

**ESCOLA SUPERIOR DE PROPAGANDA E MARKETING**

**Projeto de Graduação em Tecnologia**

**Mineração de texto em conteúdo esportivo baseado  
em *web scraping* através de rede social**

Caique Thadeu Faria de Almeida  
Lucca Negrini Garcia

**São Paulo  
2024**

**ESCOLA SUPERIOR DE PROPAGANDA E MARKETING**

**Projeto de Graduação em Tecnologia**

**Mineração de texto em conteúdo esportivo baseado  
em *web scrapping* através de rede social**

Caique Thadeu Faria de Almeida

Lucca Negrini Garcia

Trabalho de conclusão de curso apresentado à  
Escola Superior de Propaganda e Marketing, como  
requisito para o recebimento do bacharelado em  
Sistemas de Informação.

**Orientador:** prof. Dr. Flavio Marques Azevedo

**São Paulo**

**2024**

Dedico este trabalho a todas as pessoas que compreendem o valor dos dados atribuídos aos negócios e reconhecem a importância das informações na vida cotidiana. Seja para direcionar estratégias de negócios, facilitar tarefas do dia a dia ou contribuir para avanços na área da saúde, como o diagnóstico precoce de doenças por meio de modelos de aprendizagem profunda. Expresso minha sincera gratidão a todos aqueles que compreendem o impacto transformador dos dados e sua aplicação na sociedade.

Além disso, dedico este trabalho com profunda gratidão aos meus pais, cujo apoio e estímulo ao longo de minha jornada acadêmica foram inestimáveis. Sua constante dedicação e incentivo foram fundamentais para meu progresso e sucesso.

Com estima,

**Lucca Negrini**

Dedico este trabalho aos meus pais, amigos e familiares que, ao longo da minha jornada acadêmica, me concederam apoio incondicional e estímulo constante. Seu suporte desempenhou um papel crucial no desenvolvimento deste trabalho de conclusão de curso.

Agradeço a todos que estiveram ao meu lado, proporcionando orientação, incentivo e compreensão. Este trabalho é fruto do apoio de cada um de vocês, e dedico-o em reconhecimento a essa valiosa contribuição.

Com gratidão,

**Caique Thadeu Faria de Almeida**

## **AGRADECIMENTOS**

Gostaria de expressar meus sinceros agradecimentos a todos aqueles que desempenharam um papel fundamental na realização deste projeto. Em primeiro lugar, expresso minha gratidão à instituição acadêmica que proporcionou um ambiente propício ao aprendizado, oferecendo uma infraestrutura robusta e contínuo apoio aos alunos.

Meus agradecimentos também se estendem aos estimados professores que desempenharam um papel crucial na minha formação acadêmica. Sua orientação, estímulos e conselhos moldaram minhas escolhas e deixaram uma marca indelével na minha trajetória.

Quero dedicar um agradecimento especial ao nosso orientador, cuja paciência, auxílio e orientação foram fundamentais para a realização deste trabalho com excelência.

Por fim, estendo meus agradecimentos às nossas famílias, que têm sido uma fonte constante de apoio e encorajamento em todos os momentos. Suas palavras de incentivo são inestimáveis e motivam nosso contínuo crescimento.

Com gratidão,

**Lucca Negrini e Caique Thadeu Faria de Almeida**

Negrini, Lucca Garcia

Mineração de texto em conteúdo esportivo baseado em web scrapping através de rede social / Lucca Garcia Negrini, Caique Thadeu Almeida. - São Paulo, 2024.

60 f. : il. p&b.

Trabalho de conclusão de curso (graduação) – Escola Superior de Propaganda e Marketing, Curso de Sistemas de Informação em Comunicação e Gestão, São Paulo, 2024.

Orientador: Flavio Marques Azevedo

1. redes sociais. 2. dados. 3. esportivo. I. Almeida, Caique Thadeu. II. Azevedo, Flavio Marques. III. Escola Superior de Propaganda e Marketing. IV. Título.

Ficha catalográfica elaborada pelo autor por meio do Sistema de Geração Automático da Biblioteca  
ESPM

## **RESUMO**

As redes sociais passaram por uma grande transformação desde seu surgimento, deixando de ser apenas plataformas simples de compartilhamento para se tornarem ecossistemas complexos com impactos significativos na vida das pessoas e na sociedade. Esse avanço tem reconfigurado a interação no cenário esportivo, permitindo que atletas e equipes se conectem diretamente com seus fãs e promovam eventos esportivos globalmente. No entanto, junto com os benefícios, surgem desafios como disseminação de mensagens negativas e assédio. Este estudo examina esses aspectos nas redes sociais, com foco no contexto do futebol, com o objetivo de analisar as manifestações de publicações negativas e positivas durante eventos esportivos. A pesquisa visa fornecer insights sobre melhores práticas de utilização de técnicas de Data Mining para otimizar o ambiente esportivo.

## SUMÁRIO

1. INTRODUÇÃO.....	10
1.1. Redes sociais e o futebol como esporte profissional .....	12
1.2. Web scrapping .....	13
1.3. Mineração de texto .....	14
1.4. Justificativa e hipóteses .....	15
1.5. Objetivo geral e objetivos específicos .....	16
2. REVISÃO DA LITERATURA .....	17
2.1 Mensagens de ódio nas redes sociais ligadas ao esporte .....	18
2.2. Natureza e tipologia das mensagens de ódio .....	18
3. MATERIAIS E MÉTODOS.....	20
3.1. Metodologia CRISP DM .....	20
4. DESENVOLVIMENTO DO PROJETO.....	23
4.1. Entendimento do negócio .....	23
4.2. Entendimento dos dados.....	23
4.3. Modelagem .....	30
4.4. Avaliação do Modelo.....	39
4.5. Implementação.....	40
5. RESULTADOS OBTIDOS .....	43
5.1. Análise de Sentimentos .....	43
5.2. Insights .....	44
5.3. Publicações Durante o Jogo.....	44
5.4. Exploração dos Tweets mais interativos .....	47
5.5. Avaliação do modelo com novas entradas de dados .....	48
6. Considerações Finais .....	51
6.1. Limitações do projeto .....	51
6.2. Futuras aplicações.....	51
7. REFERÊNCIAS BIBLIOGRÁFICAS .....	52
8. ANEXOS .....	56



## LISTA DE FIGURAS

Figura 1: Ilustração do modelo CRISP-DM. ....	20
Figura 2: Ilustração dos passos do planejamento. ....	21
Figura 3: As colunas que formam a estrutura da análise de dados. ....	25
Figura 4: Resultado da execução do scrapping de dados ....	26
Figura 5: As palavras-chave que serão buscadas na raspagem de dados. ....	26
Figura 6: Outras palavras-chave que serão buscadas na raspagem de dados. ....	27
Figura 7: Configurações de duração da raspagem no Apify ....	27
Figura 8: Opção dos dias em que os tweets foram realizados ....	28
Figura 9: Escolha de filtros para os tweets. ....	28
Figura 10: Retorno da raspagem de dados. ....	30
Figura 11: Import's das bibliotecas utilizadas ....	31
Figura 12: Dados analisados e processados com “sentimentos negativos” ....	32
Figura 13: Dados analisados e processados com sentimentos positivos. ....	33
Figura 14: Quantitativos de tweets positivos e negativos. ....	34
Figura 15: Passos utilizados para o tratamento dos dados. ....	35
Figura 16: Etapas para divisão dos dados entre treinamento e teste. ....	35
Figura 17: Modelo utilizado para lógica de regressão logística. ....	36
Figura 18: Modelo utilizado para lógica de árvore de decisão. ....	36
Figura 19: Modelo utilizado para lógica de floresta aleatória. ....	37
Figura 20: Modelo utilizado para lógica de SVM. ....	37
Figura 21: Modelo utilizado para lógica de KNN. ....	38
Figura 22: Gráfico que representa uma matriz de confusão e permite uma análise visual do desempenho do modelo. ....	39
Figura 23: Avaliação de desempenho do modelo treinado. ....	40
Figura 24: Conteúdo textual dos tweets, obtidos na raspagem. ....	41
Figura 25: Tweets sendo armazenados e classificados em positivos e negativos. ....	41
Figura 26: tweets positivos processados pelo modelo de análise. ....	43
Figura 27: tweets negativos processados pelo modelo de análise. ....	43
Figura 28: Resultado da análise de dados realizada no Power BI. ....	44
Figura 29: Divisão de resultados entre sentimentos positivos e negativos. ....	45
Figura 30: Distribuição dos resultados por positivo e negativo e por horário (antes e depois do início da partida). ....	45
Figura 31: Histórico de quantidade de tweets por hora (começando das 20:00). ....	46
Figura 32: Histórico de quantidade de tweets por hora (começando das 22:00). ....	46
Figura 33: Tweets que receberam mais interação de retweets (compartilhamento). ....	47
Figura 34: Top 10 tweets que receberam mais likes. ....	48
Figura 35: Top 10 tweets que receberam mais comentários. ....	48
Figura 36: Publicações criadas para teste. ....	49
Figura 37: Passos utilizados para chegar até o resultado final. ....	49

## 1. INTRODUÇÃO

A sociedade contemporânea tem sido profundamente influenciada pelo avanço tecnológico, que transformou a maneira como vivemos e interagimos no mundo. A tecnologia tornou-se um elemento essencial na vida das pessoas, impactando diversas esferas, desde o trabalho e o estudo até as atividades diárias.

Um dos desenvolvimentos mais marcantes desse cenário é a proliferação das redes sociais, que desempenham um papel significativo na forma como os indivíduos se conectam, compartilham informações e promovem produtos e serviços. Este trabalho aborda a influência das redes sociais na sociedade contemporânea, com foco nas mudanças que ocorreram no esporte e nos desafios que surgem em meio a essa evolução tecnológica (BOYD e ELLISON, 2007).

As redes sociais se estabeleceram como um canal de comunicação fundamental, permitindo que os usuários se conectem com amigos e familiares, independentemente da distância física. Além disso, elas proporcionam uma plataforma para vendedores promoverem produtos e serviços de maneira eficaz, alcançando um público amplo (BOYD e ELLISON, 2007).

É relevante notar que muitos usuários agora dependem das redes sociais como fonte de renda, criando um impacto notável em suas vidas. A organização de eventos, venda de produtos e serviços e até a busca por parceiros tornaram-se possíveis através dessas plataformas.

O mundo do esporte não escapou às transformações geradas pelas redes sociais. Elas revolucionaram a forma como atletas, equipes esportivas e organizações interagem com seus fãs. As redes sociais permitiram que atletas e equipes estabelecessem conexões diretas com seus seguidores, compartilhando não apenas seus desempenhos esportivos, mas também suas histórias, treinamentos e sucessos pessoais. Além disso, as redes sociais tornaram os eventos esportivos mais acessíveis ao público global, com a possibilidade de transmissões ao vivo, ampliando o alcance desses eventos (HINDUJA e PATCHIN, 2018).

As redes sociais se tornaram uma parte essencial do marketing esportivo. Elas oferecem uma plataforma eficaz para a promoção de marcas pessoais e o engajamento de comunidades de fãs. Atletas e equipes podem aproveitar as redes sociais para atrair patrocínios e promover marcas relacionadas ao esporte, criando oportunidades econômicas significativas (PRIMACK et al, 2017).

No entanto, o uso das redes sociais não está isento de desafios. A disseminação de discursos de ódio na internet é um problema crescente, com mensagens ofensivas que fomentam a discriminação e a hostilidade. Essas mensagens muitas vezes são publicadas anonimamente, e seu impacto pode levar a conflitos sociais, *cyberbullying* e até mesmo violência *offline*. Controlar esses discursos de ódio sem infringir a liberdade de expressão é uma tarefa complexa.

Além disso, o uso excessivo das redes sociais pode ter implicações negativas na saúde mental, levando ao isolamento social, ao vício digital e a problemas como ansiedade e depressão. Portanto, apesar dos benefícios das redes sociais em uma sociedade cada vez mais conectada, é fundamental abordar esses desafios com cautela e responsabilidade (Primack et al., 2017).

Desde o surgimento das redes sociais na década de 2000, com plataformas como o Facebook e o Instagram, até a proliferação de aplicativos de compartilhamento de fotos como o Instagram e o Snapchat, as redes sociais têm se tornado uma parte intrínseca da cultura esportiva. A capacidade de atletas, equipes esportivas, organizações esportivas e fãs de se conectarem, compartilharem conteúdo em tempo real e participarem de discussões sobre eventos esportivos, revolucionou a experiência esportiva. Como exemplos, de acordo com o autor, a influência das redes sociais nos esportes pode ser vista em diversas áreas:

- a) Engajamento dos Fãs: As redes sociais permitem que os fãs se envolvam diretamente com seus ídolos esportivos, acompanhem notícias e atualizações em tempo real e participem de conversas sobre seus esportes favoritos. Isso fortaleceu a relação entre atletas e fãs, bem como entre equipes e seus seguidores.
- b) Marketing Esportivo: As redes sociais se tornaram uma plataforma vital para o marketing esportivo. Através de campanhas, conteúdo exclusivo e parcerias estratégicas, equipes esportivas e marcas aproveitam o alcance das redes sociais para promover seus produtos e eventos (Boyd & Ellison, 2007).
- c) Transmissões e Cobertura: As transmissões ao vivo de eventos esportivos por meio de plataformas de redes sociais democratizaram o acesso a conteúdo esportivo. Isso ampliou a audiência global e criou oportunidades de receita para organizações esportivas.

Além dos benefícios evidentes, as redes sociais também apresentam desafios e preocupações. A disseminação de discursos de ódio, *cyberbullying* e o impacto negativo na saúde mental devido ao uso excessivo das redes sociais são questões críticas que exigem

atenção (Primack et al., 2017). Portanto, entender a dinâmica das redes sociais no contexto esportivo é essencial para promover um ambiente online saudável e construtivo (Hinduja & Patchin, 2018).

A evolução das redes sociais e sua influência nos esportes representam um campo de estudo interdisciplinar de grande relevância. Compreender como as redes sociais afetam a experiência esportiva, tanto positiva quanto negativamente, é fundamental para profissionais de marketing esportivo, acadêmicos e entusiastas dos esportes. O impacto das redes sociais nos esportes transcende a esfera digital e tem implicações sociais, culturais, econômicas e tecnológicas que merecem uma análise aprofundada.

