nearest Neighbors (KNN) 15](#_Toc194008327)

[Equação 2 – Equação da regressão logística 16](#_Toc194008328)

**Lista de gráficos**

[Gráfico 1 – Distribuição de valores por sentimento 18](#_Toc194009082)

[Gráfico 2 – Histograma da distribuição do número de palavras 19](#_Toc194009083)

[Gráfico 3 – Comprimento do texto versus sentimento sem outliers 21](#_Toc194009084)

[Gráfico 4 – Quantidade total de palavras por sentimento com uso do pré-processamento textual NLTK 23](#_Toc194009085)

[Gráfico 5 – Top 20 palavras mais frequentes em avaliações positivas 24](#_Toc194009086)

[Gráfico 6 – Top 20 palavras mais frequentes em avaliações negativas 26](#_Toc194009087)

# **NOMENCLATURA**

Tokenização (Tokenization) – Técnica que divide um texto em unidades menores, como palavras ou frases.

Insights – Compreensão interna ou súbita de algo.

IMDb – Internet Movie Database

Remoção de Stopwords – Processo de eliminação de palavras irrelevantes para a análise, como preposições e artigos.

Dataset – banco de dados.

NLP – Processamento de Linguagem Natural

IQR – Intervalo Interquatílico

TF-IDF – Frequência do Termo - Frequência Inversa do Documento.

EAD – Análise Exploratória de Dados

Overfitting – Super ajuste do modelo ao dataset

# **INTRODUÇÃO**

A ciência de dados tem se tornado um dos campos mais relevantes da atualidade, impulsionando avanços em diversas áreas. O crescimento exponencial da informação gerada diariamente exige a extração de *insights* valiosos. Métodos estatísticos e aprendizado de máquina permitem compreender padrões, prever tendências e otimizar decisões. A análise de sentimentos destaca-se como ferramenta essencial para entender a opinião pública e embasar estratégias.

Os dados têm uma importância fundamental na ciência de dados e podem dividir-se em estruturados e não estruturados. Estruturados possuem formato organizado, armazenados em tabelas e bancos relacionais. Já os não estruturados incluem imagens, áudio, vídeos e textos, exigindo abordagens específicas para análise. A análise de sentimentos trabalha majoritariamente com textos, tornando-se um desafio para a classificação de opiniões.

O Processamento de Linguagem Natural (NLP) é fundamental para extrair significado de textos. Ele permite que máquinas compreendam e gerem linguagem humana. Técnicas como tokenização, remoção de *stopwords* e aprendizado profundo viabilizam a extração de sentimentos e emoções, impulsionando a automação da classificação textual.

A análise de sentimentos tem aplicações diversas, como avaliação de produtos, serviços e conteúdos digitais. Redes sociais e e-commerce utilizam essa tecnologia para melhorar experiências e adaptar estratégias. No conjunto de dados IMDB, que foi utilizado nesse trabalho, a classificação identifica automaticamente se uma avaliação é positiva ou negativa, otimizando a compreensão de opiniões.

A aplicação de aprendizado de máquina na classificação de sentimentos envolve pré-processamento, extração de características e treinamento. Modelos como *Naive Bayes*, Regressão Logística e redes neurais são utilizados.

Este trabalho explora a classificação de sentimentos na base IMDB, abordando do pré-processamento à avaliação de resultados. A investigação busca comparar a aplicação de algoritmos para análise de sentimentos. Espera-se demonstrar a relevância da ciência de dados na extração de *insights* a partir de textos.

# **TIPO DE ORGANIZAÇÃO**

O *Internet Movie Database* (IMDb) é uma das maiores e mais populares bases de dados sobre filmes, séries e entretenimento. Criado em 1990 por Col Needham como um hobby, o site começou como uma lista compartilhada de filmes e rapidamente cresceu em popularidade. Em 1998, a plataforma foi adquirida pela *Amazon*, tornando-se uma ferramenta essencial para cinéfilos e profissionais da indústria.



Figura 1 - Logo da Organização

O IMDb contém informações detalhadas sobre produções audiovisuais, incluindo elencos, diretores, sinopses e avaliações do público. Um dos seus diferenciais é o sistema de notas, onde usuários podem classificar filmes e séries, gerando rankings influentes na indústria. Além disso, críticos e especialistas também contribuem com análises e reviews aprofundados.

Ao longo dos anos, o IMDb expandiu seus serviços, incluindo o IMDbPro, voltado para profissionais do entretenimento, oferecendo dados sobre contratações e estatísticas do mercado. Outro destaque é o IMDb TV, um serviço de streaming gratuito com filmes e séries licenciados, disponível em alguns países.

