Tecnólogo em Ciência de Dados

Projeto Aplicado II
ANÁLISE DE SENTIMENTOS EM REVIEWS DE FILMES DO IMDB

DIANNA MAYUMI SANTOS KATAYAMA RODRIGUES – 22501762

LUCAS SILVA RIBEIRO – 22520473

PEDRO AUGUSTO ALVES MACENA LIMA – 22509747

RICARDO MANOEL FERREIRA DE OLIVEIRA – 22504141

RICARDO PARDONO – 22517189

São Paulo

2023

Universidade Presbiteriana Mackenzie

Tecnólogo em Ciência de Dados

Projeto Aplicado II

ANÁLISE DE SENTIMENTOS EM REVIEWS DE FILMES DO IMDB

Pré – Projeto apresentado para a disciplina de Projeto Aplicado II - Etapa 3

Orientador: Prof. Dr. Anderson Adaime de Borba

Tecnólogo em Ciência de Dados

DIANNA MAYUMI SANTOS KATAYAMA RODRIGUES – 22501762

LUCAS SILVA RIBEIRO – 22520473

PEDRO AUGUSTO ALVES MACENA LIMA – 22509747

RICARDO MANOEL FERREIRA DE OLIVEIRA – 22504141

RICARDO PARDONO – 22517189

São Paulo

2023

Resumo

A análise de sentimentos tem como objetivo definir técnicas automáticas capazes de extrair informações subjetivas de textos em linguagem natural, como opiniões e sentimentos, a fim de criar conhecimento estruturado que possa ser utilizado por um sistema de apoio ou tomador de decisão. A identificação de sentimentos em textos é uma das áreas de pesquisa mais destacadas em Processamento de Linguagem Natural e mineração de texto.

Atualmente, diversos espectadores, quando desejam assistir a algum filme em sua casa ou no cinema, recorrem a sites especializados, em busca de críticas sobre uma determinada obra, onde estas opiniões possuem forte influência sobre a decisão do usuário de assistir, ou não, a um determinado filme. Tais opiniões vêm sendo cada vez mais utilizadas na obtenção de feedback diretamente do público-alvo durante os últimos anos, com a aplicação da análise de sentimentos.

Neste sentido, insere-se este projeto, cujo objetivo principal é desenvolver um modelo preditivo baseado em técnicas de aprendizado de máquina supervisionado pela equipe de Ciência de Dados da empresa de streaming "Mackflix", que em parceria com o site especializado em cinema Internet Movie Database (IMDB), usará esse modelo como apoio na tomada de decisão e análise de tendências de gênero e temas de filmes preferidos pelo público, visando a aquisição de conteúdo cinematográfico de qualidade para seu catálogo, alinhado com as preferências do público.

Assim sendo, esse modelo será capaz de classificar a polaridade dos comentários nos reviews dos usuários do IMDB, utilizando como base de treino e teste para seus algoritmos o Dataset "imdbmovies.csv", com cerca de 3883 comentários sobre diversos filmes extraídos do site imdb.com.

Palavras-chaves: IMDB. Processamento de Linguagem Natural. Análise de sentimentos.

Lista de ilustrações

Lista de gráficos

Sumário

1 INTRODUÇÃO

Quando um consumidor tem interesse por um produto ou serviço, é comum que ele procure referencias ou opiniões. Empresas que vendem produtos e disponibilizam serviços também são motivadas a ter conhecimento das opiniões dos consumidores, tendo que procurar formas de analisar essas informações para conduzir ações de marketing e tomada de decisão.

Nesse contexto, segundo Anchiêta et al. (2021), pesquisadores da área de Processamento de Linguagem Natural (PLN) tem buscado extrair informações úteis de dados não estruturados. É possível predizer sentimentos sem necessitar um olhar humano em textos, áudio ou vídeos. As pessoas hoje em dia tendem a se expressar muito por meio de sites e redes sociais, e empresas como Google e Facebook, por exemplo, já utilizam o conteúdo produzido por seus usuários para criar modelos preditivos poderosos. O termo Big Data se refere ao massivo conjunto de informações digitais coletadas. Todos os dias são criados 2.5 quinquilhões de bytes de dados, tanto que 90% dos dados do mundo hoje foram criados apenas nos últimos dois anos. Essa grande quantidade de dados faz com que a análise manual se torne uma tarefa impossível, sendo necessária a criação de métodos automáticos para analisar os dados (LIU et al., 2010).

A análise de sentimentos, também conhecida como mineração de opinião é um campo que consiste em extrair e analisar emoções, opiniões, sentimentos, avaliações sobre produtos, eventos, serviços ou qualquer outro assunto que seja possível se ter opinião (BECKER, 2013).

Os estudos em análises automáticas de documentos vêm permitindo avanços no reconhecimento de aspectos subjetivos. Dentre eles está a classificação de polaridade do texto, ou melhor, o quão positiva e negativamente são as opiniões descritas por pessoa.

A análise de sentimentos pode ser usada em algum sistema ou como ferramenta de apoio a decisão. Isso devido a propagação de opiniões, comentários e avaliações, sobre produtos, filmes e empresas. Empresas que oferecem serviços de Streaming, podem identificar quais filmes são mais populares e bem recebidos pelo público, contribuindo na seleção e aquisição de conteúdo de alta qualidade para seus catálogos.

