

ANÁLISE DAS PERCEPÇÕES DO CONTEÚDO DO TWITTER

BRENNO MONTEIRO DE OLIVEIRA – 10435051

BRUNO FERNANDES MASCARINI – 10433508

LUCAS CAMARGO SPINELI – 10442957

MARINA CAMARGO SPINELI – 10441143

Introdução

O suporte ao cliente no X (Twitter), tornou-se uma importante forma de comunicação entre empresas e consumidores. Este estudo analisa mais de 3 milhões de interações para entender melhor essa dinâmica. O foco está na percepção dos consumidores e na eficácia das respostas corporativas. A análise visa aprimorar o atendimento ao cliente e desenvolver modelos de linguagem natural. Também são exploradas aplicações como previsão de respostas, agrupamento de tópicos e identificação de novos problemas.

Objetivos do Estudo

Desenvolver Modelo

Construir um modelo capaz de analisar e classificar sentimentos em tweets, identificando padrões e tendências nas opiniões expressas pelos usuários.

Aplicar Técnicas de ML

Realizar 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

Construir Ferramenta Preditiva

Ao final, construir uma ferramenta eficiente para prever a polaridade dos sentimentos em novas postagens no Twitter.

Compreender Comportamento

Contribuir para uma melhor compreensão do comportamento dos usuários na plataforma do Twitter

Contexto do estudo

- O X (Twitter) se destaca como uma das principais redes sociais para compartilhamento de opiniões sobre produtos, serviços, eventos e figuras públicas.
- O grande volume diário de postagens torna inviável uma análise manual eficaz.
- Empresas utilizam a técnica de análise automatizada de sentimentos para monitorar a aceitação de seus produtos, avaliar a satisfação do público e antecipar possíveis crises.
- Com o avanço do PLN e do aprendizado de máquina, é possível detectar nuances como sarcasmo e ironia, tornando a análise mais precisa.

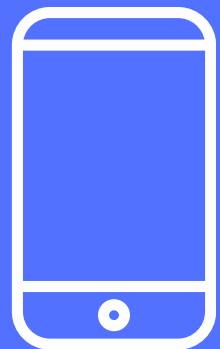
Twitter (X)

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

Promover e proteger a conversa pública, ser a praça pública da internet.



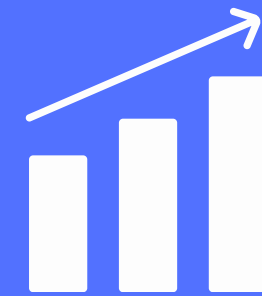
Visão

Ser o melhor aplicativo para tudo que gira em torno de áudio, vídeo, mensagens e serviços financeiros.



Valores

Destacar o compromisso com a comunicação aberta e inclusiva. Promover a liberdade de expressão para construir confiança. Incentivar a comunicação sem medo e valorizar a voz do usuário.



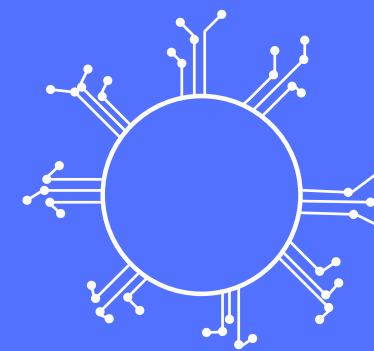
Segmento de Atuação

O X opera no setor de tecnologia e redes sociais, oferecendo um espaço para debates, notícias, entretenimento e interação entre usuários, empresas e governos.



Market Share

O X perdeu cerca de 79% de seu valor desde que foi comprado por Elon Musk e teve queda no número de usuários ativos mensais, de 368 para 335 milhões em 2024.



Data Science

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.

Descrição do dataset

- O estudo utiliza o dataset "Twitter Sentiment Analysis", que contém informações extraídas do X.
- Cada entrada no conjunto de dados possui um identificador único que diferencia os tweets.
- Inclui uma coluna que representa a categoria ou tópico principal abordado na postagem (ex: nome de um produto, jogo ou assunto específico).
- Cada postagem é rotulada com a classificação do sentimento expresso, podendo ser positiva, negativa, neutra ou irrelevante.
- O conjunto de dados também inclui o texto do tweet, que contém a opinião do usuário sobre o tema abordado.
- Estes dados permitem a aplicação de técnicas de Processamento de Linguagem Natural para compreender e classificar automaticamente os sentimentos expressos.