Atualmente, o IMDb é referência global, sendo acessado por milhões de usuários diariamente. A plataforma mantém uma presença ativa nas redes sociais e realiza premiações como o IMDb Top 10, que lista os filmes e séries mais populares do ano. Seu banco de dados continua crescendo, acompanhando as tendências do cinema e da TV.

Com mais de 30 anos de história, o IMDb segue como uma ferramenta indispensável para fãs e profissionais do entretenimento. Sua influência no setor audiovisual se mantém forte, impactando a visibilidade e a recepção de diversas produções ao redor do mundo.

# **OBJETIVO DO ESTUDO**

O objetivo principal deste estudo é realizar a análise de sentimentos em um conjunto de dados de avaliações de filmes, por meio de uma base de dados IMDB disponível no Kaggle.

O estudo visa investigar como diferentes técnicas de aprendizado supervisionado podem ser aplicadas para classificar os sentimentos expressos nas avaliações deixadas por diversos usuários sobre filmes, identificando-as como positivas ou negativas. A partir disso, espera-se demonstrar como a ciência de dados pode ser útil na extração de *insights* a partir de textos não estruturado.

Por fim, este estudo busca demonstrar como algoritmos de aprendizado supervisionado podem ser aplicados em diferentes domínios, indo além do meio acadêmico. O entendimento destas técnicas visa impactar áreas como sistemas de recomendação, saúde, finanças, política e educação, entre outras. Assim, ao explorar sua viabilidade e desafios, esta pesquisa pode servir de base para estudos futuros e avanços na automação da análise textual.

# **FUNDAMENTAÇÃO TEÓRICA**

A fundamentação teórica é uma etapa essencial para a construção de um trabalho acadêmico, pois fornece embasamento conceitual e metodológico ao estudo. Nessa seção, são apresentados os principais conceitos, modelos e teorias que sustentam a pesquisa, permitindo uma compreensão mais ampla do tema investigado. Portanto, neste capítulo, abordaremos os conceitos utilizados no desenvolvimento deste trabalho.

# **ANÁLISE EXPLORATÓRIA DE DADOS**

A Análise Exploratória de Dados (AED) consiste em técnicas gráficas e estatísticas para entender dados, identificar padrões, tendências e anomalias. Seu objetivo é gerar insights para facilitar decisões estratégicas, melhorando etapas posteriores de modelagem e análise.

Segundo Bussab e Morettin (2017), essa etapa inclui técnicas gráficas e estatísticas para identificar padrões, tendências, relações e possíveis anomalias nos dados. O objetivo principal é gerar insights relevantes que orientem adequadamente as próximas etapas da análise e a escolha de métodos adequados.

# **METADADOS**

Os metadados são informações estruturadas que descrevem características importantes de outros dados, ajudando na sua organização, recuperação, interpretação e utilização adequada. Seu uso garante maior eficiência em sistemas de informação, facilitando buscas e garantindo a qualidade das informações armazenadas.

Segundo Sayão e Marcondes (2005), eles facilitam o entendimento do contexto, origem, estrutura e características essenciais das informações armazenadas. Seu principal objetivo é auxiliar na recuperação eficiente dos dados, aumentando a precisão, interoperabilidade e qualidade da gestão informacional.

# **OUTLIERS**

Os outliers são valores que se destacam muito dos demais em um conjunto de dados, sugerindo possíveis erros ou comportamentos atípicos. Identificá-los é essencial para garantir análises confiáveis e resultados precisos.

Segundo Bussab e Morettin (2017), esses valores discrepantes exigem atenção especial, pois influenciam diretamente na confiabilidade e precisão das análises estatísticas. O uso destas técnicas estatísticas e gráficas são importantes para identificá-los corretamente. Dessa forma, ressaltam que seu tratamento adequado garante resultados mais robustos e seguros.

# **ANÁLISE DE DISTRIBUIÇÃO DOS DADOS**

A análise de distribuição dos dados busca compreender como os valores de uma variável estão distribuídos dentro de um conjunto de informações, identificando padrões como simetria, assimetria e dispersão. A partir disso, é possível aplicar testes estatísticos adequados e tomar decisões assertivas sobre o uso dessas informações.

De acordo com Levine, Stephan e Szabat (2016), a análise da distribuição dos dados permite compreender aspectos essenciais como a concentração dos valores em torno de medidas centrais, dispersão e simetria. Os autores destacam que essa análise é indispensável para validar suposições estatísticas e assegurar a aplicação correta de métodos quantitativos, garantindo resultados confiáveis nas análises realizadas.

# **IQR**

O intervalo interquartílico (IQR - Interquartile Range) é uma medida estatística que representa a dispersão dos dados em torno da mediana, sendo calculado pela diferença entre o terceiro e o primeiro quartil (Q3-Q1). Esse método é amplamente utilizado por ser resistente à presença de outliers.

