UNIVERSIDADE PRESBITERIANA MACKENZIE

BRENNO MONTEIRO DE OLIVEIRA
BRUNO FERNANDES MASCARINI
LUCAS CAMARGO SPINELI
MARINA CAMARGO SPINELI

PROJETO APLICADO II

ANÁLISE DAS PERCEPÇÕES DO CONTEÚDO DO TWITTER

São Paulo
2025

BRENNO MONTEIRO DE OLIVEIRA
BRUNO FERNANDES MASCARINI
LUCAS CAMARGO SPINELI
MARINA CAMARGO SPINELI

PROJETO APLICADO II

ANÁLISE DAS PERCEPÇÕES DO CONTEÚDO DO TWITTER

Projeto Aplicado II, do terceiro período do curso de Tecnologia em Ciência de Dados da Universidade Presbiteriana Mackenzie.

Prof. Dr. Felipe Albino dos Santos

São Paulo
2025

SUMÁRIO

1. INTRODUÇÃO

O suporte ao cliente nas redes sociais tem se tornado uma das principais formas de interação entre empresas e consumidores, e o X (Twitter), com sua natureza dinâmica e interativa, se destaca como uma das plataformas mais utilizadas para esse fim. Com mais de 3 milhões de tweets e respostas de grandes marcas, o conjunto de dados analisado neste estudo oferece uma oportunidade de explorar a comunicação entre empresas e consumidores, além de fornecer insights sobre a eficácia das práticas de suporte ao cliente modernas.

Este trabalho se propõe a investigar esses dados, visando compreender melhor as interações em tempo real, a análise da percepção do consumidor e o impacto das respostas oferecidas pelas empresas. A análise desses dados é relevante não apenas para aprimorar as estratégias de atendimento ao cliente, mas também para contribuir com a evolução da compreensão da linguagem natural e o desenvolvimento de modelos conversacionais. Ao longo deste estudo, exploraremos as características do conjunto de dados, suas vantagens em relação a outros corpora conversacionais e as possibilidades de aplicação, incluindo a previsão de respostas, o agrupamento de tópicos e a identificação de novos problemas que possam surgir no contexto do atendimento online.

Toda a documentação, ficará disponível no github, no link a seguir:
https://github.com/Brennu/Projeto-Aplicado-II

2. CRONOGRAMA

O cronograma a seguir ilustra as etapas a serem realizadas neste projeto. Cada integrante do grupo está representado por uma cor, indicando as respectivas responsabilidades de cada um; o "X" marca a expectativa da semana do mês de realização de cada parte do trabalho.

LEGENDA					
	BRENNO MONTEIRO DE OLIVEIRA		LUCAS CAMARGO SPINELI		TODOS
	BRUNO FERNANDES MASCARINI		MARINA CAMARGO SPINELI		

	CRONOGRAMA - PROJETO APLICADO II																	
	Lista de Atividades	**RESP.**	**FEVEREIRO**				**MARÇO**				**ABRIL**				**MAIO**			
Etapa 1	Definição da Organização			X														
	Apresentação do Dados (Metadados)			X														
	Objetivos e Metas			X														
	Cronograma de Atividades				X													
	Estrutura do Documento				X	X												
	Link para o Github				X													
Etapa 2	Definição da Linguagem de Programação						X											
	Análise Exploratória da Base de Dados							X	X									
	Tratamento da Base de Dados							X										
	Definição e Descrição das Bases Teóricas dos Métodos									X								
	Definição e Descrição da Acurácia									X								
Etapa 3	Aplicação do Método Analítico										X							
	Medidas de Acurácia											X						
	Descrição dos Resultados Preliminares												X					
	Esboço do Storytelling													X				
Etapa 4	Relatório Técnico														X			
	Apresentação do Storytelling															X	X	
	Repositório no GitHub																X	
	Vídeo da Apresentação																X	X

3. OBJETIVOS DO ESTUDO

O presente estudo tem como objetivo desenvolver um modelo capaz de analisar e classificar sentimentos em tweets, identificando padrões e tendências nas opiniões expressas pelos usuários. Para isso, será realizada uma investigação detalhada dos dados, incluindo a limpeza e o pré-processamento dos textos, a aplicação de algoritmos de aprendizado de máquina e a avaliação do desempenho dos modelos. Ao final, pretende-se construir uma ferramenta eficiente para prever a polaridade dos sentimentos em novas postagens no Twitter, contribuindo para uma melhor compreensão do comportamento dos usuários na plataforma.