Pré-processamento de Texto

Padronização de texto:

- Conversão para letras minúsculas
- Garantia de que todos os registros são strings

Limpeza com expressões regulares:

- Remoção de pontuações e caracteres especiais
- Exclusão de emojis e URLs
- Filtragem de símbolos e ruídos visuais

Objetivo:

- Reduzir ruídos nos dados
- Melhorar a qualidade dos textos para vetorização
- Deixar os dados prontos para a modelagem de sentimentos

Visualizações e Análise por Marca

Para entender melhor os padrões de linguagem e o comportamento do público, foram geradas visualizações com base nos tweets classificados.

A análise foi dividida em duas abordagens principais:

- **Nuvens de Palavras** – destacam os termos mais frequentes de acordo com o tipo de sentimento.
- **Gráficos de Distribuição** – mostram o volume de tweets por marca e sentimento, permitindo comparações claras.

Essas representações ajudam a identificar padrões, tendências e diferenças no engajamento entre as marcas.

Nuvens de Palavras por Sentimento

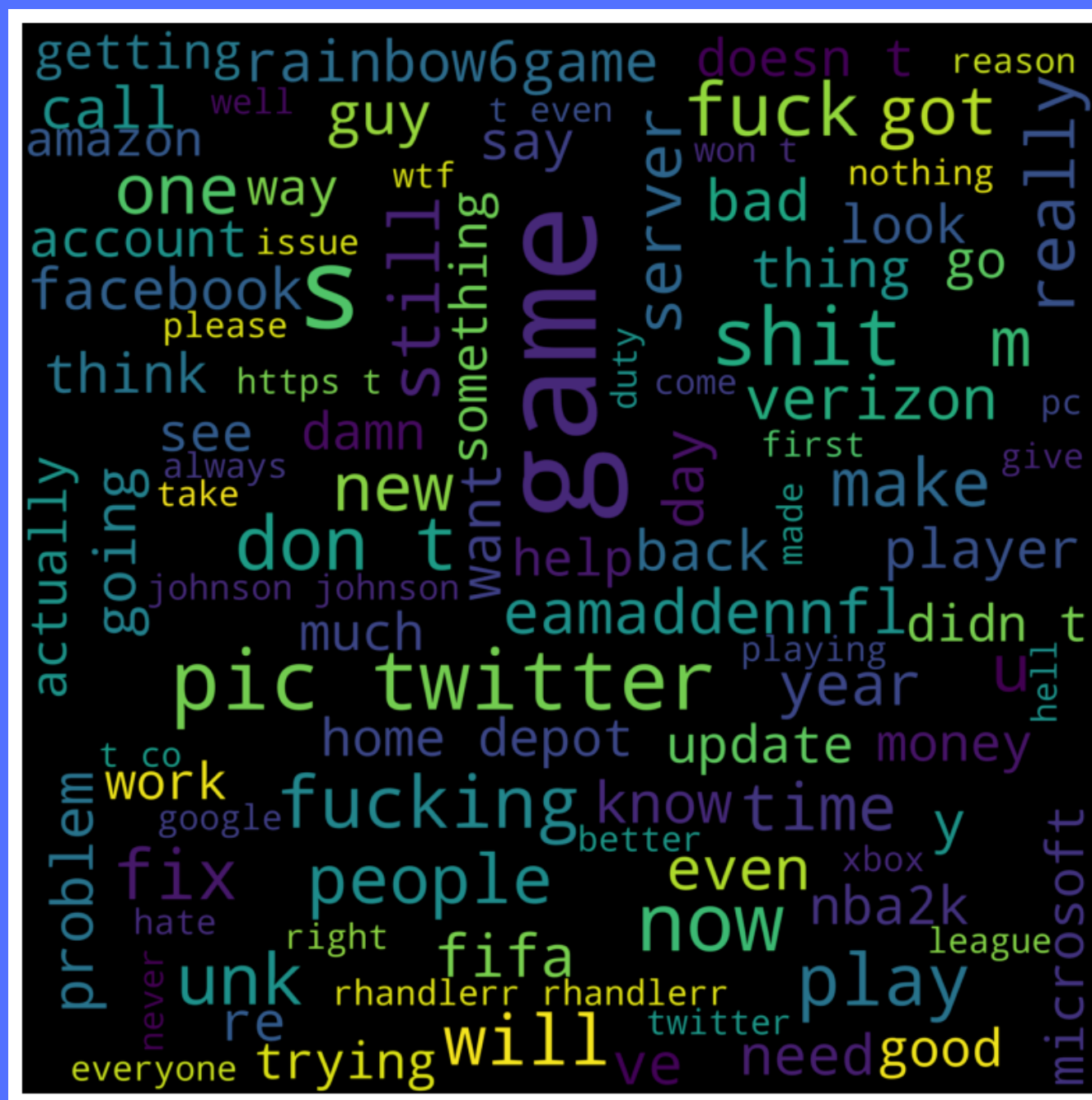
Gráfico 1 – Positive Tweets Word Cloud



Palavras mais frequentes em tweets positivos incluem termos como "love" e "game", além de outras palavras associadas a sentimentos positivos.