Segundo Triola (2017), o intervalo interquartílico (IQR) é uma medida eficaz para avaliar a dispersão dos dados sem sofrer influência significativa dos valores extremos ou outliers. Ao utilizar o IQR, obtém-se uma visão robusta da variação das observações, favorecendo análises mais consistentes, especialmente em situações que envolvem dados assimétricos ou com anomalias.

# **PRÉ-PROCESSAMENTO DOS DADOS**

O pré-processamento de dados é uma etapa fundamental na análise de dados, sendo responsável pela limpeza, transformação e normalização das informações antes da aplicação de modelos de aprendizado de máquina.

Segundo Han, Kamber e Pei (2011), essa fase envolve alguns tratamentos nos dados, como a conversão de variáveis categóricas e padronização dos dados para garantir maior qualidade e consistência. O objetivo principal é minimizar problemas que possam comprometer a performance dos algoritmos e melhorar a capacidade preditiva dos modelos utilizados.

# **VETORIZAÇÃO DE TEXTO**

A vetorização de texto é um dos principais processos do pré-processamento de dados textuais, permitindo que textos sejam convertidos em representações numéricas para utilização em modelos de aprendizado de máquina.

Segundo Jurafsky e Martin (2021), essa transformação é essencial, pois os algoritmos computacionais operam sobre números, tornando necessário mapear palavras ou documentos para um espaço vetorial. Métodos tradicionais incluem o *Bag of Words* (BoW) e o *Term Frequency-Inverse Document Frequency* (TF-IDF), que representam os textos com base na frequência das palavras, capturando padrões estatísticos do vocabulário.

# **TF-IDF**

O método *Term Frequency-Inverse Document Frequency* (TF-IDF) é uma técnica amplamente utilizada na vetorização de texto, permitindo a transformação de documentos em representações numéricas para análise e modelagem computacional.

Segundo Manning, Raghavan e Schütze (2008), o TF-IDF combina duas métricas principais: a Frequência do Termo (TF), que mede a recorrência de uma palavra em um documento, e a Frequência Inversa do Documento (IDF), que avalia a importância do termo considerando sua presença em vários documentos. Esse equilíbrio garante que palavras comuns, como artigos e preposições, tenham menor peso, enquanto termos mais relevantes recebam maior ênfase na representação vetorial.

# **CODIFICAÇÃO DE VARIÁVEIS CATEGÓRICAS**

A codificação de variáveis categóricas é uma etapa essencial do pré-processamento de dados, especialmente em modelos de aprendizado de máquina, que operam com valores numéricos.

Segundo Kuhn e Johnson (2013), variáveis categóricas representam atributos qualitativos que precisam ser convertidos em uma forma numérica para serem interpretados por algoritmos estatísticos e de machine learning. Os principais métodos de codificação incluem One-Hot Encoding e Label Encoding. A escolha do método de codificação pode impactar diretamente o desempenho do modelo.

# **Label Encoding**

O Label Encoding é um método de codificação de variáveis categóricas que converte valores não numéricos em representações inteiras, sendo uma abordagem fundamental no pré-processamento de dados para aprendizado de máquina.

De acordo com Kuhn e Johnson (2013), esse método mapeia cada categoria única de uma variável para um número inteiro específico, mantendo a quantidade original de informações e evitando o aumento da dimensionalidade dos dados.

Segundo James et al. (2021), enquanto Label Encoding pode ser eficiente para modelos baseados em árvores, como Random Forest e XGBoost, o One-Hot Encoding é mais adequado para algoritmos lineares

# **PARTICIONAMENTO DE CONJUNTOS DE DADOS**

O particionamento de conjuntos de dados é um processo fundamental no aprendizado de máquina, permitindo a avaliação e generalização de modelos preditivos.

De acordo com Hastie, Tibshirani e Friedman (2009), a divisão dos dados em conjunto de treinamento (training set), validação (*validation set*) e teste (*test set*) é essencial para evitar *overfitting* e garantir que o modelo aprenda padrões generalizáveis.

# **Train\_test\_split**

O método train\_test\_split da biblioteca Scikit-Learn é amplamente utilizado para particionar conjuntos de dados em subconjuntos de treino e teste de maneira eficiente e aleatória.

Segundo Géron (2019), essa técnica é essencial para avaliar o desempenho de modelos de aprendizado de máquina, pois permite separar uma fração dos dados para teste, garantindo que a avaliação do modelo seja feita em exemplos que ele não viu durante o treinamento.

Além disso, conforme Pedregosa et al. (2011), o train\_test\_split suporta a opção stratify, garantindo que a distribuição das classes na variável-alvo seja preservada entre os conjuntos de treino e teste. Isso é especialmente importante em problemas de classificação desbalanceados, onde a proporção de classes minoritárias deve ser mantida para evitar viés no modelo.