### **1.1. Redes sociais e o futebol como esporte profissional**

As redes sociais tornaram-se uma parte integral do cenário do futebol profissional, transformando a maneira como os clubes, jogadores e torcedores interagem e se envolvem com o esporte mais popular do mundo (Boyle & Haynes, 2009). Um aspecto crucial é o impacto das redes sociais no futebol profissional. Anteriormente, a comunicação entre clubes e torcedores era principalmente unidirecional, mas atualmente, plataformas como Twitter, Instagram e Facebook possibilitam uma interação direta e instantânea (Thompson, 2020). Clubes utilizam essas plataformas para compartilhar notícias, atualizações de jogos e conteúdo exclusivo, enquanto os jogadores constroem suas marcas pessoais, promovem produtos e interagem com os fãs (Hutchins & Rowe, 2012).

A transformação da comunicação entre clubes de futebol e torcedores, impulsionada pelas redes sociais, tem aberto novas possibilidades para a análise de dados. Antes, a comunicação era predominantemente unidirecional, com clubes disseminando informações por meio de canais tradicionais como jornais, televisão e rádio (Smith, 2016). Hoje, plataformas como Twitter, Instagram e Facebook permitem uma interação direta e instantânea, tornando o diálogo bidirecional (Panjwani, 2021). Clubes utilizam essas redes para compartilhar notícias, atualizações de jogos e conteúdo exclusivo, enquanto os jogadores constroem suas marcas pessoais, promovem produtos e interagem com os fãs (Williams, 2013).

Diante desse cenário, surge a necessidade de ferramentas avançadas como *web scraping*, mineração de texto e análise de sentimentos, o qual permite a extração automática de grandes volumes de dados dessas plataformas, coletando informações valiosas sobre as interações e o conteúdo compartilhado (Russell, 2019). A mineração de texto, por sua vez,

processa e analisa esses dados textuais, extraindo padrões e informações relevantes que podem ser utilizados para diversas finalidades (Feldman & Sanger, 2007).

A análise de sentimentos é outra área chave de aplicação da mineração de texto no contexto do futebol. Ao monitorar as conversas nas redes sociais, é possível identificar tendências de opinião, avaliar a recepção de eventos específicos e entender o sentimento dos torcedores em relação a jogadores, treinadores e até mesmo patrocinadores (Liu, 2012). Esses insights podem ser fundamentais para a tomada de decisões estratégicas, como campanhas de marketing, gestão de crises e desenvolvimento de estratégias de comunicação (Pang & Lee, 2008). Além disso, permitem aos clubes e jogadores ajustarem suas ações e discursos para melhor atender às expectativas e demandas dos fãs, fortalecendo o engajamento e a lealdade da base de torcedores (Stieglitz & Dang-Xuan, 2013).

Além disso, a mineração de texto permite o monitoramento de notícias e tendências relacionadas ao futebol. Ao analisar o conteúdo gerado pelos usuários, é possível identificar notícias importantes, acompanhar o desempenho dos clubes em tempo real e antecipar possíveis crises de imagem (Alecsa et al., 2019). Isso é especialmente relevante em um ambiente onde a informação se espalha rapidamente e pode ter um impacto significativo na reputação de um clube ou jogador (Frederick et al., 2020).

O engajamento dos fãs e o marketing esportivo também se beneficiam da mineração de texto. Ao entender o que ressoa com os torcedores nas redes sociais, os clubes podem criar conteúdo mais relevante e envolvente, aumentar o alcance de suas campanhas e fortalecer o relacionamento com os fãs (Parganas et al., 2017). Além disso, a análise de dados pode ajudar na identificação de influenciadores digitais que possam promover a marca do clube ou de seus parceiros comerciais (Mander, 2019).

No contexto da avaliação de desempenho e *scouting*, a mineração de texto pode ser uma ferramenta valiosa para identificar talentos emergentes e avaliar o potencial de jogadores (Mackenzie & Cushion, 2013). Ao analisar estatísticas, comentários de especialistas e opiniões dos torcedores, os clubes podem tomar decisões mais informadas sobre contratações e desenvolvimento de jogadores (Sarmiento et al., 2018).

## **1.2.Web scrapping**

O *web scraping*, ou raspagem de dados da web, é uma técnica utilizada para extrair informações de sites da internet de forma automatizada (Mitchell, 2018). Essa técnica é especialmente útil quando se deseja coletar dados de maneira sistemática e estruturada a partir

de páginas da web que não oferecem uma API (Interface de Programação de Aplicativos) para acesso aos dados (Russell, 2019).

Existem diferentes abordagens para realizar *web scraping*, mas, em geral, envolve o uso de programas de computador (*bots*) para visitar páginas da web, analisar o seu conteúdo e extrair os dados relevantes. Esses dados podem ser textos, imagens, links, tabelas, entre outros tipos de conteúdo (Glez-Peña et al., 2014).

No contexto do presente estudo, o *web scraping* é utilizado para coletar dados relevantes relacionados ao futebol profissional em redes sociais como Twitter, Facebook e Instagram. Por exemplo, podemos extrair postagens de clubes, jogadores e torcedores, comentários em tempo real durante os jogos e análises de especialistas sobre eventos esportivos (Liu et al., 2020).

Apesar de ser uma técnica poderosa, o *web scraping* também apresenta desafios e considerações éticas. É importante garantir que a coleta de dados seja feita de forma ética e legal, respeitando os termos de serviço dos websites e a privacidade dos usuários (Krotov & Silva, 2018). Além disso, a qualidade e a precisão dos dados extraídos podem variar, exigindo cuidado na interpretação dos resultados (Hale, 2017).

### **1.3. Mineração de texto**

A mineração de texto, também conhecida como mineração de dados de texto, refere-se ao processo de extrair informações úteis e de alta qualidade de fontes de texto (Aggarwal & Zhai, 2012). Ela usa várias técnicas de análise, como processamento de linguagem natural (PLN), aprendizado de máquina, estatísticas e outros métodos para descobrir padrões e tendências em grandes conjuntos de dados de texto (Manning et al., 2008).

A mineração de texto pode ser usada para diversas aplicações, incluindo análise de sentimentos, extração de entidades nomeadas, sumarização automática de texto, tradução automática e detecção de tópicos (Cambria & White, 2014). É uma ferramenta poderosa para transformar dados não estruturados em informações estruturadas, facilitando a tomada de decisões informadas (Miner et al., 2012). Esse processo utiliza inteligência artificial e estatísticas para analisar grandes volumes de dados, descobrindo informações valiosas (Weiss et al., 2010). À medida que o *big data* e a tecnologia de armazenamento de dados evoluem, essa abordagem, também conhecida como descoberta de conhecimento em banco de dados (KDD), tem se tornado cada vez mais relevante (Fayyad et al., 1996).

#### 1.4. Justificativa e hipóteses

A realização deste estudo se justifica pela crescente importância das redes sociais como plataforma de expressão para os torcedores de futebol, refletindo suas emoções e opiniões em tempo real. Com a popularização do X e outras mídias sociais, compreender o sentimento dos torcedores se torna crucial não apenas para os clubes e atletas, mas também para analistas de mídia, publicitários e profissionais de marketing esportivo.

A análise de sentimento fornece insights valiosos sobre o impacto de eventos esportivos, permitindo uma resposta mais ágil e precisa às expectativas e percepções dos fãs (Cambria et al., 2017). Além disso, este trabalho contribui para o avanço das técnicas de análise preditiva aplicadas à análise de texto, um campo em crescimento que encontra aplicações em diversas áreas do conhecimento (Liu, 2012). Ao desenvolver um modelo capaz de classificar automaticamente os sentimentos expressos em publicações, a pesquisa demonstra a aplicabilidade e a eficácia dessas tecnologias na análise de grandes volumes de dados textuais (Pang & Lee, 2008), o que é essencial em um mundo cada vez mais orientado por dados (Russell, 2019).

Ademais, o estudo aborda lacunas na literatura existente sobre a reação emocional dos torcedores durante eventos esportivos específicos, além de trazer relevância prática ao oferecer subsídios para que stakeholders compreendam melhor o comportamento de seus torcedores nas redes sociais. Isso pode levar à implementação de estratégias de engajamento mais eficazes, melhoria na comunicação e fortalecimento da relação entre torcedores e clubes, promovendo uma experiência mais positiva e envolvente para os fãs do esporte.

Para alcançar esses objetivos práticos, este estudo se baseia em hipóteses que exploram a dinâmica dos sentimentos dos torcedores:

- **Distribuição de Sentimentos:** A maioria das publicações coletadas possui uma intenção negativa ou depreciativa.
- **Reação Pós-Jogo:** A maioria das publicações negativas são postadas logo após o término da partida, enquanto as publicações positivas são mais frequentes durante o jogo.
- **Em alta:** Após o término da partida, observa-se um crescimento acentuado na quantidade de publicações.

- **Comparação entre Pré-Jogo e Pós-Jogo:** As publicações coletadas antes do início da partida terão uma proporção maior de sentimentos positivos em comparação com os posts coletados após o final da partida.

### 1.5. Objetivo geral e objetivos específicos

Este trabalho visa desenvolver um modelo preditivo para análise de sentimentos em rede social ao segmento esportivo. O principal objetivo é classificar automaticamente os sentimentos expressos nas publicações como positivos ou negativos, permitindo uma compreensão mais abrangente da percepção pública em relação aos esportes. Os resultados esperados incluem a identificação da polaridade dos sentimentos associados a cada jogo, proporcionando insights valiosos sobre a opinião dos usuários nas redes sociais. Além disso, pretende-se avaliar a eficácia da análise em capturar a tendência geral dos sentimentos expressos nas publicações, contribuindo para uma análise mais abrangente do impacto emocional dos eventos esportivos nas plataformas de mídia social. O projeto visa atingir os seguintes objetivos específicos:

- **Identificar comportamento em redes sociais:** Analisar como os torcedores se expressam nas redes sociais em relação a eventos esportivos específicos, identificando tendências e padrões nos sentimentos (positivos ou negativos).
- **Avaliar a Precisão do Modelo preditivo:** Avaliar a eficácia e a precisão do modelo treinado para classificar as postagens como positivas ou negativas, contribuindo para o avanço das técnicas de análise de sentimento no contexto esportivo.
- **Compreender a Influência de Eventos Esportivos nos Sentimentos dos Torcedores:** Investigar como eventos específicos durante uma partida (gols, decisões de arbitragem, desempenho de jogadores etc.) influenciam os sentimentos expressos pelos torcedores nas redes sociais.
- **Explorar o Impacto das Redes Sociais no Comportamento dos Torcedores:** Explorar como as redes sociais moldam e refletem o comportamento e os sentimentos dos torcedores de futebol, contribuindo para a compreensão mais ampla da cultura esportiva digital.



## **2. REVISÃO DA LITERATURA**

Este capítulo trata da intersecção entre redes sociais e esportes constitui um campo de estudo multifacetado que tem ganhado relevância significativa nas últimas décadas. O advento e a proliferação das redes sociais transformaram a maneira como os esportes são consumidos, discutidos e comercializados. Este tópico de pesquisa se concentra na evolução e na influência das redes sociais no mundo dos esportes, abordando as implicações sociais, culturais, econômicas e tecnológicas dessa interação.

### **Interseção entre Redes Sociais e Esportes: Uma Perspectiva Multifacetada**

O campo de estudo que se situa na interseção entre redes sociais e esportes tem ganhado relevância significativa nas últimas décadas. O advento e a proliferação das redes sociais transformaram a maneira como os esportes são consumidos, discutidos e comercializados.

### **Evolução das Redes Sociais no Mundo dos Esportes**

As redes sociais revolucionaram a maneira como os esportes são consumidos e discutidos. Antes, os fãs dependiam de transmissões televisivas e rádio para acompanhar seus times favoritos. Hoje, as redes sociais permitem que os fãs se envolvam diretamente com os clubes e atletas, criando uma conexão mais profunda e pessoal.

Inspirado pelo trabalho de Fernanda de Alvarenga Miranda (2013) sobre marketing digital e futebol brasileiro, este estudo busca explorar ainda mais a relação entre redes sociais e esportes. Com este estímulo, nos aprofundamos mais nos dados e buscamos justificar algumas hipóteses que possuíamos em mente.

### **Implicações Sociais e Culturais**

Através do artigo, (Miranda, 2013) que conduziu o tema relacionado as redes sociais e o esporte, implicando diretamente na interação entre redes sociais e esportes, tendo resultado em implicações profundas na sociedade e na cultura. As redes sociais permitem que os fãs se envolvam diretamente com os clubes e atletas, criando uma conexão mais profunda e pessoal. Isso tem o potencial de fortalecer a identidade do clube e aumentar a lealdade dos fãs.

### **Impacto Econômico e Tecnológico**

As redes sociais têm um impacto econômico significativo no mundo dos esportes. Elas oferecem uma plataforma poderosa para a comercialização de esportes, permitindo que os clubes alcancem um público global e gerem receita através de publicidade e patrocínio. Além disso, as redes sociais fornecem uma rica fonte de dados que podem ser usados para melhorar a tomada de decisões e a estratégia de marketing dos clubes.

## **2.1 Mensagens de ódio nas redes sociais ligadas ao esporte**

A interação entre as redes sociais e o mundo dos esportes tem propiciado uma plataforma amplamente acessível para a expressão de opiniões, discussões e compartilhamento de informações sobre eventos esportivos. No entanto, essa interação também trouxe à tona uma preocupação crescente: a proliferação de mensagens de ódio nas redes sociais ligadas ao esporte.

O discurso de ódio compõe-se de dois elementos básicos: discriminação e externalidade. (Silva et al., 2011). Este tópico de pesquisa tem como objetivo explorar as diversas dimensões das mensagens de ódio nas redes sociais em relação ao esporte e analisar suas implicações.

As mensagens de ódio nas redes sociais ligadas ao esporte consistem em conteúdo online que é ofensivo, discriminatório, incitador de violência ou prejudicial a indivíduos ou grupos. A análise das mensagens de ódio nesse contexto requer uma abordagem multidimensional, abrangendo diversas áreas de pesquisa.