4. CONTEXTO DE ESTUDO

Com o crescimento das redes sociais, o Twitter tornou-se uma das principais plataformas onde os usuários compartilham suas opiniões sobre uma ampla variedade de assuntos, como produtos, serviços, eventos e figuras públicas. A grande quantidade de postagens diárias cria um ambiente rico em informações, mas ao mesmo tempo desafiador para análise manual. Diante desse cenário, a análise automatizada de sentimentos surge como uma ferramenta essencial para compreender as percepções dos usuários de forma rápida e eficiente.

Empresas e organizações utilizam a análise de sentimentos para monitorar a recepção de seus produtos, avaliar a satisfação do público e até mesmo antecipar crises de imagem. Ao identificar padrões nas emoções expressas nos tweets, é possível obter insights valiosos que auxiliam na tomada de decisões estratégicas.

Além disso, com o avanço do processamento de linguagem natural e aprendizado de máquina, é possível identificar nuances emocionais complexas, como sarcasmo e ironia. Isso torna a análise de sentimentos mais precisa e permite que empresas respondam rapidamente a problemas de clientes no Twitter. Assim, a análise de sentimentos se torna uma ferramenta estratégica, ajudando as organizações a adaptar suas estratégias e melhorar a experiência do cliente.

5. APRESENTAÇÃO DA EMPRESA

O X, anteriormente conhecido como Twitter, é uma plataforma de mídia social dedicada à comunicação em tempo real. Fundada em 2006 por Jack Dorsey, Biz Stone e Evan Williams, a empresa tem sede em São Francisco, Califórnia, EUA.

Missão: A missão do X é promover e proteger a conversa pública — ser a praça pública da internet.

Visão: Ser o melhor aplicativo para tudo que gira em torno de áudio, vídeo, mensagens, pagamentos e serviços bancários

Valores: Os valores do X refletem seu compromisso com a comunicação aberta e inclusiva. A empresa incentiva a comunicação sem medo para construir confiança, garantindo um ambiente onde os usuários possam se expressar livremente. Além disso, busca defender e respeitar a voz do usuário, promovendo um espaço de diálogo autêntico.

Segmento de atuação: O X opera no setor de tecnologia e redes sociais, oferecendo um espaço para debates, notícias, entretenimento e interação entre usuários, empresas e governos. Seu modelo de negócios inclui publicidade digital, assinaturas e monetização de conteúdo, atendendo a um público global.

Market Share: O X sofreu uma desvalorização de aproximadamente 79% desde a aquisição por Elon Musk, reduzindo seu valor de mercado para cerca de US$ 9,4 bilhões. Relatórios indicam que o número de usuários ativos mensais do X caiu para aproximadamente 335 milhões em 2024, uma redução em relação aos 368 milhões anteriores.

Iniciativas em Ciência de Dados: O X utiliza técnicas de ciência de dados para analisar sentimentos e tendências nas postagens dos usuários, auxiliando empresas a entenderem a percepção de suas marcas e produtos.

6. APRESENTAÇÃO DO DATASET

O dataset utilizado neste estudo "Twitter Sentiment Analysis" contém informações extraídas do Twitter e apresenta quatro tipos de dados principais. Cada entrada no conjunto de dados possui um identificador único que diferencia os tweets entre si. Além disso, há uma coluna que representa a categoria ou tópico principal abordado na postagem, como o nome de um produto, jogo ou assunto específico discutido pelos usuários.

Outro dado relevante presente no dataset é a classificação do sentimento expresso no tweet. Cada postagem é rotulada de acordo com a emoção predominante, podendo ser positiva, negativa, neutra ou irrelevante. Por fim, o conjunto de dados inclui o próprio texto do tweet, que contém a opinião do usuário sobre o tema abordado. Esses dados permitirão a aplicação de técnicas de Processamento de Linguagem Natural para compreender e classificar automaticamente os sentimentos expressos nas postagens.

7. SELEÇÃO DE BIBLIOTECAS EM PYTHON

Nós escolhemos Python para nossa análise exploratória e tratamento dos dados porque é uma linguagem poderosa, flexível e amplamente utilizada na área de ciência de dados. Com bibliotecas como Pandas e NumPy, conseguimos manipular e limpar os dados de forma eficiente, enquanto Matplotlib e Seaborn nos ajudaram a visualizar padrões e insights de maneira clara. Além disso, utilizamos a Scikit-Learn, uma biblioteca fundamental para a implementação de algoritmos de machine learning e avaliação do desempenho do modelo, fornecendo métricas como acurácia, precisão e F1-score. Outro ponto importante é a integração com ferramentas de banco de dados e aprendizado de máquina, o que nos permitiu trabalhar desde a preparação dos dados até análises mais avançadas, tudo no mesmo ambiente. Por fim, a grande comunidade e os recursos disponíveis online foram essenciais para resolver desafios que surgiram durante o projeto.