# **PROCESSAMENTO DOS DADOS**

O processamento de dados é uma etapa essencial na ciência de dados, envolvendo a transformação de informações brutas em um formato adequado para análise. Segundo Provost e Fawcett (2013). Essa fase pode incluir a limpeza, integração, transformação e redução de dados, garantindo qualidade e consistência antes da modelagem. Métodos como normalização, padronização e tratamento de valores ausentes são frequentemente empregados para melhorar a performance de algoritmos de aprendizado de máquina.

Além disso, dependendo da natureza dos dados, diferentes abordagens são aplicadas. Para séries temporais, técnicas como suavização e diferenciação são utilizadas para remover tendências e sazonalidade (HYNDMAN; ATHANASOPOULOS, 2018).

# **KNN**

O método K-Nearest Neighbors (KNN) é uma técnica de aprendizado de máquina não paramétrica amplamente utilizada para problemas de classificação, onde a variável resposta é categórica, podendo assumir dois ou mais valores. Segundo James et al. (2021), o método KNN classifica uma nova observação com base nas classes das k observações mais próximas presentes no conjunto de treinamento, utilizando uma métrica de distância como a Euclidiana.

Matematicamente, a distância Euclidiana entre duas observações pode ser representada pela equação 1:

Equação 1 – Equação do método K-Nearest Neighbors (KNN)

representa a distância entre a observação nova x e uma observação existente

representa o valor do atributo j na nova observação;

representa o valor do atributo j da observação existente

é o número total de atributos ou variáveis independentes.

De acordo com Bishop (2006), a escolha do valor k influencia diretamente no desempenho do modelo, e geralmente é determinada por meio de técnicas como validação cruzada. Além disso, recomenda-se a normalização prévia dos dados para evitar que variáveis com diferentes escalas tenham influência desproporcional no resultado.

O método KNN é amplamente utilizado em áreas como diagnóstico médico, sistemas de recomendação e classificação de imagens devido à sua simplicidade conceitual e fácil implementação. Entretanto, pode apresentar desafios em conjuntos grandes ou de alta dimensionalidade, exigindo métodos auxiliares para melhorar sua eficiência.

# **REGRESSÃO LOGÍSTICA**

A Regressão Logística é um aprendizado de máquina baseado em modelo estatístico e é amplamente utilizado para problemas de classificação binária, onde a variável resposta assume apenas dois valores, como 0 e 1, sim e não, ou positivo e negativo. Segundo Hosmer, Lemeshow e Sturdivant (2013), a regressão logística modela a relação entre as variáveis preditoras X e a probabilidade de um determinado evento ocorrer, utilizando a função sigmoide para restringir os valores previstos ao intervalo [0,1].

Matematicamente, a função da regressão logística, posso ser observada através de equação 1 e é expressa como:

Equação 2 – Equação da regressão logística

onde:

P(Y=1∣X) representa a probabilidade condicional do evento de interesse ocorrer;

𝛽 é o intercepto do modelo;

β1, β2, …, βn: os coeficientes para cada variável independente;

𝑋1, 𝑋2, …, 𝑋𝑛: as variáveis independentes ou preditoras;

𝑒 é a base do logaritmo natural.

De acordo com James et al. (2021), a interpretação dos coeficientes 𝛽𝑖 na regressão logística é feita por meio do odds ratio, que indica o impacto de cada variável explicativa na probabilidade de o evento ocorrer. Além disso, técnicas como Regularização L1 (Lasso) e L2 (Ridge) podem ser aplicadas para evitar *overfitting*, especialmente quando há alta colinearidade entre as variáveis.

A Regressão Logística é amplamente empregada em diversas áreas, como medicina (para prever doenças), finanças (para análise de risco de crédito) e aprendizado de máquina (para classificação de textos e imagens). Sua simplicidade e interpretabilidade fazem dela uma escolha popular em tarefas de modelagem preditiva.

# **CRONOGRAMA**

* **Etapa 1**: metas e milestones – Data 03/03/2025
* **Etapa 2**: analytics – Data 31/03/2025
* **Etapa 3**: storytelling – Data 28/04/2025
* **Etapa 4**: apresentação e encerramento – Data 26/05/2025

# **ANÁLISE EXPLORATÓRIA DE DADOS**

A análise exploratória de dados constitui uma etapa crucial no processo de compreensão profunda dos padrões e características que permeiam um conjunto de dados, especialmente quando o foco é a classificação textual de avaliações relacionadas ao cinema. Neste projeto, foi realizada de forma minuciosa a análise do dataset do IMDB, constituído por 50.000 avaliações textuais classificadas em positivas e negativas.