## **2.2. Natureza e tipologia das mensagens de ódio**

Uma análise aprofundada das mensagens de ódio nas redes sociais ligadas ao esporte deve começar pela identificação e categorização da natureza dessas mensagens. Isso envolve a classificação de diferentes tipos de discurso de ódio, como racismo, xenofobia, homofobia, misoginia e outras formas de discriminação. É essencial entender como o contexto esportivo e as rivalidades entre equipes ou grupos de fãs podem influenciar a ocorrência de mensagens de ódio. A rivalidade esportiva muitas vezes é usada como justificativa para o discurso prejudicial, e a análise deve investigar como essas rivalidades desencadeiam mensagens de ódio.

Para compreender completamente as mensagens de ódio nas redes sociais em relação ao esporte, é necessário examinar as motivações por trás dessas mensagens. As motivações

podem incluir preconceitos, rivalidades esportivas, desejo de chamar a atenção ou até mesmo a participação em grupos online com inclinações extremistas. A proliferação de mensagens de ódio nas redes sociais ligadas ao esporte apresenta implicações significativas, incluindo a criação de um ambiente online tóxico, o fomento de conflitos sociais e o impacto na saúde mental. Portanto, é crucial desenvolver estratégias de mitigação eficazes para lidar com esse problema. Estratégias podem incluir educação, moderação de conteúdo, monitoramento e a promoção de um ambiente online mais saudável.

A análise das mensagens de ódio nas redes sociais ligadas ao esporte requer uma abordagem multidimensional que abrange a natureza das mensagens, seu contexto esportivo, motivações e implicações. Compreender e enfrentar esse fenômeno é fundamental para garantir que as redes sociais continuem sendo um espaço positivo e construtivo para a comunidade esportiva e seus seguidores.

### 3. MATERIAIS E MÉTODOS

#### 3.1. Metodologia CRISP DM

O modelo CRISP-DM (Cross-Industry Standard Process for Data Mining) emerge como uma estrutura de referência consagrada para a condução de projetos de mineração de dados e análise de dados em diversos setores (Chapman et al., 2000). Sua abordagem sistemática e iterativa compreende seis fases interligadas: Compreensão do Negócio, Compreensão dos Dados, Preparação dos Dados, Modelagem, Avaliação e Implantação.

A adaptabilidade do CRISP-DM a diferentes domínios e requisitos de negócios (Wirth & Hipp, 2000) o torna particularmente apropriado para este projeto de análise de sentimentos no contexto do futebol. Ao considerar as nuances do ambiente esportivo e a dinâmica das interações nas redes sociais, o CRISP-DM oferece uma estrutura flexível para orientar todas as etapas do ciclo de vida do projeto.

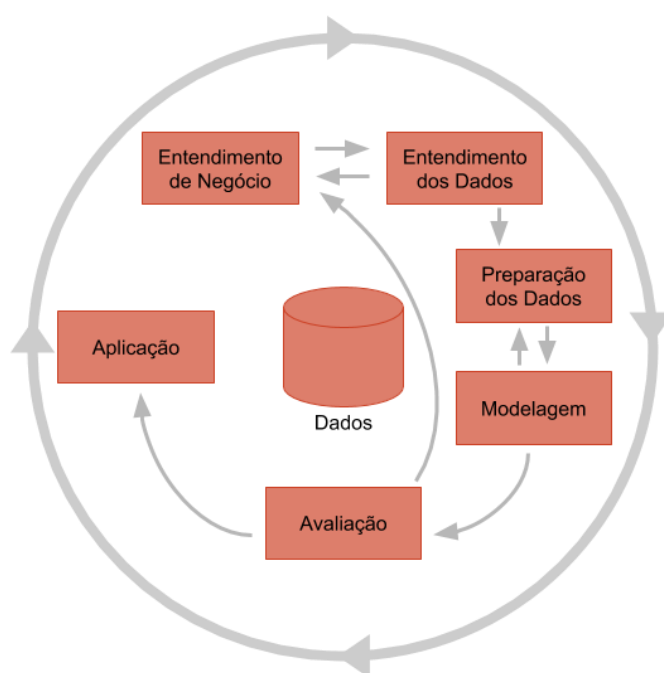


Figura 1: Ilustração do modelo CRISP-DM.

Uma das principais vantagens do modelo CRISP-DM é sua ênfase na compreensão do domínio do problema e na definição clara dos objetivos do projeto desde o início (Tan et al., 2016). Essa abordagem contribui para garantir que as análises realizadas estejam alinhadas com

as necessidades reais do negócio e que os resultados obtidos sejam diretamente aplicáveis na tomada de decisões estratégicas.

Além disso, a natureza iterativa do CRISP-DM, com sua ênfase na interação contínua entre as diferentes fases do processo, permite uma abordagem ágil e adaptável que pode responder eficazmente às mudanças no ambiente externo e aos novos insights descobertos durante o projeto (Hand et al., 2001).

Portanto, ao adotar o modelo CRISP-DM como metodologia para o projeto, é possível aproveitar sua estrutura robusta, foco na compreensão do domínio do problema e natureza iterativa para garantir uma abordagem sistemática e eficaz na obtenção de insights valiosos a partir dos dados disponíveis.

Inicialmente, focaremos no entendimento do contexto do negócio, priorizando a identificação de sentimentos positivos e negativos em publicações sobre eventos esportivos. Em seguida, faremos uma análise detalhada dos dados disponíveis, coletando uma base inicial de publicações já rotuladas como positivas ou negativas, como pode ser observado na figura a seguir.

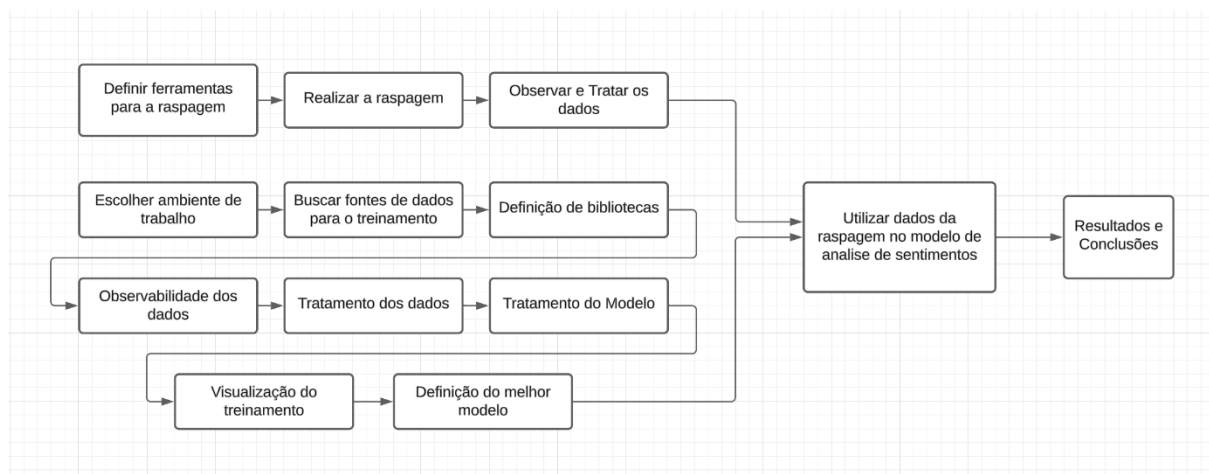


Figura 2: Ilustração dos passos do planejamento.

Após a fase de entendimento dos dados, entra a de preparação, onde serão realizados a limpeza e o processamento dos textos. Isso incluirá a remoção de palavras irrelevantes, a normalização de termos e outras técnicas para garantir a qualidade dos dados de entrada. Na etapa de modelagem, serão selecionados algoritmos de aprendizado de máquina adequados para a tarefa de classificação de sentimentos. O conjunto de dados será dividido em conjuntos de treino e teste para avaliar o desempenho do modelo.

A avaliação do modelo será feita utilizando métricas padrão, como acurácia, precisão e recall. Isso nos permitirá entender o quão bem o modelo está performando em identificar corretamente os sentimentos expressos nas publicações.

Por fim, após a fase de implementação, será usado o modelo treinado com novas publicações em redes sociais (novas entradas), automatizando a classificação de sentimentos. Este processo será continuamente monitorado e refinado para garantir resultados precisos e atualizados. Essa abordagem sistemática garantirá a qualidade e a eficácia da análise de sentimentos em larga escala.

## 4. DESENVOLVIMENTO DO PROJETO

### 4.1. Entendimento do negócio

A geração textual e a comunicação digital, especialmente na plataforma X, têm um impacto significativo na polarização dentro do setor futebolístico. Publicações se espalham rapidamente, amplificando emoções positivas ou negativas e influenciando as percepções dos torcedores em tempo real. Algoritmos das redes sociais criam bolhas de filtragem, mostrando conteúdo que reforça opiniões existentes e fortalecendo a polarização. Torcedores se agrupam em comunidades online com opiniões semelhantes, aumentando a resistência a pontos de vista divergentes.

Os jogadores e influenciadores utilizam a plataforma para promover suas narrativas, moldando percepções e potencialmente polarizando ainda mais os torcedores. Hashtags e campanhas organizadas mobilizam grupos em torno de certas narrativas, reforçando divisões. Tal polarização pode criar divisões significativas entre grupos de torcedores, levando a conflitos verbais online e até mesmo a confrontos físicos em eventos. Críticas intensas e apoio incondicional afetam a moral e desempenho de jogadores e clubes. Em suma, a polarização influencia decisões administrativas e de gestão nos clubes, incluindo contratações e demissões.

Entender como as opiniões e sentimentos são gerados e propagados no Twitter pode ajudar clubes e stakeholders a gerenciar melhor suas estratégias de comunicação e engajamento, minimizando os efeitos negativos da polarização e promovendo um ambiente mais inclusivo e positivo para todos os torcedores.

### 4.2. Entendimento dos dados

Para utilizar um algoritmo de análise de sentimentos eficaz, é essencial ter um conjunto de dados abrangente para treinamento e um conjunto de teste para avaliar a precisão do algoritmo. Por este motivo, utilizamos o dataset UMICH SI650 - *Sentiment Classification*, fornecido pela Universidade de Michigan para competições do Kaggle<sup>1</sup>, como pode ser visualizado da próxima figura.

---

<sup>1</sup> <https://www.kaggle.com/c/si650winter11>

Este conjunto de dados foi projetado para um curso de aprendizado de máquina e inclui exemplos de texto de redes sociais e outras fontes online, tornando-o ideal para tarefas de análise de sentimentos.

A base apresentada possui uma estrutura organizada em quatro colunas, cada uma desempenhando um papel específico na análise dos dados. A primeira coluna, denominada "ItemID", serve como identificador único para cada entrada de dado, garantindo a individualização de cada registro e facilitando sua referência ao longo do estudo.

A segunda coluna, intitulada "Sentiment", é crucial para a classificação dos textos analisados. Esta coluna contém valores binários, onde '1' indica que o texto correspondente possui um sentimento positivo, enquanto '0' denota um sentimento negativo. Esta categorização é essencial para a análise de sentimentos, permitindo uma avaliação clara e objetiva do conteúdo emocional dos textos.

Embora a terceira coluna, chamada "SentimentSource", registre a origem do sentimento atribuído, ela não será utilizada diretamente na análise presente. Esta coluna poderia, em outros contextos, fornecer informações adicionais sobre a procedência dos dados, potencialmente enriquecendo a compreensão das fontes de sentimentos.

Finalmente, a quarta coluna, "SentimentText", contém os textos que foram submetidos à análise de sentimentos. Estes textos são os elementos centrais da tabela, sendo o conteúdo sobre o qual se aplica a classificação positiva ou negativa descrita na coluna Sentiment. A estrutura organizada da tabela permite uma análise sistemática e detalhada dos dados textuais, fundamental para a investigação proposta.





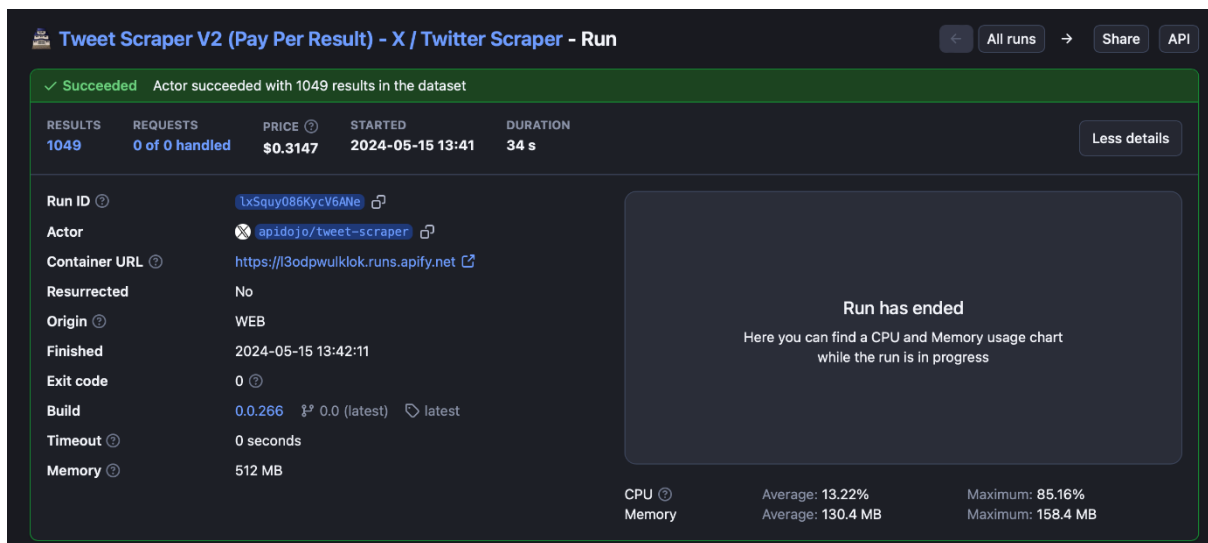


Figura 4: Resultado da execução do scrapping de dados

Além disso, a busca por palavras-chave ajuda a garantir que os dados coletados sejam relevantes e úteis para a análise posterior. Ao filtrar os resultados da raspagem com base em palavras-chave específicas, é possível obter informações mais pertinentes e significativas, contribuindo para uma análise mais precisa e informada. No projeto atual optamos por filtrar algumas palavras e hashtags que possuam relação com a partida de futebol escolhida. Entre elas #Realmadrid, #ManCity, #Champions, #Real Madrid, #Manchester City, #Vini Jr, #Rodrygo e #Foden.