8. MÉTODOS UTILIZADOS

Para realizar essa análise de sentimentos em tweets, o projeto seguiu uma abordagem cuidadosa e estruturada, combinando várias técnicas de processamento de linguagem natural. Tudo começou com uma etapa fundamental de preparação dos dados, onde os tweets passaram por um processo de limpeza e padronização. Primeiro, todos os textos foram convertidos para letras minúsculas, eliminando diferenças entre maiúsculas e minúsculas que poderiam atrapalhar a análise. Em seguida, caracteres especiais, links e símbolos foram removidos usando expressões regulares, deixando apenas o conteúdo textual relevante. Palavras muito comuns, como artigos e preposições, foram filtradas por não agregarem significado para a análise de sentimentos.

Com os dados limpos, o sistema criou um dicionário completo das palavras presentes nos tweets, contando sua frequência de aparição. Inicialmente, trabalhou-se com palavras isoladas (chamadas de unigramas), mas logo se percebeu que analisar sequências de palavras (bigramas, trigramas e até quadrigramas) permitia capturar melhor o sentido das frases. Essa abordagem de n-grams foi crucial para diferenciar expressões como "não gosto" de simplesmente "gosto", mostrando como o contexto muda completamente o significado. Para ajudar na compreensão dos padrões linguísticos, foram geradas nuvens de palavras que visualizavam graficamente os termos mais frequentes em cada categoria de sentimento. Esses gráficos revelavam claramente como certas palavras estavam associadas a sentimentos positivos ou negativos.

O coração da análise foi um modelo de regressão logística, que aprendeu a associar padrões de palavras a sentimentos específicos. Esse modelo funcionava como um classificador que, ao encontrar certas palavras ou combinações de palavras, conseguia determinar com boa precisão se o tweet expressava uma opinião positiva, negativa ou neutra. A eficácia desse modelo foi impressionante, alcançando 90,79% de acerto nos testes e validação com 98,6%, demonstrando que mesmo técnicas relativamente simples podem ser extremamente eficazes quando bem aplicadas.

Sobre os códigos utilizados para o modelo, estarão no github do projeto.

9. CÁLCULO DE ACURÁCIA

Para avaliar o desempenho do modelo de classificação de sentimentos em tweets, será utilizada a métrica de acurácia. A acurácia é definida como a proporção de previsões corretas feitas pelo modelo em relação ao total de amostras analisadas. Sua fórmula é dada por:

Acurácia = Número de previsões corretas / Número total de previsões

Para calcular essa métrica, será utilizada a função *accuracy_score* da biblioteca Scikit-Learn, que compara os rótulos reais dos sentimentos com as previsões geradas pelo modelo. O cálculo será realizado conforme o código abaixo:

```
from sklearn.metrics import accuracy_score

# Rótulos reais (valores esperados)

y_true = ["positive", "neutral", "negative", "positive", "negative", "neutral", "irrelevant"]


# Previsões do modelo

y_pred = ["positive", "neutral", "negative", "neutral", "negative", "positive", "irrelevant"]


# Cálculo da acurácia

acuracia = accuracy_score(y_true, y_pred)


# Exibir o resultado formatado

print(f"Acurácia do modelo: {acuracia:.4f}")
```

10. ANÁLISE EXPLORATÓRIA DE DADOS

Segundo a análise exploratória dos dados detalhada no github, as palavras mais comuns em tweets positivos incluem termos de apreço e empolgação, enquanto tweets negativos frequentemente mencionam marcas específicas, sugerindo que são reclamações diretas. Além disso, a categoria "Irrelevant" apresenta certa semelhança com a negativa, o que pode impactar a precisão da classificação.

A modelagem com Regressão Logística demonstrou um desempenho satisfatório, com boa precisão na classificação dos tweets. No entanto, a distribuição das classes mostrou um leve desbalanceamento, o que pode influenciar a performance do modelo. A implementação de outras técnicas, como TF-IDF ou redes neurais, poderia ser explorada para aprimorar os resultados.