UNIVERSIDADE FEDERAL RURAL DO RIO DE JANEIRO
INSTITUTO MULTIDISCIPLINAR

JULIA DOS SANTOS BARTOLO

Análise de Dados do Twitter sobre Agentes Dopantes na Sociedade

Prof. Filipe Braida do Carmo, D.Sc.
Orientador

Nova Iguaçu, Julho de 2023

Análise de Dados do Twitter sobre Agentes Dopantes na Sociedade

Julia dos Santos Bartolo

Projeto Final de Curso submetido ao Departamento de Ciência da Computação do Instituto Multidisciplinar da Universidade Federal Rural do Rio de Janeiro como parte dos requisitos necessários para obtenção do grau de Bacharel em Ciência da Computação.

Apresentado por:

_____

Julia dos Santos Bartolo

Aprovado por:

_____

Prof. Filipe Braida do Carmo, D.Sc.

_____

Prof. Leandro Guimaraes Marques Alvim, D.Sc.

_____

Prof. Bruno José Dembogurski, D.Sc.

NOVA IGUAÇU, RJ - BRASIL

Julho de 2023

Emitido em 31/07/2023

DOCUMENTOS COMPROBATÓRIOS Nº 13577/2023 - CoordCGCC (12.28.01.00.00.98)

(Nº do Protocolo: NÃO PROTOCOLADO)

(Assinado digitalmente em 07/08/2023 20:21)
BRUNO JOSE DEMBOGURSKI
PROFESSOR DO MAGISTERIO SUPERIOR
DeptCC/IM (12.28.01.00.00.83)
Matrícula: ###249#4

(Assinado digitalmente em 04/08/2023 21:27)
FILIPE BRAIDA DO CARMO
PROFESSOR DO MAGISTERIO SUPERIOR
DeptCC/IM (12.28.01.00.00.83)
Matrícula: ###295#4

(Assinado digitalmente em 07/08/2023 17:23)
LEANDRO GUIMARAES MARQUES ALVIM
PROFESSOR DO MAGISTERIO SUPERIOR
DeptCC/IM (12.28.01.00.00.83)
Matrícula: ###008#2

(Assinado digitalmente em 05/08/2023 11:13)
JULIA DOS SANTOS BARTOLO
DISCENTE
Matrícula: 2019######6

Visualize o documento original em https://sipac.ufrrj.br/documentos/ informando seu número: **13577**, ano: **2023**, tipo: **DOCUMENTOS COMPROBATÓRIOS**, data de emissão: **04/08/2023** e o código de verificação: **dcd01f28ed**

Agradecimentos

Quero, primeiramente agradecer a Deus por nunca me desamparar e por me permitir chegar até aqui, me surpreendendo positivamente em toda essa jornada.

Gostaria de agradecer à minha família, em especial aos meus pais, Autelina e Paulo, por sempre me incentivarem, me compreenderem e me ajudarem durante todo o processo de chegar à graduação até a conclusão. Sem o apoio de vocês, eu não teria conseguido. Sou grata também ao meu irmão Julio, que é minha inspiração desde pequena e me impulsionou em toda caminhada acadêmica.

Agradeço ao meu companheiro Luan, que me incentivou a ingressar nesta graduação e esteve ao meu lado em cada desafio que enfrentei. Seu apoio, confiança e acolhimento foram fundamentais para o que eu chegasse até aqui. Sou grata também aos meus sogros, Giovania e Reinaldo, por me acompanharem ao longo desses anos, sempre preocupados e torcendo pelo meu sucesso.

Queria muito agradecer aos meus amigos, tanto aos que vieram antes da graduação, como a Evelyn e a Christyne, que sempre torceram muito por mim, quanto aos amigos que fiz durante a graduação, que fizeram toda essa jornada mais leve e alegre. Em especial, agradeço aos meus amigos Ana Luiza e Lucas. Nunca esquecerei as noites viradas em meio aos estudos e as muitas risadas compartilhadas.

Gostaria de agradecer a todo Departamento de Ciência da Computação da (UFRRJ-IM), em especial ao meu orientador, Filipe Braida, que sempre admirei e que me motivou durante todo o trabalho sempre muito solicito e compreensivo.

Meus sinceros agradecimentos a todos vocês.

RESUMO

Análise de Dados do Twitter sobre Agentes Dopantes na Sociedade

Julia dos Santos Bartolo

Julho/2023

Orientador: Filipe Braida do Carmo, D.Sc.