1	#Realmadrid	×
2	#ManCity	×
3	#Champions	×
4	Real Madrid	×

Figura 5: As palavras-chave que serão buscadas na raspagem de dados.

5	Manchester City	X
6	Vini Jr	X
7	Rodrygo	X
8	Foden	X

Figura 6: Outras palavras-chave que serão buscadas na raspagem de dados.

Em seguida optamos por trazer a última versão (build), além de não inserir limite de tempo para que o número de resultados não seja impactado. Foi configurada a memória da CPU, e o máximo de resultados, ou seja, tweets da raspagem.

Run options

MAXIMUM RESULTS

Unlimited

BUILD

latest

TIMEOUT

0s

MEMORY

512 MB

Actor running slowly or crashing? Try increasing memory.

**Build**

latest

**Timeout** ?

0 seconds   ☒ No timeout ?

No timeout can result in infinite runs.

**Memory** ?

512 MB

More memory means more CPU horsepower.  
512 MB gives you 0.125 CPU cores.

**Maximum charged results** (optional) ?

1000 tweets   ☒ No maximum limit ?

Figura 7: Configurações de duração da raspagem no Apify

Em seguida, temos as opções de filtrar a data de publicação dos tweets, ou seja, será a data alvo da raspagem de dados. A partida ocorreu no dia 17/04/2024, portanto utilizamos a mesma data no filtro para atingir um melhor resultado.

Start date (optional) ?

2024-04-17 X

End date (optional) ?

2024-04-18 X

Figura 8: Opção dos dias em que os tweets foram realizados

Para obter uma maior acurácia, a plataforma fornece a possibilidade de filtrar os tweets por idioma, usuários verificados, pode-se filtrar apenas imagens, vídeos ou apenas citações. A escolha do idioma em inglês é indispensável, uma vez que o modelo foi treinado a partir de textos em inglês, isso vai garantir a eficiência na hora de estudar os tweets raspados.

Filter Tweets

You can filter tweets by using these options.

Tweet language (optional) ?

English X | v

☐ Only verified users ?

☐ Only Twitter Blue ?

☐ Only image ?

☐ Only video ?

☐ Only quote ?

Figura 9: Escolha de filtros para os tweets.

O resultado da raspagem de dados foi uma tabela estruturada em seis colunas, cada uma desempenhando um papel específico na análise dos dados coletados. Essa tabela retorna os tweets em inglês publicados referente a uma partida de futebol, especificamente relacionada ao jogo entre Real Madrid e Manchester City.

A primeira coluna, denominada "text", contém o conteúdo textual dos tweets. Esta coluna é fundamental para a análise qualitativa, pois registra exatamente o que está sendo dito nos tweets, permitindo a aplicação de técnicas de processamento de linguagem natural para classificar e interpretar os sentimentos expressos.

A segunda coluna, "viewCount", registra o número de visualizações de cada tweet. Esta métrica é importante para avaliar o alcance e a visibilidade dos tweets, ajudando a entender quantas pessoas foram expostas ao conteúdo, sejam ele positivo ou negativo. Já a terceira coluna, "retweetCount", indica o número de retweets que cada tweet recebeu. Essa informação é crucial para medir o engajamento dos usuários e a disseminação do conteúdo, revelando quais tweets tiveram maior repercussão na rede.

A quarta coluna, "quoteCount", contabiliza o número de vezes que o tweet foi citado por outros usuários. Esse dado fornece insights sobre o nível de interação e discussão gerada pelo tweet, além de destacar quais conteúdos motivaram respostas diretas e análises por parte da comunidade.

A quinta coluna, "replyCount", mostra o número de respostas diretas ao tweet. Esta métrica é relevante para entender o engajamento direto dos usuários com o tweet original, indicando a interação e o debate gerado. Finalmente, a sexta coluna, "likeCount", registra a quantidade de curtidas recebidas pelo tweet. Este dado é um indicador claro da popularidade e aceitação do conteúdo pelos usuários da plataforma.

Além dessas colunas, há também a coluna "createdAt", que registra a data e hora em que cada tweet foi publicado. Essa informação temporal é essencial para a análise cronológica dos dados, permitindo a identificação de padrões e tendências ao longo do tempo. A estrutura organizada dessa tabela permite uma análise detalhada e sistemática dos dados, fundamental para a investigação proposta.

text	viewCount	retweetCount	quoteCount	replyCount	likeCount	createdAt
Rodrygo: "Foden thinks I	6008	14	2	20	207	Wed Apr 17 23:45:45 +0000 2024
my club still employs phil	386	0	0	0	12	Wed Apr 17 23:13:44 +0000 2024
@MrDtAFC Since Pep Joii	41456	27	7	14	616	Wed Apr 17 23:59:00 +0000 2024
Foden vs Real Madrid / H	373	1	0	0	4	Wed Apr 17 23:44:59 +0000 2024
So we're like just not goir	596520	797	263	75	9009	Wed Apr 17 23:05:16 +0000 2024
wE lOsT bUt FoDeN sCOrE	5566	5	3	0	64	Wed Apr 17 23:42:00 +0000 2024
I posted this on IG expect	60	0	0	0	0	Wed Apr 17 23:10:20 +0000 2024
Your City side to face Rea	1842920	3480	2564	1001	20906	Wed Apr 17 17:45:16 +0000 2024
I don't think this was any	125	1	1	3	5	Wed Apr 17 22:57:41 +0000 2024
Foden stinker when the li	95	0	0	0	0	Wed Apr 17 23:59:33 +0000 2024
Real Madrid knocks out M	479	0	0	0	1	Wed Apr 17 22:49:18 +0000 2024
We started an actual lb a	85	0	0	0	3	Wed Apr 17 23:49:01 +0000 2024
Been hearing "Saka this"	247	0	1	0	3	Wed Apr 17 23:21:24 +0000 2024
Odegard who??? Foden v	91	0	0	0	0	Wed Apr 17 23:25:52 +0000 2024
Also need to figure out w	105	0	0	0	2	Wed Apr 17 23:15:08 +0000 2024
Not even tryna ride out fr	21758	9	5	14	257	Wed Apr 17 23:29:45 +0000 2024
thank fuck foden didn't n	1957	1	2	4	34	Wed Apr 17 23:07:41 +0000 2024
Thought the same. It felt	407	0	0	1	2	Wed Apr 17 23:55:26 +0000 2024

Figura 10: Retorno da raspagem de dados

### 4.3. Modelagem

Para realizar a análise de sentimentos, foi utilizada a linguagem Python, através da plataforma Google Colab. Esse ambiente fornece recursos poderosos e acesso direto aos dados armazenados no Google Drive. Com o suporte às principais bibliotecas, como pandas e NLTK, conseguimos explorar, pré-processar e modelar os dados de texto de maneira eficiente e colaborativa. Tal abordagem proporcionou um ambiente robusto e eficaz para conduzir a pesquisa de análise de sentimentos.

A montagem do Google Drive é crucial para integrar os dados armazenados no Google Drive ao ambiente de execução do Google Colab. Ao montar o Google Drive, estabelecemos uma conexão direta entre o notebook Colab e os arquivos no Drive, permitindo o acesso aos dados sem a necessidade de download ou upload manual. Isso é particularmente útil quando lidamos com conjuntos de dados grandes ou quando queremos salvar modelos treinados ou resultados de análises para acesso posterior.

Antes de começar a trabalhar no projeto, é necessário garantir que todas as bibliotecas e pacotes necessários estejam instalados. Nesta etapa, são instalados os pacotes adicionais que serão utilizados no projeto, como o `nltk` e o `pyspark`. Tais pacotes fornecem funcionalidades

específicas que serão úteis para tarefas como pré-processamento de texto (NLTK) e processamento distribuído de grandes conjuntos de dados (PySpark).

Com os pacotes instalados, as bibliotecas necessárias para análise e processamento de dados foram importadas, o que inclui bibliotecas como `numpy`, `pandas`, `matplotlib`, `seaborn`, `datetime` e outras. Essas bibliotecas fornecem uma ampla gama de funcionalidades para manipulação de dados, visualização, análise estatística, entre outras tarefas, tornando possível realizar análises detalhadas e criar modelos de aprendizado de máquina sofisticados.

```
import os, sys
import numpy as np
import pandas as pd
from pyspark.sql import SparkSession

import matplotlib.pyplot as plt
%matplotlib inline
import seaborn as sns
from datetime import datetime
from IPython.display import display, Markdown as md

from transformers import AutoModelForSequenceClassification, AutoTokenizer
import nltk

# Random, seed..
from numpy.random import normal, seed, uniform
np.random.seed(42)
```

Figura 11: Import's das bibliotecas utilizadas

O próximo passo é carregar os dados que serão analisados e processados. Neste caso, os dados são lidos de um arquivo Parquet armazenado no Google Drive que traz as informações do dataset UMICH SI650 - Sentiment Classification que posteriormente será utilizado para treinar o modelo. O formato Parquet é uma escolha comum para armazenar grandes conjuntos de dados, pois oferece eficiência de leitura e compactação de dados.

Após carregar os dados, é necessário realizar uma análise exploratória para entender melhor suas características e identificar possíveis padrões ou tendências. Isso inclui a verificação de informações básicas sobre o conjunto de dados, como o número de registros e variáveis, tipos de dados, valores ausentes, além de estatísticas descritivas e visualizações para entender a distribuição e correlação entre as variáveis. Nesta etapa, podemos identificar possíveis problemas nos dados, como outliers ou inconsistências, que podem exigir tratamento durante o pré-processamento.

Neste trecho do processo foram identificadas algumas informações importantes do conjunto de dados, apresentando informações básicas sobre as entradas e examinando a distribuição dos sentimentos. Descobriu-se que os valores únicos da coluna "Sentiment", o número de registros com sentimentos positivos e negativos, além de visualizar essa distribuição por meio de um gráfico. Por fim, exibe exemplos dos dois tipos de sentimentos, positivos e negativos, na coluna alvo Sentiment.

Examples with negative sentiments:

	ItemID	Sentiment	SentimentSource	SentimentText
0	1	0	Sentiment140	is so sad for my APL frie...
1	2	0	Sentiment140	I missed the New Moon trail...
3	4	0	Sentiment140	.. Omgaga. Im sooo im gunna CRy. I'...
4	5	0	Sentiment140	i think mi bf is cheating on me!!! ...
5	6	0	Sentiment140	or i just worry too much?
7	8	0	Sentiment140	Sunny Again Work Tomorrow :-I ...
10	11	0	Sentiment140	I must think about positive..
12	13	0	Sentiment140	this weekend has sucked so far
13	14	0	Sentiment140	jb isnt showing in australia any more!
14	15	0	Sentiment140	ok thats it you win.

Figura 12: Dados analisados e processados com "sentimentos negativos".



Examples with positive sentiments:

ItemID	Sentiment	SentimentSource	SentimentText
2	3	1	Sentiment140 omg its already 7:30 :O
6	7	1	Sentiment140 Juuuuuuuuuuuuuuuuuussssst Chillin!!
8	9	1	Sentiment140 handed in my uniform today . i miss you ...
9	10	1	Sentiment140 hmmm.... i wonder how she my number @-)
11	12	1	Sentiment140 thanks to all the haters up in my face a...
17	18	1	Sentiment140 Feeling strangely fine. Now I'm gonna go l...
22	23	1	Sentiment140 You're the only one who can see this cause...
28	29	1	Sentiment140 goodbye exams, HELLO ALCOHOL TONIGHT
38	39	1	Sentiment140 uploading pictures on friendster
41	42	1	Sentiment140 (: !!!!! - so i wrote something last week. ...

Figura 13: Dados analisados e processados com sentimentos positivos.

Ainda neste processo selecionamos aleatoriamente uma amostra de 100.000 entradas do DataFrame original (número de linhas utilizadas para o treinamento), excluindo as colunas 'ItemID' e 'SentimentSource' e renomeando as colunas 'Sentiment' para 'sentiment' e 'SentimentText' para 'sentimentText'. Essa amostra reduzida é armazenada no DataFrame ``df_amos``. Em seguida, o código calcula o número de registros com sentimentos positivos e negativos nesse DataFrame ``df_amos``, contando o número de ocorrências onde o valor da coluna 'sentiment' é igual a 1 (positivo) e 0 (negativo), respectivamente. Os resultados são então impressos, fornecendo informações sobre o número de registros com sentimentos positivos e negativos na amostra de dados selecionada.

```
df_amos = (df.sample(100000)
.....drop(columns=['ItemID', 'SentimentSource'])
.....rename(columns={'Sentiment': 'sentiment', 'SentimentText': 'sentimentText'}))

pos_count = df_amos[df_amos.sentiment==1].count()
print("\n Record count with positive sentiments:\n", pos_count)

neg_count = df_amos[df_amos.sentiment==0].count()
print("\n Record count with negative sentiments:\n", neg_count)
```

```
Record count with positive sentiments:
sentiment      50303
sentimentText   50303
dtype: int64
```

```
Record count with negative sentiments:
sentiment      49697
sentimentText   49697
dtype: int64
```

Figura 14: Quantitativos de tweets positivos e negativos.

Ter uma quantidade equilibrada de textos negativos e positivos para o treinamento de um modelo de análise de sentimentos é crucial, comparável à função de um para-raios em uma tempestade elétrica. Assim como um para-raios distribui igualmente a carga elétrica para prevenir descargas descontroladas, um conjunto de dados balanceado assegura que o modelo seja exposto a uma representação equitativa de ambas as classes de sentimento.

Isso promove uma aprendizagem mais precisa e imparcial, minimizando o risco de viés e permitindo que o modelo generalize melhor para novos dados. Em termos acadêmicos, a igualdade na quantidade de exemplos positivos e negativos no conjunto de treinamento é essencial para mitigar a distorção do modelo, garantindo uma avaliação justa e uma capacidade eficaz de classificação de sentimentos.