Nuvens de Palavras por Sentimento

Gráfico 2 - Negative Tweets Word Cloud



Tweets negativos apresentam palavras com frequência, além de menções a empresas/games específicos como 'facebook' e 'eamaddennfl'.

Nuvens de Palavras por Sentimento

Gráfico 3 - Irrelevant Tweets Word Cloud



Palavras um tanto similares às da categoria Negative.

Precisam de maior rigor na distinção entre essas classes durante o pré processamento ou rotulagem .

Nuvens de Palavras por Sentimento

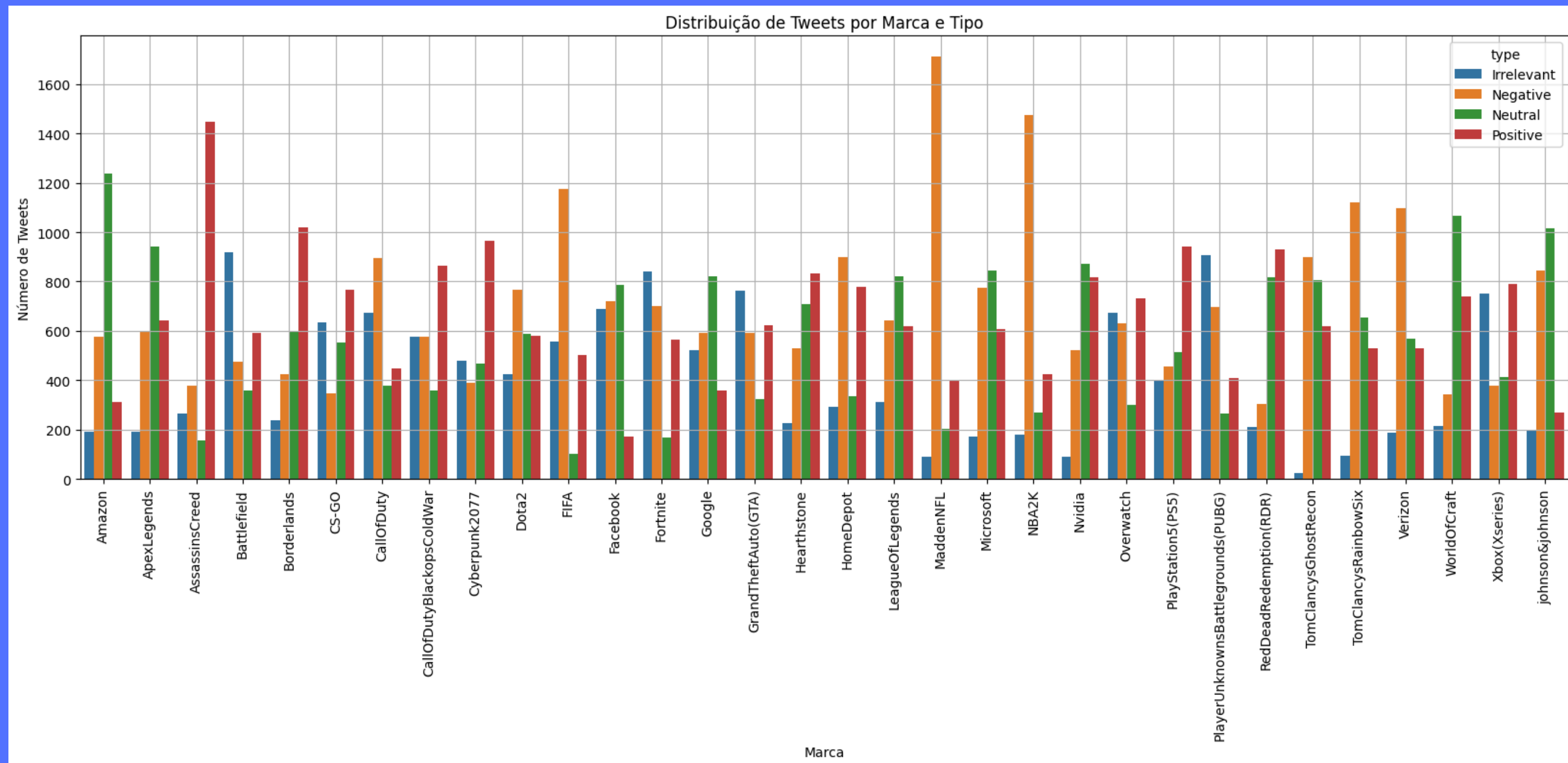
Gráfico 4 – Neutral Tweets Word Cloud



Apresenta perfil lexical distinto, com quase nenhum palavrão e palavras-chave diferentes das outras categorias, indicando maior neutralidade.

