relatório de atividades – 16/10/25

Disciplina: Mineração de Textos

Equipe:

Anderson de Matos Guimarães

Gustavo Stefano Thomazinho

Renan Ost

# objetivo da atividade

O objetivo da atividade de hoje foi desenvolver e validar a Fase 2 do projeto “Análise de Sentimentos em Comentários de Vídeos do YouTube”, aplicando técnicas de Processamento de Linguagem Natural (NLP) para classificar os comentários coletados na Fase 1 em positivos, negativos e neutros.

Essa etapa representa o estágio de Modeling dentro da metodologia CRISP-DM, transformando os dados textuais brutos em informações analíticas que permitirão, na próxima fase, a construção dos dashboards interativos e a integração com Flask.

# Atividades desenvolvidas

Durante a aula, as seguintes etapas foram implementadas e testadas:

## Estruturação do notebook analise.ipynb

Criação e organização do notebook da Fase 2, seguindo a estrutura modular e padronizada do projeto.

Inclusão de seções em Markdown explicando cada passo, conforme o modelo CRISP-DM.

## Carregamento dos dados da Fase 1

Importação do dataset comentarios\_top5.csv, gerado na pipeline de coleta.

Validação da integridade e visualização inicial dos dados.

## Pré-processamento textual

Implementação da função limpar\_texto() para remover URLs, menções, emojis e caracteres especiais, garantindo que o texto estivesse pronto para análise.

Criação de uma nova coluna comentario\_limpo no DataFrame.

## Detecção e filtragem de idioma

Aplicação da biblioteca langdetect para identificar o idioma de cada comentário.

Filtragem dos comentários em português (pt), assegurando a compatibilidade com o modelo de análise de sentimentos escolhido.

## Aplicação do modelo de NLP (HuggingFace)

Utilização do modelo “pierreguillou/bert-base-cased-sentiment-br”, específico para a língua portuguesa.

Geração das colunas sentimento (classificação textual) e score (confiança do modelo).

Testes adicionais com modelo multilíngue comentado (nlptown/bert-base-multilingual-uncased-sentiment), mantendo o padrão em português como o principal.

## Geração de estatísticas e visualizações

Cálculo da distribuição percentual dos sentimentos (positivos, negativos e neutros).

Criação de gráficos com Seaborn e Matplotlib, mostrando:

Distribuição geral dos sentimentos;

Comparativo entre vídeos analisados.

## Exportação dos resultados

Criação e salvamento do arquivo comentarios\_analisados.csv contendo os resultados completos da análise.

Estrutura final das colunas:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| video\_id | video\_titulo | comentario | comentario\_limpo | idioma | sentimento | score |

# Resultados Obtidos

A Pipeline 2 foi implementada com sucesso, gerando uma análise precisa dos sentimentos expressos pelos usuários nos comentários dos vídeos.

O modelo BERT em português apresentou alta consistência nos resultados, captando nuances emocionais e contextuais de forma satisfatória.

O dataset resultante está preparado para ser explorado estatisticamente e visualizado na próxima fase.