Após a análise exploratória, teve início a etapa de limpeza e pré-processamento dos dados. Esta etapa é crucial para garantir a qualidade e consistência dos dados antes da modelagem. Aqui, aplicamos uma série de técnicas para remover informações irrelevantes, corrigir erros e padronizar o formato dos dados. Isso inclui a remoção de URLs, menções a usuários e hashtags, caracteres especiais e números. Além disso, foi feita a tokenização dos textos, quebra dos textos em palavras individuais, remoção de *stopwords* (palavras comuns que não contribuem para o significado do texto) e lematização (transformação das palavras para sua forma base). O objetivo é preparar os dados de texto de maneira consistente e padronizada para a modelagem de aprendizado de máquina.

```

# Função para limpar o tweet
def clean_tweet_advanced(tweet):
    # Removendo URLs
    tweet = re.sub(r"http\S+|www\S+|https\S+", ' ', tweet, flags=re.MULTILINE)
    # Removendo menções a usuários e hashtags
    tweet = re.sub(r'\@w+|\#', ' ', tweet)
    # Remover hashtags, mas manter o texto
    tweet = re.sub(r'#', ' ', tweet)
    # Remover caracteres especiais e números
    tweet = re.sub(r'[^a-zA-Z\s]', ' ', tweet)

    # Tokenização
    word_tokens = word_tokenize(tweet)
    # Removendo a pontuação
    word_tokens = [word for word in word_tokens if word.isalnum()]
    # Removendo stop words e palavras com menos de 3 caracteres
    stop_words = set(stopwords.words('english'))
    filtered_tweet = [word for word in word_tokens if not word.lower() in stop_words and len(word) > 2]
    # Lemmatização
    lemmatizer = WordNetLemmatizer()
    lemmatized_tweet = [lemmatizer.lemmatize(word) for word in filtered_tweet]

    # Convertendo a lista de palavras lematizadas em uma string
    cleaned_tweet = ' '.join(lemmatized_tweet)

    # Retornar None se a string estiver vazia
    return cleaned_tweet if cleaned_tweet.strip() != '' else None

# Aplicando a limpeza aos dados
df_amos['sentimentText'] = df_amos['sentimentText'].apply(clean_tweet_advanced)
df_amos.dropna(subset=['sentimentText'], inplace=True)

```

Figura 15: Passos utilizados para o tratamento dos dados.

Com os dados pré-processados, dividimos o conjunto de dados em conjuntos de treinamento e teste. Isso nos permite avaliar o desempenho do modelo em dados não vistos e evitar *overfitting*, onde o modelo se ajusta muito bem aos dados de treinamento, mas não generaliza bem para novos dados. A divisão dos dados é realizada de forma aleatória, mantendo a distribuição das classes de sentimentos nos conjuntos de treinamento e teste.

```

# Dividindo os dados em conjunto de treinamento e teste
X_train, X_test, y_train, y_test = train_test_split(df_amos['sentimentText'], df_amos['sentiment'], test_size=0.3, random_state=42)

# Vetorização dos textos usando TfidfVectorizer
# Escolha o número máximo de features conforme necessário
tfidf_vectorizer = TfidfVectorizer(max_features=5000)

```

Figura 16: Etapas para divisão dos dados entre treinamento e teste.

Antes de treinar os modelos de classificação, precisamos converter os textos em representações numéricas que possam ser processadas pelos algoritmos de aprendizado de máquina. Para isso, utilizamos o **TfidfVectorizer**, que converte os textos em vetores de *features* com base na frequência dos termos e na inversa da frequência do documento. Isso permite que os algoritmos de *machine learning* trabalhem com os textos de forma eficiente, capturando as relações semânticas entre as palavras.

Com os dados vetorizados, treinamos vários modelos de classificação, incluindo Regressão Logística, Árvore de Decisão, Floresta Aleatória, SVM e KNN. Utilizamos a técnica de GridSearchCV para encontrar os melhores hiperparâmetros para cada modelo, otimizando seu desempenho. Isso envolve testar diferentes combinações de hiperparâmetros e selecionar aquelas que resultam no melhor desempenho de acordo com uma métrica de avaliação definida, como a acurácia.

A Regressão Logística é um modelo estatístico amplamente utilizado para análise de dados em que a variável-alvo é categórica. Baseia-se na estimação da probabilidade de ocorrência de um evento por meio de uma função logística, sendo comumente aplicado em problemas de classificação binária (Hosmer Jr, Lemeshow, & Sturdivant, 2013).

```
models = {  
    'logistic_regression': {  
        'model': LogisticRegression(),  
        'params': {  
            'clf__C': [1, 5, 10]  
        }  
    }
```

Figura 17: Modelo utilizado para lógica de regressão logística.

A Árvore de Decisão é um modelo de aprendizado supervisionado que representa uma estrutura hierárquica de decisões, dividindo o conjunto de dados em subconjuntos menores com base nos atributos mais relevantes para a classificação (Breiman, Friedman, Olshen, & Stone, 1984).

```
'decision_tree': {  
    'model': DecisionTreeClassifier(),  
    'params': {  
        'clf__max_depth': [5, 10, 20],  
        'clf__min_samples_leaf': [1, 2, 5]  
    }  
}
```

Figura 18: Modelo utilizado para lógica de árvore de decisão.

A Floresta Aleatória é um algoritmo de aprendizado de conjunto que combina múltiplas árvores de decisão para melhorar a precisão e evitar o sobre ajuste. Ele utiliza técnicas de

amostragem aleatória para construir várias árvores e, em seguida, combina suas previsões (Breiman, 2001).

```
'random_forest': {  
    'model': RandomForestClassifier(),  
    'params': {  
        'clf__n_estimators': [10, 50, 100],  
        'clf__max_depth': [5, 10, 20]  
    }  
}
```

Figura 19: Modelo utilizado para lógica de floresta aleatória.

A Máquina de Vetores de Suporte (SVM) é um modelo de aprendizado supervisionado que mapeia os dados em um espaço dimensional superior para encontrar um hiperplano de separação ótimo entre as classes. Ele é eficaz em conjuntos de dados de alta dimensionalidade e pode lidar com problemas de classificação e regressão (Cortes & Vapnik, 1995).

```
'SVM': {  
    'model': SVC(),  
    'params': {  
        'clf__C': [0.1, 1, 10],  
        'clf__kernel': ['rbf', 'linear']  
    }  
}
```

Figura 20: Modelo utilizado para lógica de SVM.

O K-Vizinhos Mais Próximos (KNN) é um algoritmo simples de aprendizado supervisionado que classifica um ponto de dados com base na maioria dos votos dos seus k vizinhos mais próximos no espaço de atributos. Ele é intuitivo e fácil de implementar, mas pode ser computacionalmente custoso em grandes conjuntos de dados (Cover & Hart, 1967).

```

},
'knn': {
    'model': KNeighborsClassifier(),
    'params': {
        'clf__n_neighbors': [3, 5, 7]
    }
}

```

Figura 21: Modelo utilizado para lógica de KNN.

Após o treinamento, avaliamos o desempenho de cada modelo utilizando métricas de avaliação, como acurácia, precisão, recall e F1-score. Também buscamos exibir uma matriz de confusão, uma ferramenta crucial na avaliação do desempenho de modelos de classificação. Ele calcula a matriz de confusão com base nas previsões do modelo e nos valores verdadeiros do conjunto de teste, utilizando a biblioteca *scikit-learn*. Em seguida, o gráfico de mapa de calor é plotado com *seaborn*, mostrando os valores numéricos dentro das células. Isso permite uma análise visual do desempenho do modelo na classificação de sentimentos, com as classes 'Negativo' e 'Positivo' representadas nos eixos x e y.

Para encontrar o melhor modelo para classificação de sentimentos, foi implementada uma função denominada `find_best_model_with_grid_search`. Esta função utiliza um método conhecido como busca em grade (grid search), que testa múltiplos modelos e combinações de hiper parâmetros para identificar aquele que apresenta o melhor desempenho.

Para cada modelo, é construído um pipeline que inclui um transformador de vetorização TF-IDF para converter os textos em representações numéricas e o classificador correspondente. Em seguida, é realizado o treinamento de cada modelo utilizando a função `GridSearchCV`, que busca as melhores combinações de hiper parâmetros através de validação cruzada com 5 folds. Essa abordagem sistemática permite a identificação do modelo mais adequado para a tarefa de classificação de sentimentos em publicações de redes sociais, garantindo resultados mais precisos e confiáveis. A técnica de busca em grade é usada para encontrar o melhor modelo de classificação de sentimentos em publicações de redes sociais. Diferentes modelos são testados com várias combinações de hiper parâmetros, e o modelo com melhor desempenho é selecionado. Após essa análise, o código imprime o nome do melhor modelo, sua acurácia e os hiper parâmetros otimizados.

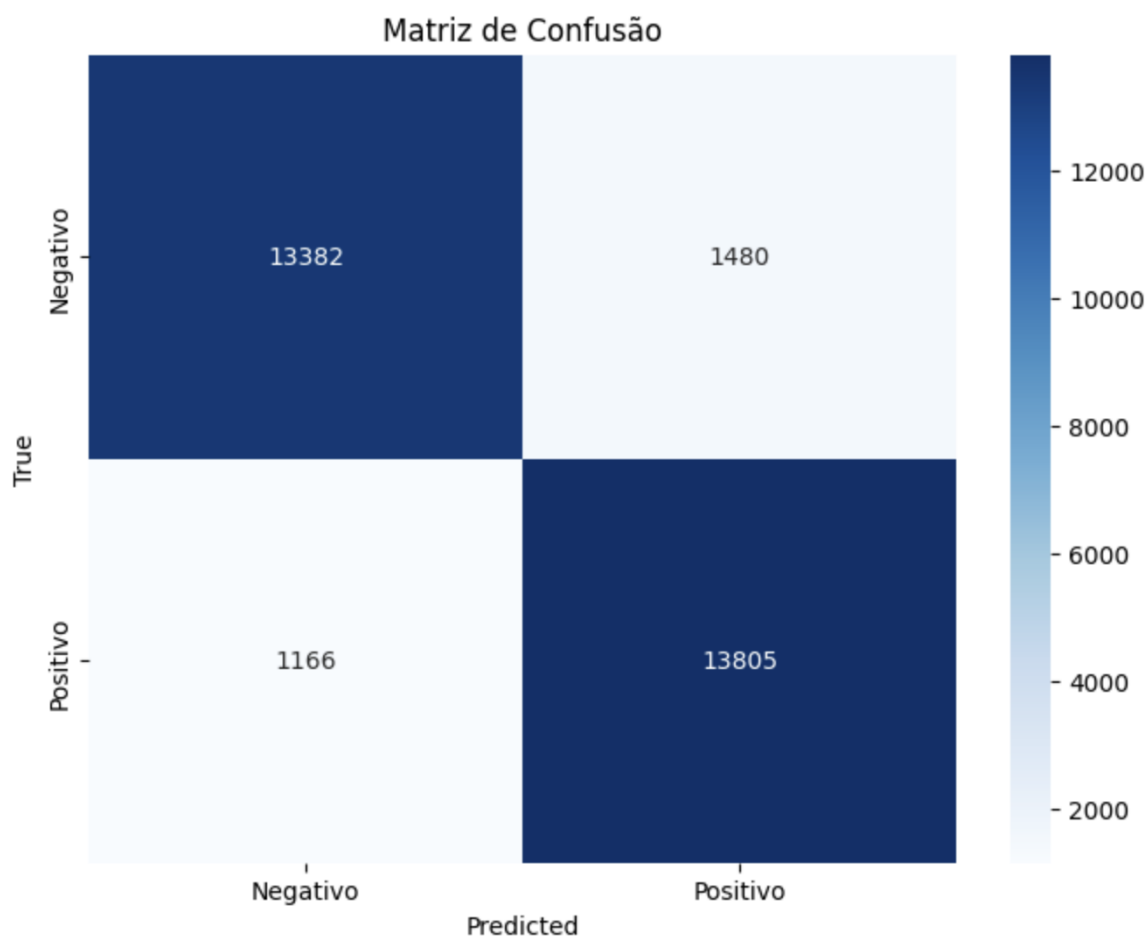


Figura 22: Gráfico que representa uma matriz de confusão e permite uma análise visual do desempenho do modelo.

#### 4.4. Avaliação do Modelo

Seguindo todas as métricas e técnicas de avaliação, o modelo que obteve o melhor desempenho foi o SVM, uma técnica de aprendizado supervisionado que mapeia os dados em um espaço dimensional superior para encontrar um hiperplano de separação ótimo entre as classes.

Sua abordagem é especialmente valiosa em cenários nos quais a precisão e a capacidade de generalização são fundamentais. A avaliação do desempenho do modelo de análise de sentimentos revelou uma acurácia de 91%, indicando uma taxa significativa de precisão na classificação dos sentimentos dos tweets, o relatório de classificação apresentou métricas

detalhadas para cada classe, demonstrando uma alta precisão e recall para ambas as classes de sentimento.

Acurácia: 0.91

Relatório de Classificação:

	precision	recall	f1-score	support
0	0.92	0.90	0.91	14862
1	0.90	0.92	0.91	14971
accuracy			0.91	29833
macro avg	0.91	0.91	0.91	29833
weighted avg	0.91	0.91	0.91	29833

Figura 23: Avaliação de desempenho do modelo treinado.

Para a classe 'Negativo', a precisão foi de 92% e o recall foi de 90%, enquanto para a classe 'Positivo', a precisão foi de 90% e o recall foi de 92%. Esses resultados indicam uma capacidade consistente do modelo em identificar corretamente os tweets com diferentes polaridades de sentimentos.

A análise adicional das métricas macros *avg* e *weighted avg* também confirmou a robustez geral do modelo, com uma média ponderada de precisão, recall e pontuação F1 de 91%. Esses achados corroboram a eficácia do modelo SVM na classificação precisa de sentimentos em textos, destacando sua utilidade potencial em diversas aplicações práticas, desde análise de mídias sociais até tomada de decisões empresariais baseadas em dados.

