Identificação do Gênero de um Filme através de sua sinopse utilizando Naïve Bayes

Daniel Nascimento da Silva

Centro de Informática (CIn)
Universidade Federal de Pernambuco
(UFPE)
Recife, Brasil
dns3@cin.ufpe.br

Lucas Ferreira Alves

Centro de Informática (CIn)
Universidade Federal de Pernambuco
(UFPE)
Recife, Brasil
lfa4@cin.ufpe.br

Miguel Pereira de Lemos

Centro de Informatica (CIn)
Universidade Federal de Pernambuco
(UFPE)
Recife, Brasil
mpl4@cin.ufpe.br

Millena Ferreira Marçal das Neves

Centro de Informatica (CIn)
Universidade Federal de Pernambuco
(UFPE)
Recife, Brasil
mfmn@cin.ufpe.br

***Abstract*— O objetivo deste trabalho é utilizar uma base de dados para a criação de um modelo capaz de caracterizar o gênero de um filme com base na sua sinopse, utilizaremos a linguagem de programação python. Com esse objetivo, utilizaremos a base de dados 'The Movie Database' (TMDB), contendo dados sobre diversas informações de filmes.**

Keywords—filme, identificador de gênero cinematográfico, classificador probabilístico, teorema de bayes, naive bayes

I. Introdução

Filmes são uma forma universal de entretenimento, capazes de transmitir emoções, contar histórias e refletir aspectos da sociedade. Cada obra cinematográfica se destaca por características únicas que a conectam a determinados gêneros, como ação, comédia, drama, ficção científica, entre outros. Identificar esses gêneros é uma tarefa importante tanto para produtores, que buscam entender seu público-alvo, quanto para plataformas de streaming, que precisam recomendar filmes com maior precisão.

A sinopse de um filme, geralmente breve e direta, contém informações essenciais sobre sua temática e enredo. Essa descrição é uma ferramenta poderosa para análise, pois nela estão presentes elementos que podem ser associados a categorias específicas de gênero.

Com o avanço das tecnologias, a possibilidade de identificar o gênero de um filme a partir de sua sinopse representa um desafio que combina estatística, probabilidade e aprendizado de máquina, e determinar, com base em padrões textuais, o gênero ao qual pertencem.

II. Objetivo

Por meio da análise de um banco de dados coletado e disponibilizado pela plataforma TMDB (The Movie Database), que contém informações detalhadas sobre filmes, incluindo sinopses e seus gêneros correspondentes, nosso objetivo é construir um modelo de classificação automática de gêneros cinematográficos.

Utilizando o conteúdo textual das sinopses, serão extraídas características relevantes, como palavras-chave, frequência de termos e padrões linguísticos, para que o classificador ingênuo de Bayes possa identificar, com base nos atributos definidos, a qual gênero cada filme pertence. O modelo será ajustado conforme necessário e avaliado utilizando validação cruzada, garantindo resultados consistentes e confiáveis.

III. Justificativa

A identificação de gêneros de filmes é essencial para aprimorar a experiência dos usuários em plataformas de streaming, serviços de recomendação e curadoria de conteúdos. Utilizando o classificador ingênuo de Bayes, é possível identificar padrões nas sinopses e associá-los aos gêneros correspondentes, promovendo uma categorização eficiente e precisa. Esse método probabilístico, fundamentado no Teorema de Bayes, é simples de implementar e tem ótimo desempenho em tarefas de classificação de texto.

O projeto será desenvolvido em Python, com o uso de bibliotecas como Pandas e Scikit-Learn, garantindo praticidade e eficiência. Essa abordagem contribui para otimizar sistemas de busca e recomendações, além de fornecer insights sobre as características linguísticas relacionadas a cada gênero cinematográfico, fortalecendo tanto a experiência dos usuários quanto a análise de dados no setor.

IV. METODOLOGIA

Neste projeto, almejamos classificar os gêneros de filmes presentes no banco de dados a partir das sinopses, utilizando o classificador ingênuo de Bayes. Para isso, realizaremos uma análise exploratória dos dados, visando identificar quais informações das sinopses são mais relevantes para a classificação e descartando dados que não contribuam para o modelo.