O uso de agentes dopantes deixou de ser um problema restrito apenas aos atletas de elite e tornou-se uma questão ampla de saúde pública. O aumento preocupante do uso deliberado dessas substâncias por jovens atletas e praticantes de musculação em busca de um padrão estético merece destaque, considerando os diversos riscos à saúde que essas substâncias acarretam. Neste contexto, o presente trabalho utiliza o *Twitter* como fonte de dados, escolhido por seu grande alcance e pelas publicações instantâneas que refletem diversos contextos sociais. A coleta de dados teve como objetivo principal extrair *tweets* que contivessem os termos relacionados a substâncias proibidas no esporte e algumas modalidades esportivas. Com base nesses dados, foram realizadas análises quantitativas e qualitativas, buscando entender tanto os dados quanto o contexto em que a maioria das publicações dos usuários se encontrava. Essa análise é fundamental para compreender a visão social em relação ao uso desses agentes dopantes e pode auxiliar na proposição de medidas de conscientização e prevenção.

ABSTRACT

Análise de Dados do Twitter sobre Agentes Dopantes na Sociedade

Julia dos Santos Bartolo

Julho/2023

Advisor: Filipe Braida do Carmo, D.Sc.

The use of doping agents is no longer a problem restricted to elite athletes and has become a broad public health issue. The worrying increase in the deliberate use of these substances by young athletes and bodybuilders in search of an aesthetic standard deserves to be highlighted, considering the various health risks that these substances entail. In this context, the present work uses Twitter as a data source, chosen for its great reach and for the instantaneous publications that reflect different social contexts. The main objective of data collection was to extract tweets that contained terms related to prohibited substances in sport and some sports. Based on these data, quantitative and qualitative analyzes were carried out, seeking to understand both the data and the context in which most of the users' publications were found. This analysis is essential to understand the social view regarding the use of these doping agents and can help propose awareness and prevention measures.

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas e Siglas

API	Application Programming Interface
AAS	Anabolic Androgenic Steroids
IDC	International Data Corporation
KDD	Knowledge Discovery in Databases
WADA	World Anti-Doping Agency
ABCD	Autoridade Brasileira de Controle de Dopagem
SARM	Selective Androgen Receptor Modulator
EPO	Eritropoietinas
COB	Comitê Olímpico do Brasil
CFM	Conselho Federal de Medicina

Sumário

Capítulo 1

Introdução

O uso de substâncias farmacologicamente ativas que aprimoram o desempenho físico, aumentam a massa muscular e a força ocorrem há séculos. No entanto, nos últimos anos, tem havido um aumento considerável nessa prática (SJöQVIST; GARLE; RANE, 2008),(TRAJANO F, 2023). Essas substâncias têm sido amplamente utilizadas por indivíduos que buscam melhorar sua aparência estética ou obter resultados mais rápidos, não se limitando apenas a competidores de fisiculturismo e atletismo.

Com a crescente popularização das redes sociais, a busca por um corpo esteticamente atraente tem se intensificado, uma vez que o tempo dedicado a essas plataformas e a exposição a padrões de beleza são muito superiores em comparação à década anterior. Esse contexto impulsionou ainda mais a procura por métodos que possam auxiliar no alcance desse padrão idealizado. Durante os meses de novembro a janeiro por exemplo, observar-se um aumento significativo na busca por agentes anabolizantes, período em que as pessoas geralmente se mostram mais preocupadas com a aparência física, devido à aproximação do verão (MORAES1, 2015).

Em 11/04/2023, o Conselho Federal de Medicina (CFM) publicou a resolução nº 2.333/2023 (MEDICINA, 2023), proibindo a prescrição médica de esteroides e anabolizantes para fins estéticos, devido aos diversos efeitos colaterais associados a essas substâncias. Entretanto, mesmo com essa proibição, a busca por resultados

estéticos continua sendo uma preocupação constante, resultando no crescimento da produção falsificada dessas substâncias. Entre os anos de 2007 e 2010, os esteroides anabolizantes ocuparam o segundo lugar entre os medicamentos mais apreendidos pela Polícia Federal no país, conforme relatado em (AMES; SOUZA, 2012), evidenciando a relevância dessa problemática.

