**CURSO:** Ciência de Dados

**Disciplina:** Engenharia de Dados

**ANÁLISE DE SENTIMENTOS: TWITTER**

04/12/2023

**TURMA:** Grupo 10

**MEMBROS:**

**Alison de Almeida Sales:**(Ass) \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Beatriz Monteiro:**(Ass) \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Giovanna Paola Lunetta:**(Ass) \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Leonardo Moreno:**(Ass) \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Mashara Arambasic:**(Ass) \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Victor Fagundes:**(Ass) \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Yago Angelini Candido:**(Ass) \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Sumário

[Introdução 4](#_Toc152791096)

[Resumo 4](#_Toc152791097)

[Abstract 4](#_Toc152791098)

[Objetivos 5](#_Toc152791099)

[Descrição geral do sistema 5](#_Toc152791100)

[Revisão do Tema 5](#_Toc152791101)

[Arquitetura 6](#_Toc152791102)

[Processos 6](#_Toc152791103)

[Visão Geral 6](#_Toc152791104)

[Fontes dos dados 6](#_Toc152791105)

[Base "train.csv" 6](#_Toc152791106)

[Base "Twitter\_Data.csv" 7](#_Toc152791107)

[Base "twitter\_training.csv" 7](#_Toc152791108)

[Técnicas de Processamento Aplicadas 8](#_Toc152791109)

[Consolidação de arquivos 8](#_Toc152791110)

[Padronização de classes 8](#_Toc152791111)

[Conversão de Tipo 8](#_Toc152791112)

[Transformação de Caixa de Texto 8](#_Toc152791113)

[Pontuação 8](#_Toc152791114)

[Stopwords 9](#_Toc152791115)

[Vetorização 9](#_Toc152791116)

[TF-IDF 9](#_Toc152791117)

[Treino/Teste 9](#_Toc152791118)

[Pipeline 9](#_Toc152791119)

[Análise 10](#_Toc152791120)

[Sistemas de análise 10](#_Toc152791121)

[Python 10](#_Toc152791122)

[SciKit Learn 10](#_Toc152791123)

[Jupyter Notebooks 10](#_Toc152791124)

[Resultados 10](#_Toc152791125)

[Conclusões 11](#_Toc152791126)

[Referências 11](#_Toc152791127)

[Bases de dados 11](#_Toc152791128)

[GitHub 11](#_Toc152791129)

# Introdução

## Resumo

O projeto de Análise de Sentimentos de Textos no Twitter se origina da necessidade de compreender as opiniões e emoções expressas em um dos maiores repositórios de informações em tempo real disponíveis - o Twitter. Identificar sentimentos em tweets pode ser vital para diversas aplicações, desde monitorar a satisfação do cliente até entender tendências de mercado e reações a eventos em tempo real.

## Abstract

The Text Sentiment Analysis project on Twitter originates from the need to understand the opinions and emotions expressed in one of the largest real-time information repositories available - Twitter. Identifying sentiments in tweets can be crucial for various applications, from monitoring customer satisfaction to understanding market trends and reactions to real-time events.

# Objetivos

O objetivo principal deste projeto é desenvolver um sistema de análise de sentimentos que seja capaz de classificar textos do Twitter em duas categorias principais: negativo ou positivo. Isso permitirá uma avaliação abrangente do sentimento geral associado a tópicos específicos, marcas, eventos ou tendências no Twitter.

# Descrição geral do sistema

Faremos a coleta de um grande volume de tweets do Twitter, abrangendo uma ampla gama de tópicos e fontes, então realizaremos tarefas de limpeza, transformação e remoção de ruídos para preparar os dados para análise. Desenvolveremos e treinaremos um modelo de machine learning, para classificar os tweets em categorias de sentimento. Avaliaremos a precisão do modelo usando métricas de desempenho, como precisão, recall e F1-score.

## Revisão do Tema

[FAZER TEXTO]

# Arquitetura

Diagrama

Descrição gerada automaticamente

# Processos

## Visão Geral

[FAZER TEXTO]

## Fontes dos dados

### Base "train.csv"

* Arquivo de valores separados por vírgulas que nos apresenta alguns tweets;
* A base tem as seguintes informações:
  + textID (ID do texto; texto);
  + text (o Tweet; texto);
  + selected\_text (o texto selecionado que está sendo levado em consideração; texto);
  + sentiment (o sentimento que o texto está passando, neutro, negativo ou positivo; texto);
    - Time of Tweet (em que horário do dia o tweet foi publicado, manhã, tarde, noite; texto);
    - Age of User (idade do usuário que publicou o tweet; intervalo numérico);
    - Country (País de origem do tweet; texto);
    - Population - 2020 (a população do país em 2020; número);
    - Land Area (Km²) (área do país; número);
    - Density (P/Km²) (Densidade do país, população dividida pela área; número).

Tela de computador com texto preto sobre fundo branco

Descrição gerada automaticamente

### Base "Twitter\_Data.csv"

* Arquivo de valores separados por vírgulas que nos apresenta alguns tweets;
* A base tem as seguintes informações:
  + clean\_text (o tweet inteiro; texto);
  + category (sentimento do texto, se é positivo, negativo ou neutro, sendo 1 positivo, 0 neutro e -1 negativo; número).

Interface gráfica do usuário, Aplicativo, Teams

Descrição gerada automaticamente

### Base "twitter\_training.csv"

* Arquivo de valores separados por vírgulas que nos apresenta alguns tweets que são as opiniões de usuários sobre alguns jogos.
* A base tem as seguintes informações:
  + TweetId (código de 4 números; texto);
  + Entity (nome da entidade em foco no tweet; texto);
  + Sentiment (sentimento do tweet, se é posivito, negativo, neutro ou irrelevante; texto);
  + TweetContent (tweet na íntegra; texto).