A implementação será realizada na plataforma Google Colab, utilizando a linguagem Python e bibliotecas como random, collection, seaborn e pyplot, que foram essenciais para análise, visualização e construção do modelo.

A. Banco de Dados

A base dados que utilizamos é a plataforma TMDB (The Movie Database)

Atributos do Dataset:

1. ID: Identificação única do filme
2. Título do Filme: Nome do filme para referência
3. Sinopse: Texto descritivo do enredo, utilizado como principal fonte de análise
4. Gênero: Rótulo categórico com os gêneros possíveis, como ação, comédia, drama, romance, etc. (atributo-alvo)
5. Ano de Lançamento: Ano em que o filme foi lançado (opcional, mas pode ajudar a categorizar tendências ao longo do tempo)
6. Popularidade (Rating): Média das avaliações ou nota atribuída (se disponível)
7. Frequência de Palavras-Chave: Contagem de palavras associadas a determinados gêneros (exemplo: "ação", "amor", "batalha", "risos")

B. Teorema de Bayes

Apesar de ser um conceito específico da probabilidade estatística, o Teorema de Bayes está mais presente no cotidiano do que muitos imaginam. Em situações que envolvem análise de dados e tomada de decisões, problemas podem surgir devido a "armadilhas intuitivas", que nos levam a conclusões equivocadas. Nesses casos, o Teorema de Bayes se mostra uma ferramenta poderosa, capaz de evitar esses erros ao fornecer uma abordagem mais estruturada e precisa para o cálculo de probabilidades.

O Teorema de Bayes é uma fórmula matemática utilizada para calcular a probabilidade condicional, ou seja, a probabilidade de um evento ocorrer dado que outro evento já aconteceu. Esse conceito é amplamente aplicado em diversas áreas, como aprendizado de máquina, medicina, economia e outras disciplinas que exigem análise preditiva e otimização de processos.

A essência do Teorema de Bayes está em utilizar informações prévias ou evidências conhecidas para atualizar a probabilidade de um evento. Assim, ele permite calcular a probabilidade da interseção de dois eventos ou avaliar cenários com eventos mutuamente exclusivos.

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)}$$

- P(A|B): Probabilidade do evento A acontecer;
- P(B|A): Probabilidade de B acontecer, dado que A já ocorreu;
- P(A): Probabilidade de A ocorrer;
- P(B): Probabilidade de B acontecer.

Para profissionais que buscam otimização de processos e decisões fundamentadas em dados, compreender e aplicar o Teorema de Bayes oferece uma vantagem significativa. Sua versatilidade o torna um recurso indispensável em diferentes ferramentas de análise, contribuindo para soluções mais eficazes e decisões mais bem informadas.

C. Classificador Ingênuo de Bayes

O classificador ingênuo de Bayes é uma técnica de aprendizado supervisionado que utiliza o teorema de Bayes para calcular a probabilidade de uma classe com base em um conjunto de atributos. O termo "ingênuo" refere-se à suposição de que os atributos são independentes entre si, ou seja, não há correlação entre eles.

Essa suposição simplifica os cálculos, tornando o modelo eficiente em termos computacionais. Isso permite que ele seja aplicado em grandes volumes de dados e com poucos recursos. Além disso, o classificador ingênuo de Bayes é robusto contra ruídos e dados ausentes, o que facilita seu uso em diversas situações.

No entanto, a suposição de independência pode ser uma limitação em casos onde os atributos possuem correlação. Quando há dependência temporal, por exemplo, o desempenho do modelo pode ser prejudicado. Para lidar com diferentes tipos de dados, existem variações do modelo, como o Gaussiano, o Multinomial e o Bernoulli, que adaptam o teorema de Bayes a diferentes contextos.