Existem diferentes tipos de algoritmos para solucionar problemas com o método de classificação, como por exemplo: árvores de decisão, vizinhos mais próximos (KNN),

e regressão logística. A escolha do algoritmo dependerá do tipo e estrutura da base de dados, objetivo, quantidade de atributos, poder computacional disponível, entre outras coisas (KOTU e DESPANDE, 2019).

Neste projeto, será desenvolvido um modelo de aprendizado de máquina supervisionado para tarefas de classificação dos comentários dos reviews de filmes do site IMDB

O modelo de classificação escolhido para o projeto será o de Regressão Logística, utilizando o framework da biblioteca Scikit-Learn em Python.

2 SOBRE A MACKFLIX

A Mackflix é uma empresa fictícia, criada pelo grupo, com fins educacionais. É uma plataforma de streaming que precisa melhorar seu catálago de filmes para atrair novos clientes e aumentar a retenção desses clientes. Além disso, precisa de auxilio com as campanhas de marketing que não vem apresentando o retorno esperado, bem como o lucro.

3 OBJETIVOS E METAS

Aplicar um modelo preditivo capaz de, a partir dos comentários feitos por usuários do site IMDB (Internet Movie Database), avaliar a polaridade dos textos e classificá-los como positivos ou negativos utilizando técnicas de análise de sentimentos. Essa ferramenta desenvolvida pela equipe de Ciência de Dados da empresa Mackflix , será utilizada como apoio na tomada de decisão da plataforma de streaming Mackflix para analisar as preferências de gênero e temas de filmes pelo público, melhorando a qualidade do catálogo de filmes e aumentando a retenção de clientes. Atrair novos clientes através da elaboração de estratégias de marketing mais eficazes e aumentar os lucros.

4 METODOLOGIA

Neste capítulo serão descritas todas as etapas para a elaboração deste projeto.

4.1 COLETA DOS DADOS

O Dataset 'imdbmovies.csv' foi extraído do repositório Github do usuário Shreyas Wankhedeem. Os dados são públicos e não sensíveis. Consistem em dados do site IMDB referentes a 3.815 títulos de filmes e podem ser acessados utilizando o link:

https://github.com/shreyaswankhede/IMDb-Web-Scraping-and-Sentiment-Analysis

O Dataset 'imdbmovies.csv' apresenta as seguintes características:

Metadados:

- Formato do arquivo: CSV;
- Número de linhas: 3883;
- Número de colunas: 10.

Features:

- **IMDBID:** ID do filme;
- **Title:** Título do filme;
- **Genre:** Gênero do filme;
- **Year:** Ano de lançamento;
- **URL:** URL associada ao filme no IMDb;
- **Audience_Rating:** A classificação dada pelo público ao filme;
- **Critic_Rating:** A classificação dada pela crítica especializada ao filme;
- **Budget_In_Millions:** O orçamento do filme em milhões de dólares;

- **User_Review:** OS comentários em inglês escritos por usuários do IMDb sobre o filme;
- **Polarity:** A polaridade associada às revisões dos usuários.

Para leitura dos dados utilizamos a biblioteca Pandas.

Importamos a biblioteca Pandas e a base de dados para a estrututra de DataFrame: import pandas as pd df = pd.read_csv("/content/imdbmovies.csv", sep=";")

4.2 ANÁLISE EXPORATÓRIA DOS DADOS

A análise exploratória de dados é uma etapa fundamental em qualquer projeto de ciência de dados, incluindo a análise de sentimentos em revisões de filmes. Ela nos permite compreender melhor a natureza dos dados, identificar tendências, padrões e características importantes.

Importação de Dados: Os dados foram carregados a partir de um arquivo CSV chamado "imdbmovies.csv"usando a biblioteca Pandas. O arquivo contém informações sobre revisões de filmes, como título, gênero, ano de lançamento, avaliações de público e crítica, orçamento do filme, revisões de usuários e polaridade associada.

Visualização Inicial: As cinco primeiras linhas do DataFrame foram exibidas com o método head(). Isso fornece uma visão geral das primeiras entradas dos dados.

Limpeza de Dados: Colunas desnecessárias, como 'IMDBID', 'URL', 'Audience_Rating', 'Critic_Rating' e 'Budget_In_Millions', foram removidas com o método drop(). Isso simplifica o conjunto de dados.

Verificação de Valores Nulos: A existência de valores nulos foi verificada usando o método isnull().sum(). Haviam 67 valores nulos na coluna 'User Review', 4 valores na coluna 'Genre' e somente 1 valor na coluna 'Polarity'. Os valores nulos foram removidos com o método dropna().

Renomeação de Coluna e Mapeamento:

A coluna 'Polarity' foi renomeada para 'Label' para tornar mais clara a finalidade da coluna. Os rótulos textuais ('Positive' e 'Negative') foram mapeados para valores numéricos (1 e 0) na coluna 'Label'.

Verificação dos números de Instâncias e Atributos do Dataframe:

Verificar número de linhas e colunas após a remoção prévia de colunas desnecessárias e valores nulos utilizando o método shape.

Distribuição de Avaliações: A distribuição das avaliações foi analisada com o método value_counts(). Verificou-se que o conjunto de dados é balanceado, pois possui um número razoável de avaliações positivas e negativas.