O objetivo principal desta análise exploratória foi investigar detalhadamente a relação entre o comprimento dos textos, as palavras utilizadas nas avaliações e o sentimento expresso pelos usuários, visando identificar padrões que possam auxiliar na construção de modelos eficazes de classificação de sentimento.

Antes de ser iniciada a etapa de tratamento de outliers, foi necessário realizar a validação do balanceamento dos dados afim de eliminar qualquer novo pré-tratamento na base de dados.

A criação do gráfico de barras permitiu revelar um equilíbrio perfeito entre as avaliações, com 25.000 registros para cada classe (positiva e negativa), permitindo uma análise comparativa justa e imparcial entre as categorias, eliminando possíveis vieses decorrentes de desbalanceamentos.

Gráfico, Gráfico de barras

O conteúdo gerado por IA pode estar incorreto.

Gráfico 1 – Distribuição de valores por sentimento

Além disso, observou-se que o arquivo possui um total de 50.000 linhas e 2 colunas originais: “review” (comentários textuais sobre filmes) e “sentiment” (classificação do sentimento como positivo ou negativo).

Figura 2 – Informações do dataset via df.info()

Texto

O conteúdo gerado por IA pode estar incorreto.

# **RESULTADOS DA ANÁLISE E CONCLUSÃO DE CASO**

Para possibilitar a análise exploratória, inicialmente foram necessárias algumas etapas de pré-processamento dos dados. Primeiramente, criou-se uma coluna chamada clean\_review, que converte cada avaliação para o formato de texto puro. Em seguida, a coluna text\_length foi criada para calcular o comprimento das avaliações, considerando o número de palavras presentes em cada texto.

Adicionalmente, para facilitar análises quantitativas, a coluna sentiment, que descrevia avaliações como ‘positive’ ou ‘negative’, foi transformada em valores numéricos, gerando a coluna sentiment\_numeric (com valores 1 para positivo e 0 para negativo).

Por fim, os dados foram separados em dois grupos, avaliações positivas e negativas, viabilizando análises comparativas entre os sentimentos presentes nas avaliações.

Essas etapas permitiram uma avaliação inicial clara da distribuição e comprimento dos textos e forneceram bases consistentes para análises mais aprofundadas.

Em seguida, foi conduzida uma investigação sobre a distribuição do comprimento dos textos das avaliações, utilizando gráficos como a curva de densidade Kernel (KDE) e histogramas. O gráfico KDE apresentou uma distribuição notavelmente assimétrica positiva, com textos variando entre 0 e 1000 palavras, indicando uma concentração significativa das avaliações nessa faixa e uma cauda longa representando avaliações excepcionalmente extensas.  Gráfico, Histograma

O conteúdo gerado por IA pode estar incorreto.

Gráfico 2 – Histograma da distribuição do número de palavras

Essa observação fundamentou a escolha do método Interquartile Range (IQR) para o tratamento de outliers, por ser um método robusto para dados com distribuições assimétricas como essa, observadas no gráfico (X).

Através da visualização em gráficos de dispersão, foi possível verificar que textos classificados como positivos e negativos apresentavam valores extremos próximos a 2500 palavras, representando potenciais outliers segundo a técnica adotada.

Uma imagem contendo Interface gráfica do usuário

O conteúdo gerado por IA pode estar incorreto.

Contudo, após a aplicação do método IQR padrão (com fator 1,5), observou-se uma redução significativa dos limites superiores, passando de valores próximos a 2500 palavras para valores máximos em torno de 500 palavras, tanto para reviews positivos quanto negativos.

Uma imagem contendo Gráfico

O conteúdo gerado por IA pode estar incorreto.

Gráfico 3 – Comprimento do texto versus sentimento sem outliers

Inicialmente, foram encontrados:

• 1899 outliers positivos

• 1776 outliers negativos.

Entretanto, ao analisar o gráfico posterior à remoção desses outliers, ficou evidente um enviesamento dos dados. Antes da remoção, era possível observar diversos registros legítimos com comprimento superior a 500 palavras, chegando a textos próximos de 1000 palavras, considerados importantes do ponto de vista representativo e qualitativo das avaliações feitas pelos usuários.

Esses dados, apesar de classificados como outliers pela técnica padrão do IQR, ainda assim representam características reais dos usuários e suas percepções sobre os filmes analisados, não sendo, portanto, registros inválidos ou dados incorretos.

Dessa forma, torna-se claro que o uso do IQR com o fator padrão (1,5) para remoção dos dados resultou em uma exclusão indevida de valores representativos, gerando perdas importantes para a riqueza informacional e diversidade do *dataset* original.

Após a identificação dos problemas decorrentes da remoção de dados representativos usando o fator padrão (1.5) do método IQR, foi considerada a possibilidade de elevar esse fator para 3.0, objetivando uma redução na quantidade de registros indevidamente classificados como outliers.

