**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 4](#_Toc191842790)

[2. INTRODUÇÃO 4](#_Toc191842791)

[3. TIPO DE ORGANIZAÇÃO 5](#_Toc191842792)

[4. OBJETIVO DO ESTUDO 6](#_Toc191842793)

[5. METADADOS 6](#_Toc191842794)

[6. CRONOGRAMA 7](#_Toc191842795)

[7. LINK GITHUB 8](#_Toc191842796)

**Lista de Figuras**

[Figura 1 - Logo da Organização 5](#_Toc191842566)

[Figura 2 - Informações gerais do dataset 7](#_Toc191842567)

# **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.

# **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.

# **METADADOS**

O IMDB *dataset* reúne avaliações de usuários sobre filmes já lançados, registradas na plataforma IMDB. Essas avaliações refletem a percepção do público e incluem comentários que expressam diferentes sentimentos em relação às produções cinematográficas. Este conjunto de dados será utilizado para o treinamento de algoritmos de aprendizado supervisionado, com o propósito de prever, com base nos comentários dos usuários, os sentimentos gerados e identificar em quais aspectos esses sentimentos ocorrem. Dessa forma, a análise poderá fornecer insights úteis para aprimorar ou ajustar os filmes e serviços relacionados.

O banco de dados do estudo possui formato CSV e foi obtido através do repositório público Kaggle não contendo qualquer tipo de conteúdo sensível que comprometer a pesquisa tão pouco informações pessoas de qualquer usuário que tenha participado das pesquisas de avaliação. No que tange as características do conteúdo do arquivo, é possível encontrar 2 colunas, sendo as avaliações contidas em 50.000 mil linhas e todas tipo *string*.

Texto

O conteúdo gerado por IA pode estar incorreto.A figura 2 abaixo ilustra as informações supracitadas e evidência que o arquivo também não apresenta valores nulos ou inexistentes que demandaram algum tipo de pré-tratamento durante a análise.

Figura 2 - Informações gerais do dataset

O conjunto de informações está distribuído da seguinte forma:

* Review: Descrição detalhada sobre a avalição de todos os usuários
* Sentiment: Classificação entre positivo e negativo sobre a avaliação descrita

# **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

# **LINK GITHUB**

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