Análise da Coluna 'Year': A coluna 'Year' foi explorada para identificar os anos de lançamento dos filmes presentes no conjunto de dados. Isso pode ser útil para análises temporais, para entender como as revisões de filmes mudam ao longo do tempo e segmentação por ano para a Mackflix.

Quantidade de Filmes Únicos: A quantidade de filmes únicos foi calculada com base na coluna 'Title'. Isso nos ajudou a entender a diversidade de filmes no conjunto de dados.

Nuvens de Palavras: Foram criadas nuvens de palavras para revisões no conjunto de dados, destacando as palavras mais frequentes em revisões positivas como visto na figura 1 e negativas como visto na figura 2. Isso fornece uma representação visual das palavras-chave associadas a cada sentimento. A análise exploratória de dados é crucial para preparar os dados, entender a distribuição de sentimentos, avaliar a estrutura dos dados e identificar características-chave. Isso ajuda na escolha de técnicas de pré-processamento e modelagem apropriadas e na interpretação dos resultados finais do modelo.

Figura 1 – Nuvem de palavras para comentários positivos

Fonte: Imagem criada pelo autor

Figura 2 – Nuvem de palavras para comentários negativos

Fonte: Imagem criada pelo autor

4.3 PRÉ-PROCESSAMENTO

O pré-processamento de texto desempenha um papel fundamental na análise de sentimentos com modelos de Regressão Logística, assim como em muitos outros tipos de tarefas de Processamento de Linguagem Natural (NLP). O objetivo do pré-processamento de texto é preparar os dados de texto de entrada de forma que eles possam ser usados de maneira eficaz por um modelo de Regressão Logística ou qualquer outro modelo de aprendizado de máquina.

Detalhe das etapas de pré-processamento dos dados textuais:

- Limpeza de Dados: Remoção de caracteres especiais, pontuação, números e qualquer informação irrelevante.
- Tokenização: Divisão dos textos em palavras individuais (tokens).
- StopWords: Remoção de palavras comuns (stop words) que não contribuem significativamente para a análise.
- Normalização: Conversão de todas as letras para minúsculas para evitar diferenciação entre maiúsculas e minúsculas.

- Lematização: redução de palavras a sua forma base ou raiz, conhecida como "lemma".

Para o Pré-Processamento e posteriores tarefas utilizaremos as bibliotecas Scikit-Learn, Pandas, NTLK, Re, Spacy e String.

4.4 DIVISÃO DO CONJUNTO DE DADOS

Para evitar o sobreajuste do Modelo de Regressão Logística, temos que definir dois conjuntos diferentes de dados a partir dos dados originais: um conjunto de Treinamento e um conjunto de Teste. O conjunto de treinamento X-train, y-train será a parte dos dados empregada para o treinamento dos parâmetros do modelo. O conjunto de teste X-test, y-test será a parte dos dados empregada para avaliar o modelo preditivo ajustado. 'X' representa a variável preditora e 'Y' representa a variável objetivo do Modelo. Essa divisão do conjunto original de dados deve ser aleatória e pode ser obtida com a biblioteca Scikit-learn, empregando-se a função train-test-split().

4.5 MÉTODO ANALÍTICO: TF-IDF E REGRESSÃO LOGÍSTICA

O Método Analítico aplicado no projeto da Mackflix envolve a utilização de técnicas de Aprendizado de Máquina Supervisionado para analisar e interpretar sentimentos em textos. Neste caso, a abordagem escolhida foi a Regressão Logística, um método amplamente utilizado na classificação de dados.

Na análise de sentimentos, as variáveis independentes são tipicamente recursos derivados do texto (como frequência de palavras, presença de certas palavras-chave, etc.), enquanto a variável dependente é a polaridade do sentimento (positiva ou negativa). O desafio é transformar o texto em um formato que possa ser processado por algoritmos de aprendizado de máquina, um processo conhecido como Vetorização ou Feature Extraction.

Para a Vetorização ou Feature Extraction utilizaremos a técnica TF-IDF (Term Frequency-Inverse Document Frequency), que é uma técnica estatística usada para avaliar a importância de uma palavra em um documento, na qual faz parte de uma coleção ou corpus. O TF-IDF transforma o texto bruto em um vetor de features numéricas, representando a importância de cada palavra para a classificação do sentimento do documento.

Term Frequency (TF) calcula a frequência com que um termo ocorre em um documento. Se um termo aparece muitas vezes em um documento, é importante e recebe um alto score de TF. A fórmula para calcular o TF de um termo "t" em um documento "d" é a seguinte:

TF(t, d) = (Número de vezes que o termo "t" aparece no documento "d") / (Total de palavras no documento "d")

Inverse Document Frequency (IDF) diminui o peso para termos que aparecem frequentemente em documentos do corpus. Isso é feito porque termos como "e"ou "o"podem aparecer muitas vezes, mas têm pouca importância. O IDF é calculado a partir da seguinte fórmula:

IDF(t, D) = log[(Total de documentos no corpus “D”) / (Número de documentos que contêm o termo “t”)]

Para calcular o TF-IDF de um termo em um documento, basta multiplicar o TF pelo IDF:

TF-IDF(t, d, D) = TF(t, d) * IDF(t, D)

Quanto maior o valor do TF-IDF para um termo em um documento, mais relevante e importante ele é no contexto do corpus.