Ao realizar testes com o fator alterado (3.0), o número de outliers diminuiu significativamente, passando para:

• 584 outliers positivos (originalmente 1899).

• 602 outliers negativos (originalmente 1776).

Entretanto, apesar de inicialmente parecer benéfica a redução na quantidade de dados descartados, essa mudança provocou impactos importantes nos valores estatísticos do dataset, especialmente nos quartis, na mediana e no intervalo interquartil. Ao elevar demasiadamente os limites para identificação de outliers, diversos valores extremos—embora em menor número—permaneceram nos dados, contribuindo para o aumento da assimetria e da variabilidade interna dos registros restantes.

Dessa forma, a complexidade estatística resultante dessa decisão poderia gerar um enviesamento ainda maior, comprometendo a distribuição dos dados e dificultando a aplicação adequada de técnicas posteriores, como modelos preditivos de aprendizado supervisionado.

Em outras palavras, a elevação do fator para 3.0 não resolveu o problema central da representatividade dos dados, mas sim, deslocou o problema, permitindo que valores excessivamente extremos permanecessem no dataset, o que poderia gerar distorções ainda mais críticas além de impactar no aprendizado de máquina supervisionado utilizado no projeto tais como K-Nearest Neighbors (KNN) e Regressão Logística.

No caso do algoritmo KNN, a preservação desses dados permite que o modelo mantenha uma maior fidelidade à distribuição real dos textos, já que este método depende diretamente da proximidade entre os dados. Para a regressão logística, por sua vez, a exclusão dos outliers poderia acarretar a simplificação excessiva dos padrões, levando a um modelo menos capaz de capturar nuances importantes no comportamento dos dados.

Após a aplicação do pré-processamento textual utilizando o NLTK, no qual foram removidas apenas pontuações mantendo-se as stopwords, obteve-se uma visualização quantitativa detalhada da quantidade total de palavras utilizadas pelos usuários nas avaliações positivas e negativas no dataset IMDB.

Gráfico, Gráfico de barras

O conteúdo gerado por IA pode estar incorreto.

Gráfico 4 – Quantidade total de palavras por sentimento com uso do pré-processamento textual NLTK

O gráfico de barras gerado nesta etapa revela uma diferença expressiva entre os sentimentos manifestados pelos usuários em termos de volume textual. Avaliações classificadas como positivas alcançaram um total de aproximadamente 5.728.874 palavras, enquanto as avaliações negativas somaram aproximadamente 5.643.208 palavras.

Mesmo considerando que ambas as categorias possuem exatamente 25.000 avaliações, há uma clara tendência dos usuários em utilizar maior quantidade de palavras ao expressar opiniões positivas, com uma diferença numérica significativa de aproximadamente 85.666 palavras a mais para avaliações positivas.

Esse padrão pode indicar que usuários com experiências positivas tendem a descrever suas avaliações de maneira mais detalhada e abrangente, possivelmente motivados por um desejo intrínseco de compartilhar suas percepções favoráveis ou recomendar conteúdos que julgam merecedores. Em contraste, avaliações negativas, embora próximas quantitativamente, mostram uma ligeira preferência por comentários mais objetivos e sucintos, possivelmente devido à natureza crítica e pontual dessas avaliações.

Esses resultados têm importantes implicações estratégicas para plataformas digitais como o IMDB. Compreender esse comportamento permite um entendimento mais profundo do engajamento dos usuários, possibilitando estratégias direcionadas que reforcem aspectos valorizados pelo público em filmes positivamente avaliados. Ao mesmo tempo, a objetividade identificada nas avaliações negativas fornece indícios claros sobre pontos específicos que merecem maior atenção ou aprimoramento por parte de produtores e gestores de conteúdo.

Dessa forma, o uso do pré-processamento textual com o NLTK não apenas elevou a qualidade da análise quantitativa, como também proporcionou insights essenciais sobre as nuances do comportamento dos usuários em plataformas digitais especializadas em avaliações de filmes, como foi o caso do citado.

Gráfico

O conteúdo gerado por IA pode estar incorreto.

Gráfico 5 – Top 20 palavras mais frequentes em avaliações positivas

Já para a análise da frequência das palavras utilizadas em cada uma das categorias o gráfico evidencia que as palavras mais recorrentes especificamente nas avaliações positivas são termos diretamente relacionados à experiência cinematográfica, tais como “film” (40.765 ocorrências), “movie” (36.706), “one” (26.537) e “like” (17.231). Outros termos frequentemente mencionados foram “good” (14.492), “great” (12.811), e “story” (12.531), indicando claramente uma tendência dos usuários em destacar aspectos qualitativos específicos dos filmes ao emitir uma opinião favorável.