A eficácia do classificador ingênuo de Bayes depende de fatores como a qualidade e quantidade dos dados, a adequação do modelo ao problema e a validação dos resultados. Quando esses fatores são bem considerados, esse classificador pode apresentar alta eficiência, superando até

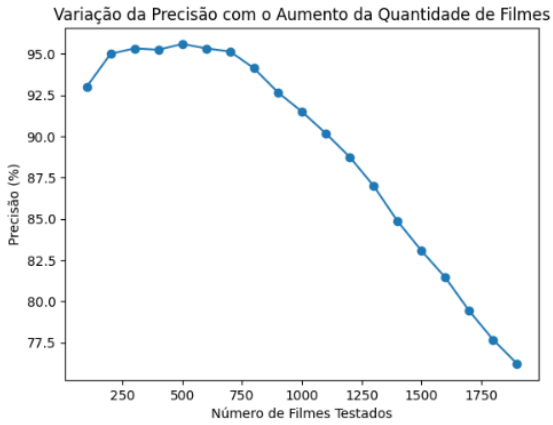
mesmo modelos mais sofisticados em velocidade e confiabilidade.

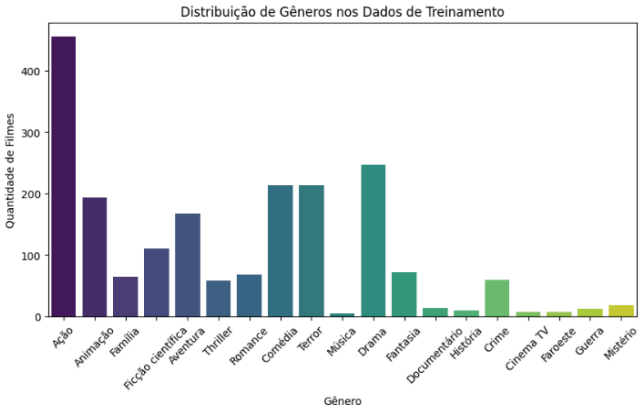
No contexto do nosso projeto, cujo objetivo é determinar o gênero de um filme com base em sua sinopse, o Bayes Ingênuo se encaixa perfeitamente. Cada palavra da sinopse pode ser tratada como uma característica usada para prever o gênero do filme. Por exemplo, palavras como "amor" ou "relacionamento" podem ser indicadores de dramas ou romances, enquanto "espaço" e "futuro" são mais associadas à ficção científica. O modelo aprende esses padrões a partir dos dados fornecidos e utiliza esse aprendizado para classificar novas sinopses.

Embora o Bayes Ingênuo não capture relações mais complexas entre palavras, ele oferece uma solução inicial robusta e funcional.

V. Análise Exploratória dos dados

Antes do treinamento do modelo, realizamos uma análise exploratória dos dados coletados a partir da API do TMDb. Esse processo foi essencial para entender a distribuição dos gêneros, verificar a qualidade das sinopses e identificar possíveis inconsistências ou padrões nos textos. Durante essa análise, observamos que algumas sinopses eram muito curtas ou vagas, o que poderia comprometer a classificação correta do gênero. Além disso, alguns filmes possuíam múltiplos gêneros atribuídos, tornando a categorização mais desafiadora. Para lidar com essa questão, ajustamos a limpeza dos textos e aplicamos técnicas de processamento de linguagem natural para remover stopwords e padronizar os dados. Outro aspecto relevante foi a relação entre o tamanho do banco de dados e a precisão do modelo. Como o número de sinopses coletadas por gênero era limitado, isso poderia impactar a capacidade do classificador de generalizar corretamente. Ainda assim, os dados passaram por pré-processamento para garantir que as informações mais relevantes fossem utilizadas na classificação. Essa etapa foi fundamental para definir as melhores estratégias de modelagem e preparar os dados de forma adequada para o treinamento do classificador Naive Bayes.