Distribuição de tweets por marca

Gráfico 5 – Distribuição de Tweets por Marca – Visualização em Barras



Tokenização

- Cada tweet foi transformado em uma lista de palavras (tokens)
- Exemplo: "Este é um tweet" → ['i', 'am', 'coming', 'to', 'the', 'borders', 'and', 'i', 'will', 'kill', 'you', 'all']
- **Resultado:**
 - Vocabulário total: **30.436 tokens únicos**
 - Alta dimensionalidade dos dados

Importância:

- Etapa essencial antes da vetorização
- Define a base da estrutura do modelo

Modelo de Regressão Logística

Configuração e Vetorização

Transformamos os tweets em vetores numéricos usando n-grams de 1 a 4 palavras, mantendo stopwords para preservar contexto.

Divisão e Análise das Classes

Os dados foram divididos em treino e teste com estratificação, revelando leve desbalanceamento entre as categorias de sentimento.

Treinamento com Avaliação dos Resultados

Aplicamos Regressão Logística com regularização. O modelo teve alta acurácia, indicando bom desempenho com possível sobreajuste.

Configuração Inicial e Vetorização

- Técnica Bag of Words com n-grams de 1 a 4 palavras
- Stopwords mantidas para preservar o contexto
- Vetorização aplicada **apenas no conjunto de treino**
- Prevenção de vazamento de informação





Matriz gerada com:

- 14.937 tweets
- 1.427.378 atributos únicos
- +803 mil ocorrências

Divisão dos Dados e Análise de Classes

Os dados foram separados em 80% para treino e 20% para teste, mantendo a mesma proporção de sentimentos em ambos os conjuntos.

A distribuição das classes mostrou um leve desequilíbrio:

-  Negativo – 29,9%
-  Positivo – 28,2%
-  Neutro – 24,6%
-  Irrelevante – 17,2%

Treinamento e Avaliação

O modelo de Regressão Logística foi treinado com regularização ($C = 0.9$) e até 1500 iterações.

Nos dados de teste, alcançou acurácia de **90.79%**.

Em validação externa, o desempenho foi ainda melhor, com **98.6%** de acurácia.

O uso de Bag of Words com n-grams se mostrou eficaz na classificação de sentimentos.

Proposta de Produto

Ferramenta de monitoramento de sentimentos em tempo real no Twitter. A plataforma permitiria que empresas acompanhassem a percepção de suas marcas por meio de um dashboard interativo, com gráficos de polaridade (positivo, negativo, neutro), tópicos mais mencionados e tendências ao longo do tempo. Além disso, seriam gerados alertas automáticos em caso de picos de menções negativas, permitindo uma resposta rápida a possíveis crises de imagem. O modelo de negócio seria baseado em planos mensais escalonados, com valores proporcionais ao volume de tweets analisados, por exemplo, um plano básico de até 10 mil tweets por R\$ 500, com opções mais avançadas para maiores demandas.



Considerações Finais

O projeto atingiu seu objetivo central de desenvolver um modelo de análise de sentimentos para classificar tweets, empregando um extenso pré-processamento textual e técnicas de vetorização como Bag of Words com n-grams em um dataset público de grande volume. O modelo demonstrou que, mesmo com métodos clássicos de PLN e machine learning, é possível construir uma ferramenta eficiente para análise de sentimentos em tempo real, alcançando alta acurácia na validação, o que indica um forte potencial para aplicação prática em ambientes corporativos. Com base nisso, foi proposto um produto escalável, um dashboard de monitoramento de sentimentos, e definiu-se como próximos passos a exploração de técnicas mais avançadas como TF-IDF e redes neurais, a expansão para outras plataformas e o aprimoramento na detecção de nuances emocionais.

Obrigado

Link da apresentação

[https://www.youtube.com/watch?
v=b_ih3C5Gbhs](https://www.youtube.com/watch?v=b_ih3C5Gbhs)