Outro ponto importante é a presença de termos que reforçam impressões pessoais positivas, como “well” (11.230), “really” (10.713), “love” (8.455) e “best” (8.350). A alta frequência desses termos sugere que usuários satisfeitos tendem a expressar não apenas uma aprovação técnica dos filmes, mas também sentimentos pessoais mais intensos e emocionalmente engajados, revelando um comportamento marcadamente subjetivo nas avaliações positivas.

Essa característica se conecta diretamente à observação feita anteriormente, onde avaliações positivas totalizaram um número significativamente maior de palavras em relação às avaliações negativas. Isso reforça a hipótese de que usuários satisfeitos são mais propensos a elaborar avaliações detalhadas, não apenas descrevendo o conteúdo assistido, mas também expressando emoções e enfatizando aspectos positivos específicos, seja na narrativa (“story”), no elenco (“people”), ou nas próprias percepções pessoais (“love”, “best”).

Gráfico, Gráfico de barras

O conteúdo gerado por IA pode estar incorreto.

Gráfico 6 – Top 20 palavras mais frequentes em avaliações negativas

Por outro lado, a análise das avaliações negativas, após o pré-processamento com o NLTK, revela um comportamento linguístico mais objetivo e crítico dos usuários.

Nas 25.000 avaliações negativas analisadas, destacam-se termos como “movie” (48.364 ocorrências) e “film” (36.154 ocorrências), frequentemente empregados em contextos depreciativos ou críticos, apontando diretamente ao objeto da insatisfação. Ao contrário das avaliações positivas, onde há uma riqueza em descrições emocionais e detalhadas, as avaliações negativas são marcadas pelo uso frequente de palavras que sugerem expectativas frustradas ou desapontamento explícito. Isso fica evidente nas palavras “bad” (14.223), “good” (14.275) e “really” (12.311), que demonstram uma forte comparação entre o que foi esperado e o que foi entregue pelo conteúdo assistido.

Além disso, termos condicionais como “would” (15.224) e “could” (11.342) revelam uma tendência de usuários insatisfeitos em apontar diretamente o potencial não realizado das produções, sugerindo melhorias específicas. Também chama atenção a frequência da palavra “time” (11.614), indicando que um fator importante de insatisfação é a percepção de tempo perdido ou experiência frustrante. A presença recorrente das palavras “see” (10.660), “get” (10.058) e “story” (9.810) reforça a ideia de problemas relacionados à narrativa e à capacidade de engajamento com o público.

Esses resultados indicam que usuários insatisfeitos expressam sua opinião de maneira objetiva e direta, contrastando claramente com o comportamento expansivo e emocional das avaliações positivas. Identificar essas palavras-chave e seus contextos de uso pode fornecer importantes direcionamentos estratégicos para plataformas digitais e produtores de conteúdo, permitindo identificar rapidamente pontos de atenção para reduzir insatisfações e aprimorar a qualidade das produções oferecidas.

Em resumo, as diversas técnicas exploratórias aplicadas ao dataset desde análises quantitativas detalhadas até visualizações gráficas permitiram obter insights profundos sobre o comportamento linguístico e emocional dos usuários.

Esses insights não apenas auxiliam na compreensão da dinâmica das avaliações, como também contribuem para estratégias eficazes de engajamento e melhoria contínua na qualidade dos conteúdos oferecidos ao público em como o aprimoramento das ferramentas de algoritmo de aprendizado de máquina supervisionado.

# **PRÉ-PROCESSAMENTO**

Na literatura, ainda há ampla discussão sobre as etapas que compõem o pré-processamento de dados. Neste trabalho, focaremos especificamente na vetorização de textos, na codificação de rótulos em variáveis numéricas e no particionamento do conjunto de dados para teste e a escolha das métricas, uma vez que outros tratamentos, como a remoção de outliers e stopwords, já foram abordados anteriormente na análise exploratória de dados.

# **VETORIZAÇÃO**

Para realizar a vetorização foi utilizado o método TF-IDF, então o importamos através da biblioteca do sklearn e vetorizamos a nossa coluna chamada de “review”, conforme mostra a figura 3.

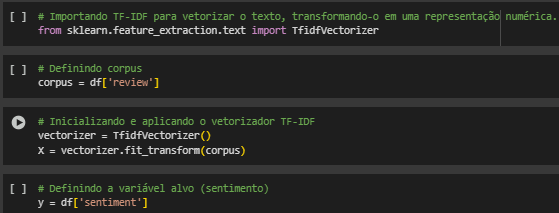


Figura 3 - Vetorização da coluna 'review'

Nota-se, também, que além da vetorização também começamos o particionamento do *dataset,* pois em uma variável “x” que serão nossos dados de input, guardamos o nosso texto vetorizado e na nossa variável “y” que será nosso output foi guardado a coluna de sentimento.