As redes sociais, apesar de atuarem como grandes impulsionadoras para o uso desses agentes, também podem ser uma ótima ferramenta para compreender como estes compostos são percebidos pela sociedade. Nesse contexto as redes sociais, especialmente o Twitter, se tornaram uma das principais plataformas de comunicação e interação. Milhões de pessoas ao redor do mundo utilizam essas plataformas para compartilhar ideias e opiniões de maneira instantânea.

Com base nos dados coletados pelo *Twitter*, a análise proporciona uma visão aprofundada dos padrões de comportamento e tendências associadas ao uso de substâncias anabolizantes. Essa análise abrangente nos permite compreender a magnitude do problema, identificar as motivações que impulsionam o uso desses compostos e explorar possíveis influências relacionadas a diferentes substâncias. Além disso, a abordagem quantitativa da análise permite mensurar os dados em grande escala, oferecendo uma perspectiva sólida sobre a amplitude e a relevância desse tema na sociedade. Por meio dessa metodologia, podemos obter conclusões valiosas para a conscientização sobre os riscos associados ao uso de anabolizantes e auxiliar na adoção de medidas preventivas eficazes.

1.1 Objetivo

O objetivo deste trabalho é realizar uma análise abrangente, abordando aspectos qualitativos e quantitativos, dos dados relacionados às substâncias proibidas na rede social do *Twitter*. Essa análise tem como propósito compreender a percepção social em relação a esses agentes dopantes e suas implicações na sociedade.

1.2 Agradecimentos

O presente trabalho contou com a valiosa colaboração do Comitê Olímpico do Brasil (COB), que generosamente concedeu acesso aos dados utilizados nesta pesquisa, os quais foram obtidos como resultado de uma parceria com a Universidade Federal do Rio de Janeiro. Essa cooperação foi essencial para o desenvolvimento e enriquecimento deste estudo, permitindo uma análise mais abrangente e fundamentada sobre o tema abordado.

1.3 Organização do Trabalho

O trabalho é composto por cinco capítulos, abrangendo introdução, mineração de dados, impacto de agentes dopantes no Twitter, análise dos dados e, por fim, a conclusão. A seguir, será apresentado o detalhamento de cada capítulo.

- Capítulo 1: Constitui uma introdução abrangente sobre o contexto relacionado ao trabalho em questão
- Capítulo 2: Constitui uma fundamentação técnica relacionada a mineração de dados e o processos de extração de conhecimento da base de dados utilizadas ao longo do trabalho.
- Capítulo 3: Constitui em uma introdução detalhada sobre o tema do *doping* e as substâncias proibidas. Explora o impacto dessas substâncias na vida de atletas e não-atletas, destacando os perigos associados ao seu uso. Além disso, analisa a influência das redes sociais nesse processo e como o uso dessas substâncias não se limita mais apenas ao ambiente esportivo. Serão discutidas também as medidas adotadas para conscientizar a população sobre os riscos relacionados ao uso dessas substâncias.
- Capítulo 4: Constitui uma análise tanto quantitativa como qualitativa a cerca dos dados. Na análise quantitativa são realizadas análise referentes todas as tabelas da base de dados, como elas se comportam e quais são as suas principais

características, já na análise qualitativa são realizadas análises referente apenas a um grupo específico de substâncias que são mais populares no Brasil.

- Capítulo 5: Constitui nas conclusões finais obtidas e nos possíveis trabalhos futuros voltados a educação, monitoração e classificações.

Capítulo 2

Mineração de Dados

Neste capítulo discutiremos o conceito de Mineração de Dados e sua importância na era da informação. Abordaremos também as etapas principais do processo Knowledge Discovery in Databases (KDD), que envolvem desde a seleção e pré-processamento dos dados até a interpretação dos resultados obtidos.

2.1 Descoberta de Conhecimento em Bases de Dados

O avanço da tecnologia tem impulsionado a revolução de grandes conjuntos de dados, possibilitando que as organizações coletem uma quantidade massiva desses registros. Em 2020, a Corporação Internacional de Dados, International Data Corporation (IDC) previu que mais de 59 zettabytes (59 trilhões de gigabytes) seriam gerados, capturados, armazenados e consumidos em todo o mundo até o final do ano. No entanto, ao término de 2020, ficou evidente que esses dados totalizaram 64,2 zettabytes, superando as expectativas (WOODIE, 2022).