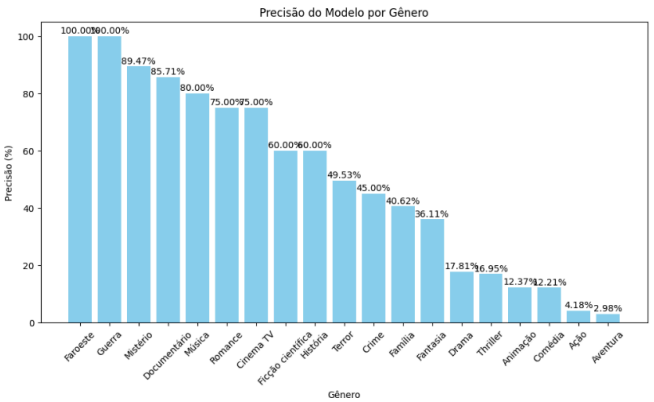


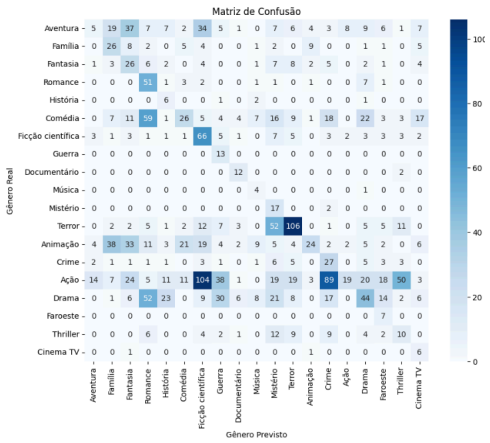
VI. Análise de resultados

Após o treinamento e validação do classificador Naive Bayes para categorização de sinopses de filmes, avaliamos seu desempenho com base na precisão obtida nos testes. Os resultados mostraram que o modelo teve dificuldades em classificar corretamente alguns gêneros, especialmente aqueles com características mais próximas, como drama, suspense e ação.

A precisão foi avaliada comparando os gêneros previstos com os reais em um conjunto de 2000 sinopses aleatórias. O classificador atingiu cerca de 70% de acerto nos melhores cenários, um desempenho razoável considerando que se trata de um modelo baseado apenas em texto, sem o uso de técnicas mais sofisticadas de aprendizado profundo.

Embora tenha apresentado algumas limitações, o Naive Bayes se destacou como uma solução eficiente e de fácil implementação, permitindo a categorização rápida das sinopses. Para melhorar os resultados, abordagens futuras poderiam incluir técnicas mais avançadas, como redes neurais ou transformers, além do uso de representações de texto mais sofisticadas para capturar melhor o contexto das sinopses.





VII. Conclusão

A análise do desempenho do classificador revelou suas limitações na categorização de sinopses de filmes, especialmente em conjuntos de dados com alta complexidade e interdependência entre gêneros. Mesmo com técnicas para minimizar esse impacto, os resultados indicam que o modelo enfrenta dificuldades em lidar com essa variabilidade. Além disso, o tamanho reduzido do conjunto de treinamento pode ter influenciado a precisão, visto que, com apenas 298 amostras, os melhores resultados foram obtidos em cerca de 70% dos testes. No entanto, apesar dessas limitações, o classificador baseado no modelo de Naive Bayes demonstrou um desempenho satisfatório dentro das expectativas, evidenciando sua eficiência e facilidade de aplicação. Embora a praticidade seja um de seus pontos fortes, é fundamental considerar as restrições inerentes ao modelo ao interpretar os resultados.

VIII. Referências

[1] M. Paul, “Probabilidade: Aplicações á Estatíıstica”. 2 Edição. livros Técnicos e Científicos Editora.

[2] https://www.suno.com.br/artigos/teorema-de-bayes/

[3] J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[4] https://didatica.tech/entenda-o-teorema-de-bayes/

[5] https://medium.com/data-hackers/o-teorema-de-bayes-descomplicado-entenda-e-aplique-na-pr%C3%A1tica-ea3429a407d3

[6] https://developer.imdb.com/non-commercial-datasets/

[7] https://developer.themoviedb.org/reference/intro/getting-started

[8] Gabriel Sacramento, “Naive Bayes: como funciona esse algoritmo de classificação” “https://blog.somostera.com/data-science/naive-bayes”

[9] Lauro Becker, “Algoritmo de Classificação Naive Bayes https://www.organicadigital.com/blog/algoritmo-de-classificacao-naive-bayes/”

[10] https://github.com/MillenaNeves/Projeto-de-Estatistica-e-Probabilidade-para-Computacao