#### 4.5. Implementação

Com o modelo selecionado, os passos estão concluídos para salvar o modelo visando utilizá-lo posteriormente e fazer previsões em novos dados. Para isso, utilizamos a tabela resultante do scraping, a qual originalmente contém sete colunas: "text", "viewCount", "retweetCount", "quoteCount", "replyCount", "likeCount" e "createdAt". No entanto, para treinar o modelo de análise de sentimentos, mantemos apenas a coluna "text", que contém o conteúdo textual dos tweets, essencial para a modelagem de processamento de linguagem



natural. As demais colunas são excluídas, pois não são necessárias para o treinamento do modelo e podem introduzir ruído nos dados.

```
df_externo = pd.read_excel("/content/drive/MyDrive/TCC/Bases/dataset_tweet-scraper_Treinamento.xlsx")
df_externo.head(2)
```

	text
0	Rodrygo: "Foden thinks he's the British Mess...
1	my club still employs phil foden https://t.co/...

Figura 24: Conteúdo textual dos tweets, obtidos na raspagem.

Aplicou-se as etapas de limpeza e pré-processamento aos novos dados, assegurando a consistência com os dados de treinamento. Em seguida, utilizando o modelo treinado, classificamos os novos textos em sentimentos positivos ou negativos. Essa abordagem nos permite obter insights valiosos sobre a opinião pública em tempo real e embasar decisões informadas com base nesses insights.

```
# Lista para armazenar os resultados
resultados = []

# Usando o pipeline para fazer previsões nos novos tweets
novas_previsoes = pipeline.predict(tweets)

# Imprimindo e salvando os resultados
for tweet, predicacao in zip(tweets, novas_previsoes):
    resultados.append([tweet, 'Positivo' if predicacao == 1 else 'Negativo'])
    print(f"Tweet: {tweet}\nPrevisão: {'Positivo' if predicacao == 1 else 'Negativo'}\n")
```

Tweet: Rodrygo Foden thinks hes the British Messi laughs  
Previsão: Positivo

Tweet: my club still employs phil foden  
Previsão: Negativo

Tweet: Since Pep Joined  
Previsão: Positivo

Tweet: Foden vs Real Madrid Highlights Le diamant de city  
Previsão: Positivo

Tweet: So were like just not going to speak on that Fodens performance Nah Cool  
Previsão: Positivo

Figura 25: Tweets sendo armazenados e classificados em positivos e negativos.

Por fim, foram exportados os resultados das previsões para um formato conveniente, como um arquivo Excel ou CSV. Isso nos permite compartilhar os resultados com outras pessoas ou integrá-los em outros sistemas ou ferramentas para análises adicionais. Os resultados exportados podem ser usados para tomar decisões estratégicas, monitorar tendências de sentimentos ao longo do tempo ou alimentar modelos de machine learning para aplicações futuras.

Ao término do capítulo de desenvolvimento, é fundamental proporcionar aos leitores acesso ao material prático desenvolvido no contexto do estudo. Esta prática é considerada uma extensão natural do trabalho acadêmico, pois permite uma compreensão mais aprofundada da implementação e uma possível replicação do estudo por outros pesquisadores interessados. Para facilitar esse acesso, o código-fonte do projeto foi disponibilizado em um repositório no GitHub.

Para explorar o código e os recursos associados ao projeto discutido neste capítulo, é possível acessar o seguinte repositório: [TCC\_AnalisedeSentimentos<sup>2</sup>]. Neste repositório, os interessados encontrarão todos os artefatos relevantes, incluindo o código-fonte, os conjuntos de dados empregados na análise, e eventuais documentações pertinentes ao estudo em questão. Esta disponibilização oferece uma oportunidade para uma análise detalhada da implementação, contribuindo assim para uma compreensão mais ampla e crítica dos resultados obtidos.

---

<sup>2</sup> [https://github.com/luccanegrini/TCC\\_AnalisedeSentimentos](https://github.com/luccanegrini/TCC_AnalisedeSentimentos)

# 5. RESULTADOS OBTIDOS

## 5.1. Análise de Sentimentos

Após a execução completa do processo de análise de sentimentos, os tweets coletados foram classificados em categorias de positivos e negativos. A seguir, apresentam-se exemplos de tweets que foram processados pelo modelo de análise, demonstrando a eficácia do método empregado. Esses exemplos ilustram claramente os resultados obtidos, fornecendo evidências concretas da aplicação prática do modelo.

TWEETS POSITIVOS	
Tweet	Previsão
Haaland and Big Games What a story	Positivo
Jude showed foden levels today fantastic	Positivo
Real Madrid for a reason	Positivo
Very biased minds everywhere Vini Jr is the most deserved player right now to win Madrid players will occupy amp positions	Positivo
Vini Jr to Jude Bellingham on IG I told you this is Real Madrid	Positivo
We are proud of you guys This win means a lot to us	Positivo

Figura 26: tweets positivos processados pelo modelo de análise.

Os *posts* classificados como positivos pelo modelo refletem reações otimistas e favoráveis dos torcedores em relação ao jogo, expressando satisfação, entusiasmo e apoio às equipes participantes. Estes exemplos evidenciam o sucesso da classificação realizada pelo modelo, ao capturar de maneira precisa o conteúdo emocionalmente positivo dos tweets. Esses resultados destacam a capacidade do modelo em discernir entre sentimentos positivos e negativos, fornecendo uma visão abrangente das reações dos usuários nas redes sociais durante o evento esportivo.

TWEETS NEGATIVOS	
Tweet	Previsão
Wdym haaland did nothing ok kdb i can understand but haaland naaah and we also had vini jr subbed off for the last minutes	Negativo
Guess who is crying now	Negativo
Who is crying now	Negativo
You played shit today too You better sit up man	Negativo
Cant believe man were arguing saka is better than foden I guess not everyone has the vision	Negativo
De bruyne gets blamed but Foden doesnt	Negativo

Figura 27: tweets negativos processados pelo modelo de análise.

Da mesma forma, os tweets classificados como negativos mostram reações de frustração e insatisfação dos torcedores. Estes exemplos confirmam a capacidade do modelo de detectar sentimentos adversos. A apresentação desses resultados destaca a precisão e a eficácia do modelo de análise de sentimentos implementado, demonstrando como os dados coletados foram processados e classificados com sucesso.

### 5.2. Insights

Com base nos resultados da análise de sentimentos dos tweets coletados sobre o jogo entre Manchester City e Real Madrid, foi possível tirar conclusões significativas sobre o comportamento dos torcedores e a dinâmica das publicações nas redes sociais durante eventos esportivos. Estas conclusões não só nos permitem entender melhor o fenômeno estudado, como também relacioná-los com as hipóteses formuladas no início do trabalho.



Figura 28: Resultado da análise de dados realizada no Power BI.

### 5.3. Publicações Durante o Jogo

Uma parte significativa, 89,04% dos tweets coletados foram publicados após o início do jogo. Este dado evidencia a crescente interação dos usuários durante eventos ao vivo. Torcedores tendem a comentar mais durante e após o jogo, refletindo um comportamento ativo e engajado nas redes sociais. Relacionando este resultado com as hipóteses, observamos uma alta atividade após o início da partida, como era previsto na hipótese de "Em alta".

#### DIVISÃO DE RESULTADOS POR SENTIMENTO POSITIVO E NEGATIVO

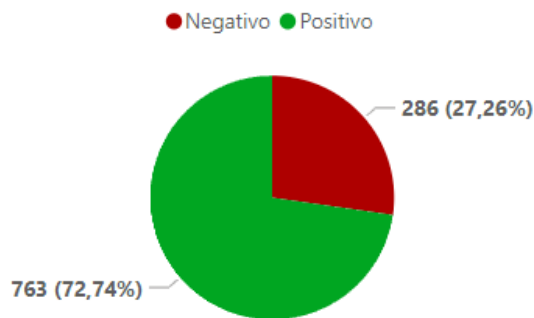


Figura 29: Divisão de resultados entre sentimentos positivos e negativos.

De acordo com o modelo treinado, 72,74% dos tweets analisados apresentaram um tom positivo. Este achado contraria a hipótese inicial de que a maioria das publicações seriam negativas ou depreciativas. O elevado percentual de sentimentos positivos indica uma perspectiva mais otimista ou entusiasta dos torcedores em relação ao jogo. Este fenômeno pode ser atribuído a diversos fatores, tais como o desempenho das equipes ou eventos específicos ocorridos durante a partida, que geraram reações favoráveis.

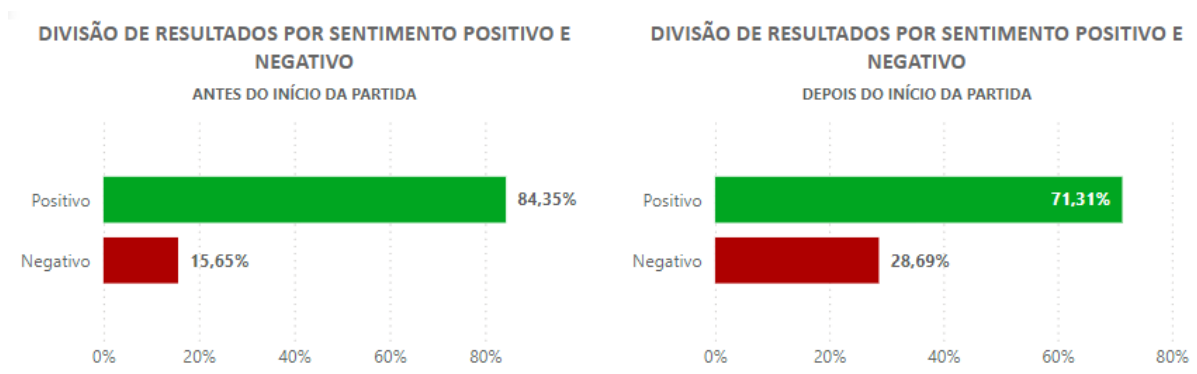


Figura 30: Distribuição dos resultados por positivo e negativo e por horário (antes e depois do início da partida).

Antes do início da partida, 84,35% dos tweets apresentavam sentimentos positivos, enquanto após o início do jogo essa proporção foi reduzida para 71,31%. Esta observação corrobora a hipótese de "Comparação entre Pré-Jogo e Pós-Jogo". Antes do jogo, há uma expectativa positiva e esperança de um bom desempenho, refletindo-se em um maior percentual de tweets positivos. Durante o jogo, especialmente diante de eventos adversos, a quantidade de tweets negativos tende a aumentar. A queda significativa na proporção de tweets positivos após

o início da partida evidencia que o desenrolar do jogo exerce uma influência direta sobre o sentimento dos torcedores.



Figura 31: Histórico de quantidade de tweets por hora (começando das 20:00).

A quantidade de tweets aumentou conforme o dia avançou, com um crescimento significativo entre 22:00 e 00:00 (horário máximo de coleta). Este dado corrobora parcialmente a hipótese de "Em Alta". Embora o maior número de publicações não tenha ocorrido especificamente após o término da partida, houve um aumento considerável durante as horas finais do jogo, indicando que os torcedores estavam altamente ativos nas redes sociais nesse período. Este comportamento sugere que momentos críticos ou emocionantes do jogo, geralmente ocorrendo nas últimas horas, incentivam mais postagens e interações.



Figura 32: Histórico de quantidade de tweets por hora (começando das 22:00).

Os resultados desta análise proporcionam uma compreensão significativa do engajamento dos torcedores nas plataformas de mídia social durante eventos esportivos. Eles indicam que a atividade nas redes sociais aumenta consideravelmente durante eventos ao vivo, atingindo um pico notável nos momentos cruciais da partida. Além disso, os sentimentos expressos pelos torcedores tendem a ser mais favoráveis e positivos antes do início do jogo, refletindo um clima de expectativa e otimismo. No entanto, é observado que eventos adversos durante a partida podem mitigar essa positividade. Apesar das expectativas iniciais de

predominância de sentimentos negativos, a análise revelou uma inclinação geral para reações positivas, sugerindo uma base de torcedores mais esperançosa ou entusiástica do que o previamente suposto.

5.4. Exploração dos Tweets mais interativos

A análise dos tweets mais interativos oferece uma visão abrangente da dinâmica da interação dos usuários nas redes sociais durante o evento esportivo em questão. Os tweets que receberam um maior número de likes, comentários e retweets são indicadores-chave do envolvimento e interesse dos usuários na plataforma (Cha, Cho, & Myers, 2010). Essas métricas de engajamento são amplamente utilizadas para avaliar a popularidade e o impacto de conteúdos nas redes sociais (Bakshy, Hofman, Mason, & Watts, 2011).

TOP 10 TWEETS POR QUANTIDADE DE RETWEETS		
Tweet	Retweet	Previsão
Against the odds Were Real Madrid	39242	Positivo
THIS IS REAL MADRID We always fight we always believe we never give up HALAAAA MADRID UCL Hustle AlwaysBelieve	19736	Positivo
Real Madrid saved football	14986	Positivo
ESPECIALLY FOR THE NEWCOMERS THIS IS REAL MADRID WE NEVER GIVE UP	13306	Positivo
Real Madrid qualified to Champions League semi finalsManchester City eliminated	12912	Positivo
Arsenal FC years without a title years without Carabao Cup UCL Europas Super cups Club World Cup Never defended a prem title The fanbase has the audacity of Real Madrid but in reality theyre closer to Blackburn Just stat padded fa cups	8427	Negativo
Antonio Rdiger Kept Haaland scoreless over two legs and converted the penalty to send Real Madrid into the semifinals	6180	Positivo
OFFICIAL Arsenal are out of the Clubs World Cup after tonight RB Salzburg access to the World Cup Manchester City Chelsea Real Madrid Atltico Madrid Bayern Borussia Dortmund Paris SaintGermain Inter Juventus Benfica Porto Salzburg	5924	Positivo
After Speed celebrated Rodrygos goal in the Man City vs Real Madrid game Man City fans got mad at him and told him to end his stream	4154	Positivo
Your City side to face Real Madrid XI Ederson Walker C Akanji Dias Gvardiol Rodrigo De Bruyne Bernardo Foden Grealish HaalandSUBS Ortega Moreno Carson Stones Ake Kovacic Doku Alvarez Gomez Nunes Bobb Lewis ManCity UCL	3480	Positivo

Figura 33: Tweets que receberam mais interação de retweets (compartilhamento).