Com o constante aumento na quantidade de dados, torna-se necessário o uso de técnicas que possibilitem compreender e extrair valor dessas informações. Nesse contexto, a Mineração de Dados desempenha um papel fundamental em estruturar e interpretar os dados. No entanto, a Mineração de Dados é apenas uma parte do processo de descoberta de conhecimento em bancos de dados, conhecido como

KDD, que envolve etapas de transformação, desde o pré-processamento até o pós-processamento dos resultados da mineração dos dados (TAN, 2013). Um exemplo dessa abordagem pode ser observado na figura 2.1.

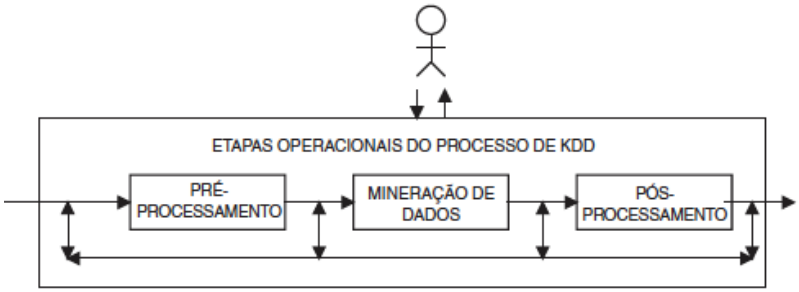

Figura 2.1: Etapas Operacionais do Processo de KDD

Fonte: Data Mining. Um guia prático.

2.1.1 Etapas do Processo de KDD

A obtenção de conhecimento a partir de uma vasta quantidade de dados é considerada um processo interativo e iterativo, não sendo tratada como um sistema de análise automática isolada, mas sim como um processo centrado na interação entre usuários, especialistas do domínio e responsáveis pela aplicação. Portanto, não se pode esperar que a extração de conhecimento seja útil simplesmente submetendo um conjunto de dados a uma *"caixa preta"* (MANNILA, 1996).

A Extração de Conhecimento da Base de Dados começa pelo entendimento do domínio da aplicação, levando em consideração fatores como o objetivo da aplicação e a fonte dos dados de entrada. Em seguida, ocorre a seleção dos dados com base na identificação do problema, preparando-os para serem submetidos aos métodos e ferramentas de extração de padrões. Na etapa de extração de padrões, o objetivo central é descobrir modelos ou conhecimento relevante a partir dos dados. O resultado dessa etapa é uma predição específica para o problema em questão, permitindo assim a tomada de decisões fundamentadas. Por fim, o resultado obtido é avaliado em relação ao problema inicial, servindo como base para processos de tomada de decisão quando os resultados são positivo.

2.1.1.1 Entendimento do Domínio

Nesta etapa, especialistas do domínio colaboram com os usuários finais para compreender os objetivos da aplicação, identificar requisitos e estabelecer o escopo do projeto de descoberta de conhecimento. Além dessa análise inicial para definir metas, restrições e objetivos principais, o conhecimento do domínio permeia todas as fases do processo. Especificamente na etapa de pré-processamento, esse entendimento pode auxiliar na seleção do conjunto de dados mais adequado para a mineração.

2.1.1.2 Pré-Processamento

O objetivo da etapa de pré-processamento é realizar uma série de transformações nos dados brutos, com o intuito de garantir que estejam com qualidade, devidamente organizados e tratados para a fase de mineração dos dados. Além disso, o pré-processamento também pode envolver ajustes específicos para que os dados se adéquem melhor a uma técnica de mineração específica, aprimorando assim os resultados obtidos.

Diversas transformações podem ser aplicadas nessa fase. A integração permite mesclar dados de diferentes bases de dados, enquanto a seleção identifica as informações relevantes para o processo. A limpeza de dados é responsável por corrigir inconsistências e remover ruídos. As transformações de dados, como a normalização, oferecem maior precisão e eficiência aos algoritmos. Já a redução de dados permite diminuir o tamanho dos dados por meio de agregação e eliminação de recursos redundantes. (HAN; KAMBER, 2011)

Essas técnicas aplicadas antes da Mineração de Dados têm o potencial de melhorar significativamente os resultados gerais da mineração. O pré processamento desempenha um papel crucial na garantia da qualidade dos dados, maximizando a eficiência dos algoritmos de mineração. Ao preparar os dados de forma adequada, essas técnicas tornam os dados prontos para a extração de conhecimento, possibilitando a descoberta de padrões e tendências.