Portanto, o TF-IDF é o produto do TF e do IDF, proporcionando uma pontuação que destaca palavras que são mais interessantes, ou seja, frequentes em um documento específico e menos comuns no corpus em geral.

O resultado é uma matriz onde cada linha representa um documento e cada coluna representa um termo, com valores que indicam a importância do termo no contexto dos documentos.

A importação da classe TfidfVectorizer da biblioteca Scikit-Learn foi utilizada neste projeto para a criação de representações numéricas de texto com base em TF-IDF.

A Regressão Logística pode então utilizar esses vetores para treinar um modelo capaz de prever a polaridade dos sentimentos de novos reviews. A combinação dessas técnicas é particularmente útil, pois a TF-IDF fornece uma maneira ponderada e normalizada de quantificar palavras em textos, enquanto a Regressão Logística utiliza essas quantificações para efetivamente distinguir entre diferentes classes de sentimentos.

A Regressão Logística, em sua essência, é um modelo estatístico que, ao contrário da regressão linear que prevê valores contínuos, é usado para previsão de variáveis categóricas, tipicamente binárias. No contexto de análise de sentimentos, essas categorias podem ser "positivo"ou "negativo".

Este modelo estima a probabilidade de uma variável dependente com base em uma ou mais variáveis independentes. O resultado é uma função logística (ou sigmóide) Essa função tem uma característica crucial: ela transforma qualquer valor de entrada em um número entre 0 e 1, o que é interpretado como probabilidade. Em termos matemáticos, se “*p*” é a probabilidade de um sentimento ser positivo, a função sigmóide pode ser representada como:

$$p = \frac{1}{1 + e^{-(b_0 + b_1 x)}}$$

Onde $b0$ e $b1$ são parâmetros do modelo e x representa a variável independente.
A probabilidade p indica a tendência de um texto ser classificado como positivo.

Este modelo é especialmente útil na análise de sentimentos porque pode lidar com

as sutilezas e ambiguidades do texto, oferecendo uma medida quantitativa da inclinação sentimental.

Além disso, abordagens de Processamento de Linguagem Natural (NLP) como a tokenização, e a remoção de stop words, a lematização são empregadas para refinar a extração de características, aumentando assim a precisão do modelo.

4.6 TREINAMENTO DO MODELO DE REGRESSÃO LOGÍSTICA

Devemos importar a classe Logistic Regression do módulo sklearn.linear-model da biblioteca Scikit-Learn, criar uma instância do modelo e treinar o modelo usando o conjunto de treinamento. Isso envolve ajustar os pesos (coeficientes) do modelo para minimizar a função de perda. Usaremos o método .fit() para realizar o treinamento. Todo o algoritmo da função para préprocessamento de texto e treinamento de modelo de regressão logística, em Python, pode ser visto no Apêndice B.

4.7 AVALIAÇÃO DO MODELO

As métricas de avaliação são usadas para medir o desempenho de modelos de classificação. Utilizaremos as seguintes métricas: acurácia, precisão, recall, F1-score e matriz de confusão.

Cada métrica avalia diferentes aspectos do desempenho do modelo, como sua capacidade de classificar corretamente as instâncias positivas e negativas.

A acurácia é uma métrica que mede a proporção de previsões corretas feitas pelo modelo em relação ao total de previsões. É calculada dividindo o número de previsões corretas pelo número total de previsões e é uma métrica básica de desempenho.

A matriz de confusão é uma tabela que descreve o desempenho de um modelo de classificação em relação aos valores reais dos dados de teste. Ela mostra quantas previsões do modelo foram corretas e quantas foram incorretas para cada classe.

A matriz de confusão é composta por quatro elementos principais: - Verdadeiro Positivo (TP): Indica que o modelo previu corretamente uma instância como positiva. - Verdadeiro Negativo (TN): Indica que o modelo previu corretamente uma instância como negativa. - Falso Positivo (FP): Indica que o modelo previu erroneamente uma instância como positiva quando, na verdade, era negativa (um erro do tipo I). - Falso Negativo (FN): Indica que o modelo previu erroneamente uma instância como negativa quando, na verdade, era positiva (um erro do tipo II). A matriz de confusão permite avaliar a qualidade das previsões de um modelo em termos de suas capacidades de classificar corretamente as instâncias em relação às classes de interesse.

O relatório de classificação é uma tabela que fornece métricas de avaliação detalhadas para cada classe em um problema de classificação. Ele inclui métricas como precisão, recall (taxa de verdadeiros positivos), F1-score e suporte.

O relatório de classificação calcula várias métricas para cada classe: - Precisão: A precisão mede a proporção de instâncias classificadas como positivas que realmente são positivas (TP / (TP + FP - O recall mede a proporção de instâncias verdadeiramente positivas que foram corretamente classificadas como positivas (TP / (TP + FN)). - O F1-score é uma média harmônica da precisão e do recall e fornece um equilíbrio entre essas duas métricas. É particularmente útil quando as classes são desequilibradas. - O suporte é o número real de instâncias de cada classe nos dados de teste. - O relatório de classificação ajuda a entender o desempenho do modelo para cada classe individualmente, além da acurácia geral do modelo.