Os likes representam a quantidade de vezes que um tweet foi marcado como favorito pelos usuários, indicando sua aceitação e apreciação (Java, Song, Finin, & Tseng, 2007). Os comentários refletem o nível de envolvimento e interação dos usuários com o conteúdo, fornecendo insights valiosos sobre as discussões e debates gerados pelo tweet (Danescu-Niculescu-Mizil, Gamon, & Dumais, 2011). Por fim, os retweets representam a quantidade de vezes que um tweet foi compartilhado por outros usuários, ampliando seu alcance e influência na plataforma (Kwak, Lee, Park, & Moon, 2010).

TOP 10 TWEETS POR QUANTIDADE DE LIKES		
Tweet	Previsão	Likes
Against the odds Were Real Madrid	Positivo	329563
Real Madrid qualified to Champions League semi finalsManchester City eliminated	Positivo	243254
THIS IS REAL MADRID We always fight we always believe we never give up HALAAAA MADRID UCL Hustle AlwaysBelieve	Positivo	184591
Real Madrid saved football	Positivo	171817
After Speed celebrated Rodrygos goal in the Man City vs Real Madrid game Man City fans got mad at him and told him to end his stream	Positivo	132192
Fede Valverde They obviously played better football than us but were Real Madrid	Positivo	124317
OFFICIAL Arsenal are out of the Clubs World Cup after tonight RB Salzburg access to the World Cup Manchester City Chelsea Real Madrid Atltico Madrid Bayern Borussia Dortmund Paris SaintGermain Inter Juventus Benfica Porto Salzburg	Positivo	114384
Antonio Rdiger Kept Haaland scoreless over two legs and converted the penalty to send Real Madrid into the semifinals	Positivo	111846
Jude Bellingham shows Real Madrid logo after scoring the penalty	Positivo	91976
Rodrygo against Man City	Positivo	87576
<b>Total</b>		<b>1591516</b>

Figura 34: Top 10 tweets que receberam mais likes.

Ao examinar esses tweets mais interativos, podemos identificar os temas mais relevantes, as opiniões mais influentes e as tendências mais significativas que emergiram durante o evento (Starbird & Palen, 2012).

TOP 10 TWEETS POR QUANTIDADE DE COMENTÁRIOS		
Tweet	Previsão	Comentários
Tonight showed Man City are the best team in Europe No team has ever dominated Real Madrid like that Proud of this team	Positivo	4298
Your City side to face Real Madrid XI Ederson Walker C Akanji Dias Gvardiol Rodrigo De Bruyne Bernardo Foden Grealish HaalandSUBS Ortega Moreno	Positivo	4004
Carson Stones Ake Kovacic Doku Alvarez Gomez Nunes Bobb Lewis ManCity UCL		
Against the odds Were Real Madrid	Positivo	3688
Real Madrid qualified to Champions League semi finalsManchester City eliminated	Positivo	3618
OFFICIAL Arsenal are out of the Clubs World Cup after tonight RB Salzburg access to the World Cup Manchester City Chelsea Real Madrid Atltico Madrid Bayern Borussia Dortmund Paris SaintGermain Inter Juventus Benfica Porto Salzburg	Positivo	3060
THIS IS REAL MADRID We always fight we always believe we never give up HALAAAA MADRID UCL Hustle AlwaysBelieve	Positivo	2782
Rudiger scores Our UCL journey comes to an end ManCity UCL	Negativo	2611
Real Madrid have parked the bus for minutes Embarrassing	Positivo	2553
Man City vs Real Madrid Bayern vs Arsenal Whoever predicts the score line for both games correct wins	Positivo	2331
This guy ghosted the entire match but its his image going viral This guys PR is insane wish they could do same for rodrygo cuz hes better	Negativo	1456
<b>Total</b>		<b>30401</b>

Figura 35: Top 10 tweets que receberam mais comentários.

## 5.5. Avaliação do modelo com novas entradas de dados

Além da validação cruzada e avaliação do modelo utilizando conjuntos de teste, foi conduzido um teste adicional para aferir a acurácia do modelo de classificação de texto. Este teste envolveu a geração de 99 publicações em texto "fakes" para serem submetidas ao modelo treinado. O objetivo foi avaliar o desempenho do modelo na classificação desses textos previamente não vistos.



Para este teste, foram criadas 99 publicações em texto simulando conteúdo realista. Esses textos abordavam uma variedade de tópicos e foram elaborados para se assemelhar ao estilo e ao conteúdo das publicações originais.

```
novos_tweets_comuns = [
    "Absolutely thrilled with my purchase! High quality and quick shipping.",
    "Fantastic customer support! They solved my issue in minutes.",
    "Just had the best meal ever at your restaurant, everything was perfect! 🍽️",
    "I'm in love with these shoes! So comfortable and stylish! 👟❤️",
    "Thanks for the incredible experience! I'll cherish these memories forever.",
    "Outstanding performance, exceeded all my expectations!",
    "The concert was phenomenal, an unforgettable night!",
    "Amazing product, I couldn't be happier with it.",
    "The hotel stay was perfect, I felt so pampered.",
    "Best coffee I've ever had, I'll be back for sure!",
    "Loved the fast service and friendly staff.",
    "Your app is so user-friendly and efficient, great job!",
    "The movie was a masterpiece, highly recommend it.",
    "Exceptional quality, worth every penny.",
    "Thank you for the prompt delivery, it arrived just in time.",
    "Absolutely beautiful location, had a wonderful time.",
    "The team was professional and very helpful.",
    "My new favorite place to shop, fantastic experience.",
    "These gadgets are lifesavers, so innovative!",
    "The food was delicious, compliments to the chef.",
    "Really impressed with the attention to detail.",
    "This book is a page-turner, couldn't put it down!",
```

Figura 36: Publicações criadas para teste.

Os 99 textos foram então submetidos ao modelo de classificação treinado previamente. Após a classificação automática, cada texto foi revisado manualmente para determinar se a classificação do modelo estava correta ou não. As classificações corretas foram contabilizadas como acertos, enquanto as incorretas foram consideradas erros.

```
# Usando o pipeline para fazer previsões nos novos tweets
novas_previsoes = pipeline.predict(novos_tweets_comuns)

# Imprimindo as previsões
for tweet, predicacao in zip(novos_tweets_comuns, novas_previsoes):
    print(f"Tweet: {tweet}\nPrevisão: {'Positivo' if predicacao == 1 else 'Negativo'}\n")

# Criar DataFrame
df = pd.DataFrame({
    "Tweet": novos_tweets_comuns,
    "Previsão": ["Positivo" if pred == 1 else "Negativo" for pred in novas_previsoes]
})

# Salvar em um arquivo Excel
df.to_excel("/content/drive/MyDrive/TCC/novas_previsoes.xlsx", index=False)

print("Arquivo Excel criado com sucesso!")
```

Figura 37: Passos utilizados para chegar até o resultado final.

Dos 99 textos submetidos ao modelo de classificação, o modelo classificou corretamente 86 textos e incorretamente 13 textos. Para calcular a acurácia do modelo neste teste adicional, aplicamos a seguinte fórmula:

$$\text{Acurácia} = (\text{Acertos} / \text{Total}) \times 100\%$$

**Substituindo os valores:**

$$\text{Acurácia} = (86 / 99) \times 100\% \approx 86.87\%$$

Portanto, a acurácia do modelo de classificação de texto neste teste adicional foi aproximadamente 86.87%. Esses resultados corroboram com a eficácia geral do modelo, demonstrando sua capacidade de generalização para textos previamente não vistos. Para explorar os recursos associados ao teste adicional discutido neste capítulo, é possível acessar o repositório no GitHub<sup>3</sup>.

---

<sup>3</sup> [github.com/luccanegrini/TCC\\_AnaliseDeSentimentos](https://github.com/luccanegrini/TCC_AnaliseDeSentimentos).

## **6. Considerações Finais**

### **6.1. Limitações do projeto**

A recente transição da API do Twitter para um modelo tarifado impõe a necessidade de pagamento por parte dos desenvolvedores para acessar determinadas funcionalidades e dados previamente disponíveis de forma gratuita. Este ajuste teve repercussões significativas no escopo de projetos dependentes da API do Twitter, demandando modificações substanciais tanto em termos financeiros quanto na reestruturação do código.

A mudança implica que, sob o novo nível gratuito, as contas têm a capacidade de postar até 1.500 tweets por mês, mas estão privadas de outros recursos e métodos que possibilitam uma extração mais flexível de tweets. Este nível é direcionado a bots e propósitos de teste. No entanto, o novo nível "básico", que implica um custo mensal de US\$100, expande as capacidades ao permitir que os desenvolvedores postem até 3.000 tweets por mês no nível do usuário e até 50.000 por mês no nível do aplicativo. Adicionalmente, oferece acesso ao V2, que incorpora diversos métodos adicionais.

É importante observar que, embora haja essa expansão em funcionalidades, o limite de leitura foi substancialmente reduzido para 10.000 tweets por mês, em comparação com a capacidade anteriormente disponível. Essas mudanças ressaltam a necessidade de uma abordagem ágil e adaptativa diante das alterações nas políticas de plataformas externas, sublinhando a importância da constante adaptação no desenvolvimento de projetos que dependem de APIs externas.

### **6.2. Futuras aplicações**

Para futuras investigações, seria interessante ampliar o período de coleta de dados para incluir o tempo após o término da partida, a fim de confirmar ou refutar a hipótese de que a maioria das publicações negativas ocorre logo após o jogo. Seria igualmente relevante analisar mais profundamente as causas específicas das mudanças de sentimentos durante o jogo, identificando eventos chave que desencadeiam reações positivas ou negativas. Além disso, expandir o estudo para outros eventos esportivos e compará-los permitiria verificar se os padrões observados são consistentes em diferentes contextos esportivos.

Essas conclusões e insights não só validam e desafiam as hipóteses iniciais, mas também fornecem uma base sólida para a compreensão do comportamento dos torcedores nas redes sociais e a aplicação de análise de sentimentos em contextos esportivos.

## **7. REFERÊNCIAS BIBLIOGRÁFICAS**

- Alecsa, D., Denev, D., Olteanu, A., Weber, I., & Gatica-Perez, D. (2019). News and User Engagement on Social Media: A Study of News Reactions on Facebook. *ACM Transactions on Social Computing*.
- Aggarwal, C. C., & Zhai, C. (2012). *Mining Text Data*. Springer.
- Bakshy, E., Hofman, J. M., Mason, W. A., & Watts, D. J. (2011). Everyone's an influencer: Quantifying influence on Twitter. *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*.
- Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1), 1-8.
- Boyd, D., & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, communication & society*, 15(5), 662-679.
- Boyd, D., & Ellison, N. (2007). Social network sites: Definition, history, and scholarship. *Journal of Computer-Mediated Communication*, 13(1), 210-230.
- Boyle, R., & Haynes, R. (2009). *Power Play: Sport, the Media and Popular Culture*. Edinburgh University Press.
- Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1), 5-32.
- Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and Regression Trees*. Chapman and Hall/CRC.
- Cambria, E., & White, B. (2014). Jumping NLP Curves: A Review of Natural Language Processing Research. *IEEE Computational Intelligence Magazine*.
- Cambria, E., Schuller, B., Xia, Y., & Havasi, C. (2017). New Avenues in Opinion Mining and Sentiment Analysis. *IEEE Intelligent Systems*.

- Cha, M., Cho, H., & Myers, S. A. (2010). Measuring user influence in Twitter: The million follower fallacy. *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*.
- Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). *CRISP-DM 1.0: Step-by-step data mining guide*. SPSS Inc.
- Citron, D. K., & Norton, H. (2011). Intermediaries and hate speech: Fostering digital citizenship for the twenty-first century. *Boston University Law Review*, 91, 1435-1464.
- Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3), 273-297.
- Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1), 21-27.
- Danescu-Niculescu-Mizil, C., Gamon, M., & Dumais, S. (2011). Mark my words!: Linguistic style accommodation in social media. *Proceedings of the 20th international conference on World Wide Web*.
- Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. *AI Magazine*.
- Feldman, R., & Sanger, J. (2007). *The Text Mining Handbook*. Cambridge University Press.
- Frederick, E. L., Pegoraro, A., & Sanderson, J. (2020). Sports and Social Media: Key Decisions and Considerations for Future Research. *Communication & Sport*.
- Glez-Peña, D., Lourenço, A., López-Fernández, H., Reboiro-Jato, M., & Fdez-Riverola, F. (2014). Web scraping technologies in an API world. *Briefings in Bioinformatics*.
- Hale, S. A. (2017). How to make sense of weak and strong ties: Social network analysis in the era of social media. *Social Networks*.
- Hand, D. J., Mannila, H., & Smyth, P. (2001). *Principles of data mining*. MIT press.
- Hinduja, S., & Patchin, J. W. (2018). *Bullying beyond the schoolyard: Preventing and responding to cyberbullying*. Sage Publications.
- Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied Logistic Regression*. John Wiley & Sons.
- Hutchins, B., & Rowe, D. (2012). *Sport Beyond Television: The Internet, Digital Media and the Rise of Networked Media Sport*. Routledge.
- Java, A., Song, X., Finin, T., & Tseng, B. (2007). Why we Twitter: Understanding microblogging usage and communities. *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis*.