Interface gráfica do usuário, Aplicativo, Tabela

Descrição gerada automaticamente

## Técnicas de Processamento Aplicadas

### Consolidação de arquivos

[FAZER TEXTO]

### Padronização de classes

[FAZER TEXTO]

### Conversão de Tipo

[FAZER TEXTO]

### Transformação de Caixa de Texto

[FAZER TEXTO]

### Pontuação

[FAZER TEXTO]

### Stopwords

A manipulação feita válida de citação foi a remoção de Stopwords. As stopwords são palavras comuns que geralmente são removidas ao processar textos, pois não contribuem significativamente para o significado em uma análise textual.

Exemplos de stopwords incluem "e", "de", "para", "o", "a", entre outras. A remoção de stopwords é comumente usada ao processar texto para análise de sentimentos, classificação de documentos, entre outros.

**Para realizar a remoção de stopwords:**

* Precisa ter uma lista de palavras que deseja remover, geralmente obtida de bibliotecas de processamento de linguagem natural.
* Em seguida, você divide o texto em palavras individuais.
* Para cada palavra, você verifica se ela está na lista de stopwords.
* Se a palavra for uma stopword, ela é removida do texto; caso contrário, é mantida.

Essa técnica ajuda a reduzir o ruído e a dimensionalidade do texto, focando nas palavras mais significativas para a análise.

### Remoção de Tweets Nulos, Irrelevantes e Duplicados

[FAZER TEXTO]

### Normalização de Classes

[FAZER TEXTO]

### Vetorização

[FAZER TEXTO]

### TF-IDF

[FAZER TEXTO]

### Treino/Teste

[FAZER TEXTO]

### Pipeline

[FAZER TEXTO]

# Análise

## Sistemas de análise

Para a análise de sentimentos em tweets, o código fornecido utiliza principalmente a linguagem de programação Python e diversas bibliotecas especializadas em processamento de texto, análise de dados e machine learning. Vamos destacar as principais ferramentas e sistemas utilizados:

### Python

A linguagem de programação Python é a espinha dorsal do projeto. Ela é amplamente utilizada para desenvolvimento de aplicações de análise de dados e machine learning devido à sua sintaxe clara, vasta comunidade de desenvolvedores e uma variedade de bibliotecas especializadas.

### SciKit Learn

O código faz extenso uso da biblioteca SciKit Learn para implementar tarefas relacionadas à análise de sentimentos. Isso inclui a criação de pipelines para o pré-processamento de texto, vetorização, treinamento de modelos e avaliação de desempenho.

### Jupyter Notebooks

Jupyter Notebooks são usados para a execução do código de forma interativa, permitindo a visualização imediata dos resultados e facilitando a análise exploratória de dados.

Em resumo, o ecossistema Python, com suas bibliotecas especializadas, serve como a base técnica para a análise de sentimentos, enquanto ferramentas como Power BI fornecem meios eficazes para visualização e comunicação dos resultados. Essa abordagem integrada permite uma análise robusta e compreensível dos sentimentos presentes nos tweets da sociedade jovem.

## Resultados

### Classification Report

[FAZER TEXTO]

### Matriz de Confusão

[FAZER TEXTO]

O código fornecido realiza uma análise abrangente de dados de sentimentos em tweets, utilizando um modelo de classificação Naive Bayes Multinomial com o SciKit Learn em Python. A etapa de pré-processamento inclui a normalização do texto, remoção de pontuações e stopwords, além da tokenização dos tweets. A distribuição do comprimento dos tweets revela uma média de aproximadamente 107 caracteres, com uma variação significativa entre mensagens mais curtas e mais longas. A análise de sentimentos por meio do modelo treinado demonstra uma precisão global de cerca de 61%, sendo que as métricas detalhadas, como precisão, recall e F1-score, variam para cada classe de sentimento.

Ao explorar a distribuição de sentimentos, observa-se uma relativa equidade entre as classes, indicando que o modelo não possui grandes desequilíbrios na capacidade de prever diferentes tipos de sentimentos. O código também destaca mensagens extremamente curtas, algumas com apenas um caractere, sugerindo a presença de dados atípicos que podem ser examinados mais detalhadamente. Como sugestões de melhorias, o código ressalta a possibilidade de ajustes nos parâmetros do modelo, a exploração de diferentes algoritmos de aprendizado de máquina, e a realização de uma Análise Exploratória de Dados (EDA) mais aprofundada para obter insights adicionais sobre o conjunto de dados.

Em termos práticos, os resultados indicam que o modelo treinado é capaz de realizar uma classificação razoável dos sentimentos presentes nos tweets, com margem para aprimoramentos. Considerando o contexto do projeto, ajustes adicionais no pré-processamento de texto e na seleção do modelo podem contribuir para melhorar ainda mais a eficácia da classificação de sentimentos em dados do Twitter.

# Conclusões

[FAZER TEXTO SIMPLES, OBJETIVO, E SEM VIAGEM]

# Referências

## Bases de dados

* Twitter\_Data.csv: [Twitter Sentiment Dataset (kaggle.com)](https://www.kaggle.com/datasets/saurabhshahane/twitter-sentiment-dataset)
* Train.csv: [Sentiment Analysis Dataset (kaggle.com)](https://www.kaggle.com/datasets/abhi8923shriv/sentiment-analysis-dataset?select=train.csv)
* Twitter\_training.csv: [Twitter Sentiment Analysis (kaggle.com)](https://www.kaggle.com/datasets/jp797498e/twitter-entity-sentiment-analysis/data)

## GitHub

* <https://github.com/leomoreno11/trabalho_engenhariaDeDados/tree/main>