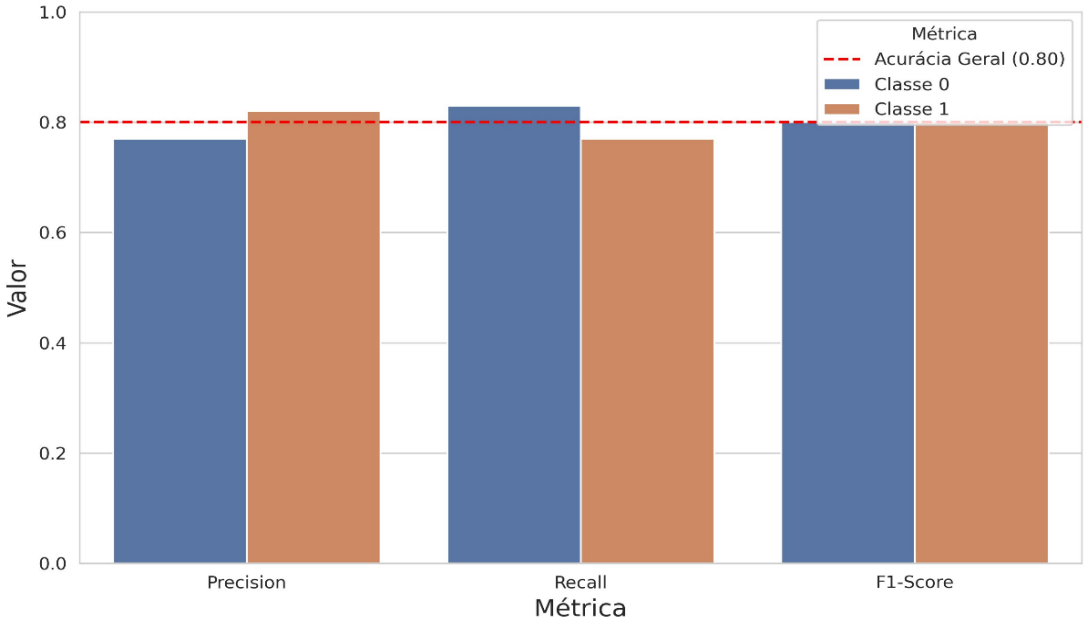
5 RESULTADOS

Um modelo preditivo foi criado, testado e validado, para realizar a análise de sentimentos e classificar as polaridades dos reviews dos filmes no site IMDB com os seguintes resultados:

- A acurácia geral do modelo é de 0.80.
- Para a classe 0 (Negativo): Precisão = 0.77, Recall = 0.83, F1-Score = 0.80.
- Para a classe 1 (Positivo): Precisão = 0.82, Recall = 0.77, F1-Score = 0.80.
- Para a Média Macro e a Média Ponderada o valor é 0.80 (80%).

O modelo tem boas métricas de precisão, recall e F1-Score, com uma acurácia geral de 80%. Isso indica que o modelo é capaz de realizar uma classificação boa do conjunto de dados, conforme podemos verificar no gráfico a seguir:

Gráfico 1 – Métricas de desempenho do Modelo



Fonte: Gráfico criado pelo autor

Com relação a Matriz de Confusão observa-se :

- Verdadeiro Negativo (VN): o valor 464 representa o número de observações que foram corretamente preditas como classe negativa.

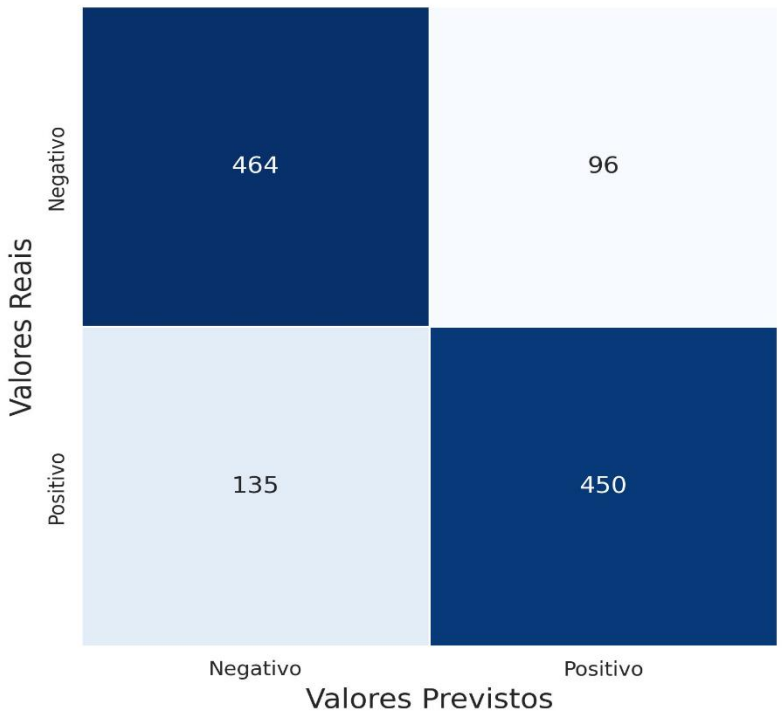
- Falso Positivo (FP): o valor 96 indica quantas vezes o modelo incorretamente predisse a classe positiva quando na verdade era negativa.

- Falso Negativo (FN): o valor 135 representa o número de vezes que o modelo predisse a classe negativa quando na realidade era positiva.