# **CODIFICAÇÃO DOS RÓTULOS**

Após realizar a vetorização, iniciou-se a conversão do rótulo do tipo categórico para uma variável do tipo numérica. Portanto, foi utilizado o método Label Encoding que foi importado da biblioteca do sklearn, conforme mostra a figura 4.

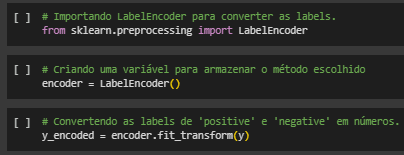


Figura 4 - Codificação da coluna "sentiment"

Note que rótulos codificados foram armazenados em uma variável chamada “y\_encoded”, a fim de se facilitar a manipulação dos dados posteriormente.

# **PARTICIONAMENTO DO *DATASET***

Para a divisão dos dados foi utilizado o método train\_test\_split importado através da biblioteca sklearn, conforme elucida a figura 5.

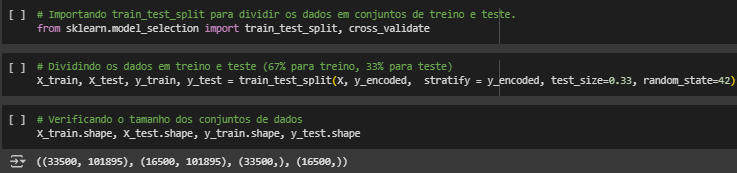


Figura 5 - Train\_Test\_Split

Pode-se observar que os dados foram divididos em treino e teste, com a maior parte destinada ao conjunto de treino, visando reduzir o risco de *overfitting* do modelo. Além disso, é importante destacar que o método de validação cruzada, também proveniente da biblioteca sklearn, já foi importado, pois será utilizado posteriormente no processamento dos dados.

# **MÉTRICAS**

A escolha de métricas adequadas é essencial para avaliar o desempenho do modelo de análise de sentimentos. Como as classes do conjunto de dados estão perfeitamente balanceadas, a acurácia se torna uma métrica confiável, pois reflete corretamente a taxa de acertos do modelo sem ser influenciada por uma classe dominante. No entanto, apenas a acurácia não é suficiente para garantir uma avaliação completa.

A precisão foi escolhida para medir a taxa de acertos ao classificar um sentimento como positivo ou negativo, reduzindo falsos positivos que poderiam comprometer a análise. Já o recall é fundamental para garantir que o modelo não deixe de identificar textos pertencentes a uma classe específica, evitando a perda de informações relevantes.

O F1-score, embora útil em cenários de classes desbalanceadas, não foi priorizado, pois não traria ganhos significativos em um conjunto de dados equilibrado.

# **LINK GITHUB**

<https://github.com/lopesita/projetoII_IMDB>

# **REFERÊNCIAS BIBLIOGRÁFICAS**

BUSSAB, Wilton de O.; MORETTIN, Pedro A. Estatística Básica. 9. ed. São Paulo: Saraiva Educação, 2017.

GÉRON, A. Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow. 2nd ed. O’Reilly Media, 2019.

HAN, J.; KAMBER, M.; PEI, J. Data Mining: Concepts and Techniques. 3rd ed. Elsevier, 2011.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd ed. Springer, 2009.

HYNDMAN, R. J.; ATHANASOPOULOS, G. Forecasting: Principles and Practice. 2nd ed. OTexts, 2018.

JAMES, G. et al. An Introduction to Statistical Learning: with Applications in R. 2nd ed. Springer, 2021

LEVINE, David M.; STEPHAN, David F.; SZABAT, Kathryn A. Estatística: Teoria e Aplicações. 7. ed. Rio de Janeiro: LTC, 2016.

JONES, K. S. A statistical interpretation of term specificity and its application in retrieval. Journal of Documentation, v. 28, n. 1, p. 11-21, 1972.

JURAFSKY, D.; MARTIN, J. H. Speech and Language Processing. 3rd ed. Pearson, 2021.

KUHN, M.; JOHNSON, K. Applied Predictive Modeling. Springer, 2013.

PEDREGOSA, F. et al. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, v. 12, p. 2825-2830, 2011.

PROVOST, F.; FAWCETT, T. Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking. O'Reilly Media, 2013.

SAYÃO, Luís Fernando; MARCONDES, Carlos Henrique. Metadados: fundamentos e aplicações. Brasília: Briquet de Lemos, 2005.

TRIOLA, Mario F. Estatística Aplicada às Ciências Sociais. 12. ed. São Paulo: Pearson Education do Brasil, 2017.

Bishop, C. M. (2006). Pattern Recognition and Machine Learning. Springer.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). An Introduction to Statistical Learning: With Applications in R (2ª ed.). Springer.