- Kumar, V., & Reddy, C. K. (Eds.). (2016). *Twitter data analytics*. Springer.
- Krotov, V., & Silva, L. (2018). Legality and ethics of web scraping. *Proceedings of the 51st Hawaii International Conference on System Sciences*.
- Kwak, H., Lee, C., Park, H., & Moon, S. (2010). What is Twitter, a social network or a news media?. In *Proceedings of the 19th international conference on World wide web* (pp. 591-600).
- Liu, B. (2012). *Sentiment Analysis and Opinion Mining*. Morgan & Claypool Publishers.
- Liu, B., Li, X., & Zhang, W. (2020). Sentiment analysis and opinion mining in social media: Challenges and applications. *Social Network Analysis and Mining*.
- Mackenzie, R., & Cushion, C. (2013). Performance Analysis in Football: A Critical Review and Implications for Future Research. *Journal of Sports Sciences*.
- Mander, J. (2019). *Global Social Media Trends 2019*. GlobalWebIndex.
- Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press.
- Miner, G., Elder, J., Fast, A., Hill, T., Nisbet, R., & Delen, D. (2012). *Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications*. Academic Press.
- Miranda, F. A. (2013). Marketing digital e o futebol brasileiro: um estudo sobre a interação entre clubes e torcida nas mídias sociais. *Esporte e Sociedade*, (22).
- Mitchell, R. (2018). *Web Scraping with Python: Collecting Data from the Modern Web*. O'Reilly Media.
- Panjwani, R. (2021). *Digital Sports Marketing*. Routledge.
- Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and trends® in information retrieval*, 2(1-2), 1-135.
- Parganas, P., Anagnostopoulos, C., & Chadwick, S. (2017). 'You'll Never Walk Alone': A Scale Development for Measuring Fan Engagement in Social Media. *Sport Management Review*.
- Primack, B. A., Shensa, A., Sidani, J. E., Whaite, E. O., Lin, L. Y., Rosen, D., ... & Miller, E. (2017). Social media use and perceived social isolation among young adults in the US. *PLoS ONE*, 12(6), e0179611.
- Russell, M. A. (2019). *Mining the Social Web*. O'Reilly Media.
- Sarmiento, H., Anguera, M. T., Pereira, A., & Araújo, D. (2018). Talent Identification and Development in Male Football: A Systematic Review. *Sports Medicine*.

- Smith, A. (2016). *Introduction to Sports Communication*. Routledge.
- Starbird, K., & Palen, L. (2012). (How) will the revolution be retweeted?: Information diffusion and the 2011 Egyptian uprising. *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*.
- Stieglitz, S., & Dang-Xuan, L. (2013). Emotions and Information Diffusion in Social Media—Sentiment of Microblogs and Sharing Behavior. *Journal of Management Information Systems*.
- Tan, P. N., Steinbach, M., & Kumar, V. (2016). *Introduction to data mining*. Pearson Education.
- Thompson, P. (2020). *Social Media and Sports*. Routledge.
- Weiss, S. M., Indurkha, N., Zhang, T., & Damerau, F. J. (2010). *Text Mining: Predictive Methods for Analyzing Unstructured Information*. Springer.
- Williams, J. (2013). *Managing Social Media in Sport*. Springer.
- Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). *Data mining: Practical machine learning tools and techniques*. Morgan Kaufmann.

## 8. ANEXOS

### ANEXO A – Retorno do scrapping de dados

```
dados[0].keys()

KeysView(<Tweet id=1719919753672716345 text='RT @Palmeiras: O TIME DA VIRADA E DO AMOR! \n\n#AvantiPalestra #BOTxPAL\n#JuntosNoBrasil'>
https://t.co/PxcMtRZCK9'>)
```


```
resposta.includes
```


```
{'users': [<User id=947280744711376901 name=Adriany Martins username=Dry_flhadeOya>,
<User id=106795441 name=Léo Gama @username=Leogama7>,
<User id=751848397519347712 name=Rafa username=rafadossantos07>,
<User id=58924356 name=Transamérica Esportes username=tresportes>,
<User id=979751678231613441 name=RUBENS @username=rubinho_fra>,
<User id=955154527778418688 name=Arte Palestrina username=ArtePalestrina>,
<User id=26940916 name=Harry Yamamoto username=harryyamamoto>,
<User id=3240648647 name=Thalles Mondin username=MondinThalles>,
<User id=54316705 name=Sandro Pontes username=Bravado21>,
<User id=1108824255884607489 name=Infiltrado username=Infiltr46596690>],
'places': [<Place id=35e1542602b65f19 full_name=Santo André, Brasil>]}
```


```
resposta.includes
```

```
{'users': [<User id=947280744711376901 name=Adriany Martins username=Dry_flhadeOya>,
<User id=106795441 name=Léo Gama @username=Leogama7>,
<User id=751848397519347712 name=Rafa username=rafadossantos07>,
<User id=58924356 name=Transamérica Esportes username=tresportes>,
<User id=979751678231613441 name=RUBENS @username=rubinho_fra>,
<User id=955154527778418688 name=Arte Palestrina username=ArtePalestrina>,
<User id=26940916 name=Harry Yamamoto username=harryyamamoto>,
<User id=3240648647 name=Thalles Mondin username=MondinThalles>,
<User id=54316705 name=Sandro Pontes username=Bravado21>,
<User id=1108824255884607489 name=Infiltrado username=Infiltr46596690>],
'places': [<Place id=35e1542602b65f19 full_name=Santo André, Brasil>]}
```






**Tweet Scraper V2 (Pay Per Result) - X / Twitter Scraper**
\$0.30 / 1,000 tweets

apidojo/tweet-scraper
1k monthly users
96% runs succeeded
Crafted by  API Dojo
Maintained by Community

Save & Start
Create task
API
...

⚡ Lightning-fast search, URL, list, and profile scraping, with customizable filters. At \$0.30 per 1000 tweets, and 30-80 tweets per second, it is ideal for researchers, entrepreneurs, and businesses! Get comprehensive insights from Twitter (X) now!

Input
Information
Runs 30
Builds 139
Integrations 0
Monitoring
Issues 0
Saved tasks 0

A ferramenta que será utilizada no site do Apify

Run options

MAXIMUM RESULTS

Unlimited

BUILD

latest

TIMEOUT

0s

MEMORY

512 MB

Actor running slowly or crashing? Try increasing memory.

**Build**

latest

**Timeout** ?

0

seconds

+

-

☒ No timeout ?

No timeout can result in infinite runs.

**Memory** ?

512

MB

+

-

More memory means more CPU horsepower.  
512 MB gives you 0.125 CPU cores.

**Maximum charged results** (optional) ?

1000

tweets

+

-

☒ No maximum limit ?

Configurações de duração da raspagem no Apify

Start date (optional) ?

2024-04-17

×

End date (optional) ?

2024-04-18

×

Opção dos dias em que os tweets foram realizados

Filter Tweets

You can filter tweets by using these options.

Tweet language (optional) ?

English x | v

☐ Only verified users ?

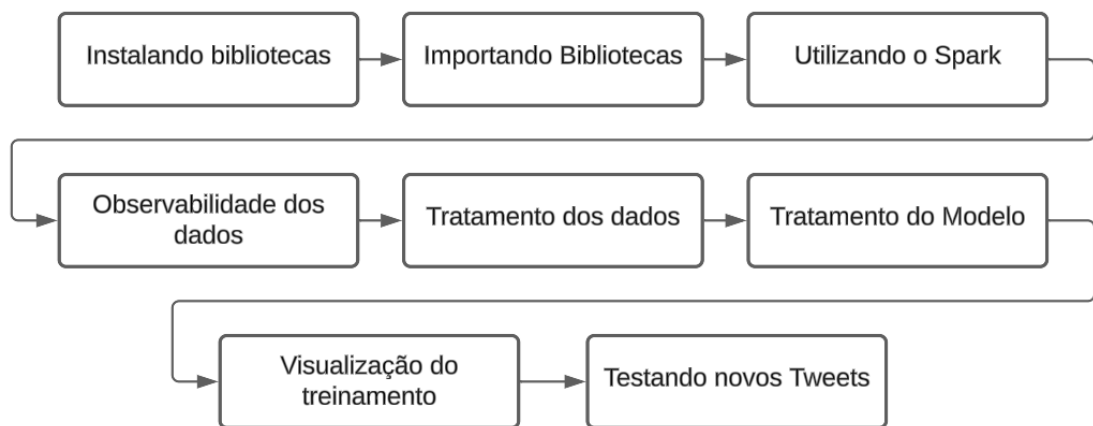
☐ Only Twitter Blue ?

☐ Only image ?

☐ Only video ?

☐ Only quote ?

Escolha de filtros para os tweets.



Fluxograma de como irá funcionar o treinamento do modelo

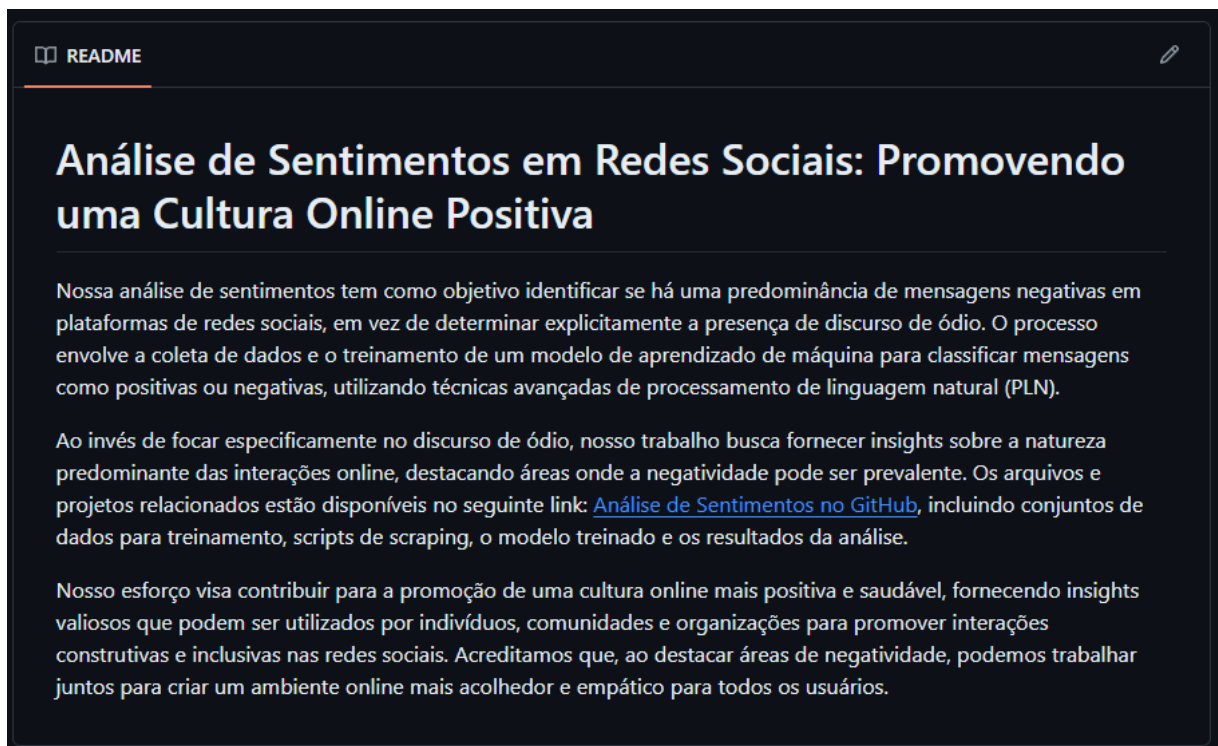
8 items

- 1  ×
- 2  ×
- 3  ×
- 4  ×
- 5  ×
- 6  ×
- 7  ×








Inserção das palavras-chaves que serão utilizadas para a raspagem de dados

✓ <b>Succeeded</b> Actor succeeded with 1049 results in the dataset				
RESULTS	REQUESTS	PRICE ?	STARTED	DURATION
1049	0 of 0 handled	\$0.3147	2024-05-15 13:41	34 s

Resultado obtidos através da raspagem

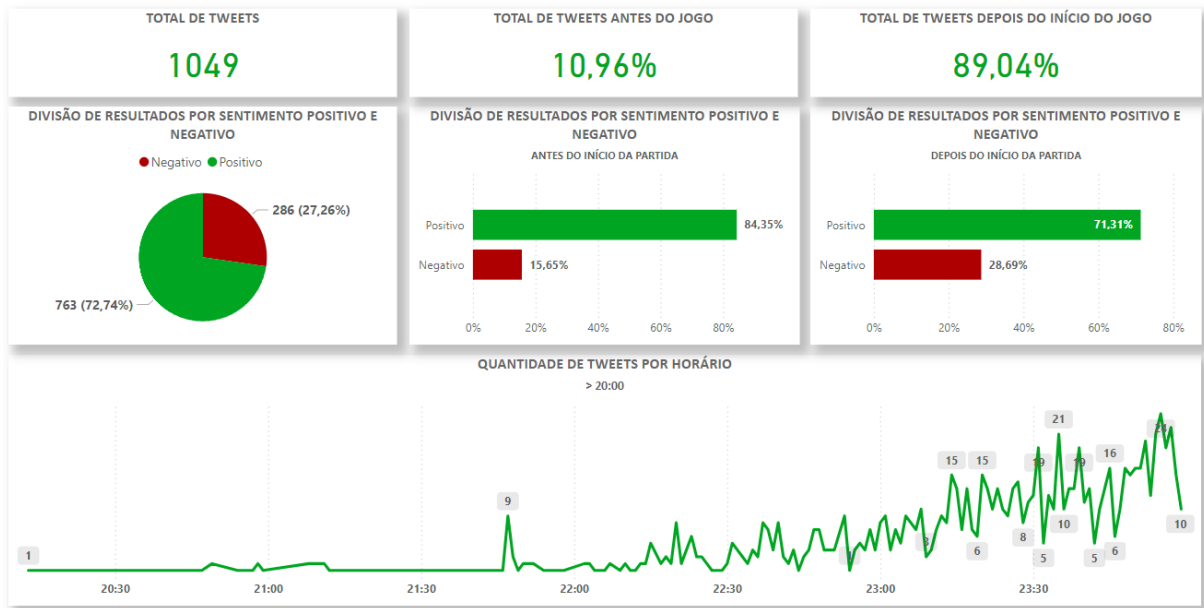


READ-ME do GitHub.

 <b>lucanegrini</b>	Add files via upload	69faf09 · now	5 Commits
 Dashboard's TCC.pbix	Add files via upload		now
 README.md	Update README.md		last week
 analise_sentimento.ipynb	Add files via upload		last week
 best_model.pkl	Add files via upload		last week
 resultados.csv	Add files via upload		last week
 resultados.xlsx	Add files via upload		last week

Arquivos do trabalho de conclusão de curso no GitHub.

## Mineração de texto em conteúdo esportivo baseado em web Scraping através de rede social



Captura de tela do dashboard feito para o trabalho.