- Verdadeiro Positivo (VP): o valor 450 mostra quantas observações foram corretamente identificadas como classe positiva.

Podemos verificar a Matriz de Confusão pelo gráfico a seguir:

Gráfico 2 – Matriz de Confusão



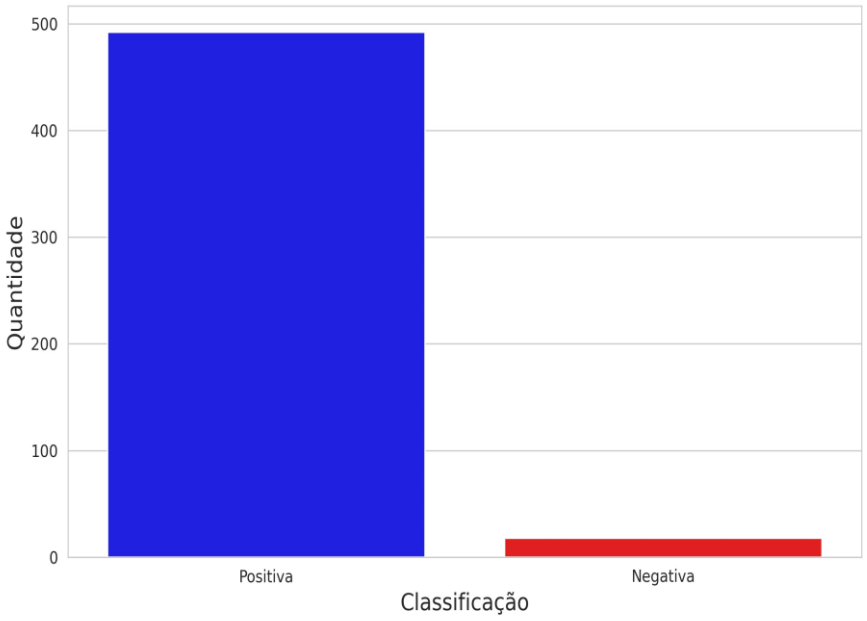
Fonte: Gráfico criado pelo autor

6 PRODUTO FINAL PARA A EMPRESA MACKFLIX

Foram selecionados três filmes do site IMDB (Internet Movie Database), para aplicação do modelo preditivo no conjunto de dados dos reviews de usuários deste site, e classificação das polaridades através da análise de sentimentos. Os filmes selecionados para a análise foram: Spider-Man: Across the Spider-Verse (2023), Tetris (2023) e John Wick: Chapter 4 (2023).

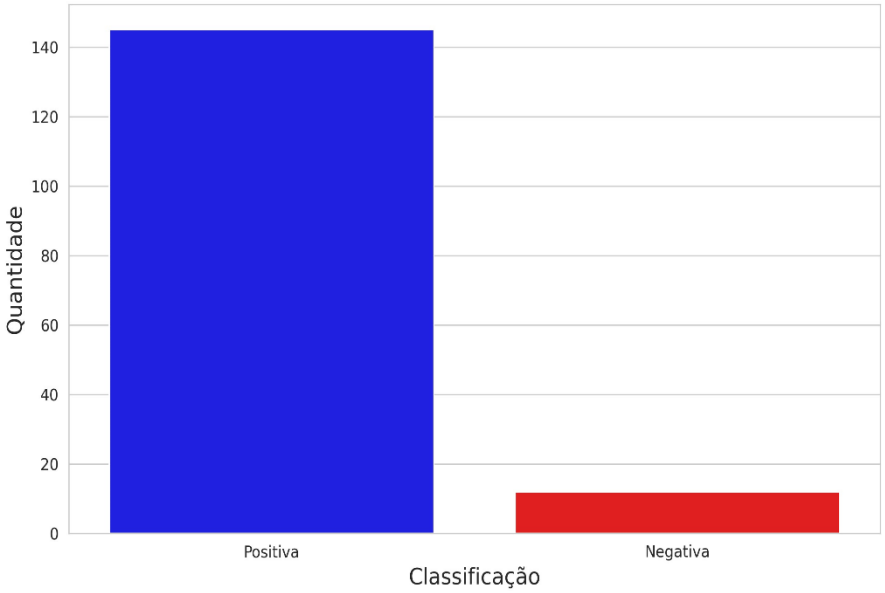
A análise de sentimentos dos três filmes visa a geração de insights valiosos para a área de marketing e aquisição de filmes da plataforma de streaming Mackflix. Os gráficos a seguir mostram o total de classificações positivas e negativas para cada filme:

Gráfico 3 – Total de classificações para Spider-Man: Across the Spider-Verse (2023)



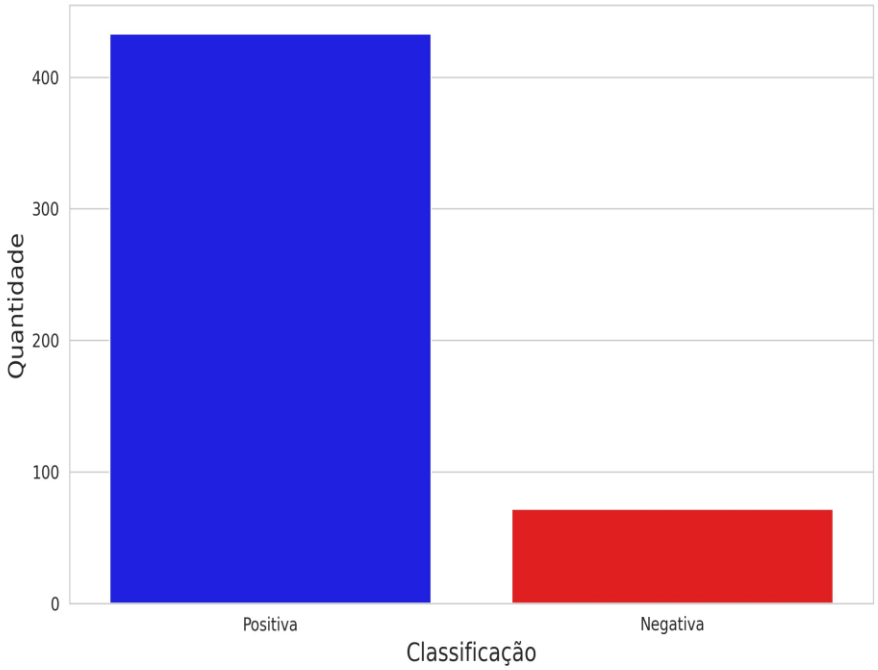
Fonte: Gráfico criado pelo autor

Gráfico 4 – Total de classificações para Tetris (2023)



Fonte: Gráfico criado pelo autor

Gráfico 5 – Total de classificações para John Wick: Chapter 4 (2023)



Fonte: Gráfico criado pelo autor

Podemos concluir que o filme Spider-Man: Across the Spider-Verse (2023), pode ser adicionado ao catálogo da Mackflix, pois obteve 492 avaliações positivas e somente 10 negativas. O filme Tetris (2023) também pode ser adicionado ao catálogo, pois obteve 145 avaliações positivas e somente 12 negativas. O filme John Wick: Chapter 4 (2023), também pode ser uma ótima aquisição para o catálogo da Mackflix pois obteve 433 avaliações positivas e somente 72 negativas. Cada filme representa um gênero diferente, o que permite a diversificação do catálogo com qualidade.

As análises revelam as preferências do público em relação a filmes. Portanto podemos identificar filmes com um grande número de revisões positivas, o que nos sugere que esses filmes podem ser populares entre os assinantes da plataforma. Além disso tendem a atrair mais espectadores. O engajamento dos usuários em torno desses filmes pode ser usado como métrica-chave para avaliar o sucesso da plataforma.

A análise temporal permite acompanhar as tendências ao longo do tempo. Identificar filmes populares ao longo dos anos pode ajudar a plataforma a criar um catálogo diversificado e atrativo e pode destacar quais gêneros de filmes têm uma recepção mais positiva. Isso ajuda na seleção de filmes que correspondem aos gostos da audiência. Sendo assim, filmes com revisões positivas podem ser promovidos ativamente na plataforma como parte de estratégias de marketing para atrair e reter assinantes.

Mediante as análises feitas, a Mackflix, pode desenvolver sistemas de recomendação, métricas de engajamento do usuário e estratégias de marketing específicas.

7 CONCLUSÕES

O projeto cujos resultados foram apresentados neste trabalho, foi proposto com o objetivo de desenvolver um modelo preditivo baseado em técnicas de aprendizado supervisionado, para avaliar a polaridade dos textos dos comentários feitos por usuários do site IMDB (Internet Movie Database), e classificá-los como positivos ou negativos utilizando técnicas de análise de sentimentos. Essa ferramenta será utilizada como apoio na tomada de decisão para analisar as preferências de gênero e temas de filmes pelo público, adquirir conteúdo de qualidade para o catálogo, aumentar a retenção de clientes, atrair novos clientes, elaborar estratégias de marketing mais eficazes e aumentar os lucros.

Para isso, utilizou-se uma base de dados obtida do repositório Github do usuário Shreyas Wankhedeem, com dados originalmente extraídos do site IMDB contendo cerca de 3.883 textos de reviews de filmes feitos por usuários. Essa base foi analisada, pré-processada, vetorizada e utilizada para treinamento e teste para um Modelo de Regressão Logística.

O pré-processamento de texto tem um papel fundamental na análise de sentimentos. Seu objetivo é preparar os dados de texto de entrada de forma que eles possam ser usados de maneira eficaz para um modelo de aprendizado de máquina. Envolve as seguintes etapas: limpeza de dados, tokenização, remoção de stopwords, normalização e lematização.

Na análise de sentimentos, as variáveis independentes são tipicamente recursos derivados do texto (como frequência de palavras, presença de certas palavras-chave, etc.), enquanto a variável dependente é a polaridade do sentimento (positiva ou negativa). O desafio é transformar o texto em um formato que possa ser processado por algoritmos de aprendizado de máquina, um processo conhecido como Vetorização ou Feature Extraction.

Para a Vetorização ou Feature Extraction utilizaremos a técnica TF-IDF (Term Frequency-Inverse Document Frequency), que é uma técnica estatística usada para avaliar a importância de uma palavra em um documento, na qual faz parte de uma coleção ou corpus. O TF-IDF transforma o texto bruto em um vetor de features numéricas, representando a importância de cada palavra para a classificação do sentimento do documento.

A Regressão Logística pode então utilizar esses vetores para treinar um modelo capaz de prever a polaridade dos sentimentos de novos reviews. A combinação dessas técnicas é particularmente útil, pois a TF-IDF fornece uma maneira ponderada e normalizada de quantificar palavras em textos, enquanto a Regressão Logística utiliza essas quantificações para efetivamente distinguir entre diferentes classes de sentimentos.

O modelo preditivo desenvolvido para analisar os sentimentos expressos nos reviews de filmes no IMDB demonstrou uma acurácia geral de 80%, equilibrando bem entre as classes de sentimentos positivos e negativos. Isso indica uma forte capacidade do modelo em classificar corretamente as opiniões dos usuários, sendo uma ferramenta valiosa para a compreensão das percepções do público em relação aos filmes.

A aplicação deste modelo nos comentários dos reviews de filmes como "Spider-Man: Across the Spider-Verse", "Tetris" e "John Wick: Chapter 4" resultou em uma predominância de avaliações positivas. Isso não apenas sugere que estes títulos seriam adições populares ao catálogo da Mackflix, mas também exemplifica como análises baseadas em dados podem orientar a seleção de conteúdo e a estratégia de aquisição de filmes.

As análises revelam as preferências do público em relação a filmes. Portanto a empresa Mackflix pode identificar filmes com um grande número de revisões positivas, o que sugere que esses filmes podem ser populares entre os assinantes da plataforma. Além disso tendem a atrair mais espectadores.

A análise temporal permite acompanhar as tendências ao longo do tempo. Identificar filmes populares ao longo dos anos pode ajudar a plataforma a criar um catálogo diversificado e atrativo. Filmes com revisões positivas podem ser promovidos ativamente na plataforma como parte de estratégias de marketing para atrair e reter assinantes. A Mackflix, pode desenvolver sistemas de recomendação, métricas de engajamento do usuário e estratégias de marketing específicas.

8 REFERÊNCIAS BIBLIOGRÁFICAS

BECKER, Karin. **Introdução à Mineração de Opiniões: Conceitos, Aplicações e Desafios**. Rio Grande do Sul: Sociedade Brasileira de Computação, 2013.

ANCHIÊTA, Rafael *et al*. PLN: Das Técnicas Tradicionais aos Modelos de Deep Learning. In: BEZERRA, Carla Ilane Moreira. **Minicursos da ERCEMAPI 2021**. Porto Alegre: Sociedade Brasileira de Computação, 2021. p. 9-33.

KOTU, V.; DESHPANDE, B. **Data Science: Concepts and Practice**. 2. ed. Cambridge: Morgan Kaufmann, 2019 | Biblioteca do Mackenzie.

LIU, B. et al. **Sentiment analysis and subjectivity. Handbook of natural language processing**, v. 2, n. 2010, 2010.

MONARD, Maria Carolina; BARANAUSKAS, José Augusto. Conceitos sobre aprendizado de máquina. São Carlos: ICMC-USP, 2002.

GAMA, João. Aprendizado de Máquina: Teoria e Aplicações. LTC Editora, 2014

Supervised Machine Learning. Geeks for Geeks, 2023. Disponível em: https://www.Geeksforgeeks.org/machine-learning. Acesso em: 15/10/2023.

HASTIE, Trevor; TIBSHIRANI, Robert; FRIEDMAN, Jerome. Elementos de Aprendizado de Máquina. Ciência Moderna, 2009.

SILVA, Diego F.; FERNANDES, David A.; CARVALHO, André C. Regressão logística: conceitos e aplicações em modelagem preditiva. Revista de Administração e Inovação, v. 14, n. 3, p. 282-295, 2017.

Logistic Regression Detailed Overview. Towards Data Science, 2023. Disponível em: https://towardsdatascience.com/logistic-regression-detailed-overview46c4da4303bc. Acesso em: 15/10/2023.

Machine Learning in Python. Scikit-learn, 2023. Disponível em: https://scikit-learn.org/stable/. Acesso em: 14/10/2023.

Python Machine Learning: Scikit-Learn Tutorial. Datacamp, 2023. Disponível em: https://www.datacamp.com/tutorial/machine-learning-python. Acesso em: 14/10/2023.

REZENDE, Solange Oliveira; RAMOS, Cláudio de Souza. Introdução à Mineração

de Textos com Aplicação à Análise de Sentimentos. Rio de Janeiro: Elsevier, 2016.

LOPES, Fábio; VIEIRA, Renata. Processamento da Linguagem Natural e suas aplicações: da teoria à prática. Bookman, 2018.

Natural Lenguage Processing. Machine Learning Mastery, 2023. Disponível em: https://machinelearningmastery.com/natural-language-processing/. Acesso em: 14/10/2023.

BATISTA, Gustavo Enrique de Almeida Prado Alves. Preprocessamento de bases de dados para aprendizado de máquina. 2003.

Métricas de Classificação: Acurácia, Precisão, Recall e F1 Score . Paulo Vasconcellos, 2023. Disponível em: https://paulovasconcellos.com.br/como-saber-se-seu-modelo-de-machine-learning-está-funcionando-mesmo-a5892f6468b. Acesso em: 14/10/2023.

8 APÊNDICE A – Link de acesso ao repositório do grupo no github:

https://github.com/lucassribs/analise_sentimentos_em_filmes