2.1.1.3 Extração de Padrões

Um grande equívoco popular é esperar que os sistemas de Mineração de Dados possam, de forma autônoma, buscar todo o conhecimento contido nos dados, sem intervenção ou orientação humana. Na realidade, se isso ocorresse, os padrões extraídos poderiam ser muito generalistas e não estariam necessariamente alinhados com o objetivo estabelecido no início do processo. (HAN; KAMBER, 2011)

Assim como em todo o processo de Mineração de Dados, essa etapa também é iterativa. Portanto, é comum realizar experimentações e refinamentos, testando diferentes configurações e parâmetros. Essas iterações possibilitam aprimorar os resultados obtidos, garantindo que sejam precisos e confiáveis (REZENDE. JB PUGLIESI.; PAULA, 2003).

No entanto, é importante ressaltar que a eficácia da etapa de Mineração de Dados depende do algoritmo escolhido para realizar a tarefa. Existem diversos algoritmos disponíveis, e a escolha adequada depende exclusivamente do problema a ser resolvido. Cada algoritmo possui características distintas e é mais adequado para certos tipos de dados e objetivos específicos.

Portanto, é necessário considerar cuidadosamente as características do problema em questão, bem como as propriedades dos dados disponíveis, antes de selecionar o algoritmo mais apropriado. A escolha correta do algoritmo desempenha um papel crucial no sucesso da Mineração de Dados, permitindo obter resultados mais precisos e relevantes para a resolução do problema em foco.

2.1.1.4 Pós-Processamento

A etapa de pós-processamento, por sua vez, engloba o tratamento do conhecimento adquirido na fase de Extração de Padrões. Nessa etapa, especialistas em KDD e no escopo principal do projeto avaliam as informações obtidas e buscam identificar outras alternativas que facilitem a compreensão dos dados.

A entrega de conhecimento é auxiliada por diversas medidas, como a Simplificação

do Modelo, que busca remover detalhes complexos sem comprometer a informação relevante. Regras de validação, como a acurácia, que indica o percentual de resultados corretos e a abrangência que está relacionada a capacidade do modelo de lidar com diferentes cenários. Essas técnicas são utilizadas para avaliar a qualidade e o poder de generalização dos modelos obtidos, garantindo um conhecimento compreensível, confiável e aplicável para a tomada de decisões (GOLDSCHMIDT; PASSOS, 2005).

Além disso, o pós-processamento pode envolver a aplicação de técnicas de visualização de dados, como gráficos, tabelas e outros recursos visuais, para auxiliar na representação clara e intuitiva das informações. Essa abordagem ajuda a comunicar e transmitir o conhecimento extraído de forma mais compreensível e acessível aos envolvidos no projeto

Após a análise do conhecimento, caso os resultados obtidos não atendam às expectativas do usuário final ou não estejam alinhados com o objetivo proposto, é possível reexecutar o processo de Extração de Padrões, ajustando os parâmetros ou aprimorando o processo de seleção dos dados. Dessa forma, é possível obter resultados mais satisfatórios e coerentes com as necessidades do projeto. (REZENDE. JB PUGLIESI.; PAULA, 2003)

Capítulo 3

Impacto de agentes dopantes no Twitter

Este capítulo tem como objetivo abordar três tópicos fundamentais relacionados ao tema do doping: seu impacto no esporte e na sociedade, como foi realizada a coleta de dados do Twitter como fonte de informações relevantes para o presente trabalho e a modelagem destes dados em um banco relacional para uma futura análise.

3.1 Doping nos Esportes e na Sociedade

A origem da palavra *doping* é controvérsia, porém está relacionada a compostos que são utilizados para melhorar o desempenho do usuário em termos físicos ou mentais. No esporte, o *doping* é caracterizado como a violação de uma ou mais regras estabelecidas pelo Código Antidopagem Mundial, que são diretrizes internacionais para combater o uso de substâncias proibidas que fornecerem vantagens injustas e comprometerem a integridade e igualdade competitiva. Os efeitos do uso dessas substâncias podem incluir aumento da força, resistência, recuperação acelerada, melhora da concentração, entre outros. No entanto, seu uso pode acarretar em sérios riscos para a saúde dos usuários e minar os princípios de jogo limpo e ética esportiva. (BIRD et al., 2016)

Ser um atleta de alto nível é uma posição muito cobiçada, o que pode levar uma pessoa a recorrer ao *doping*. Contudo, o uso de substâncias para aprimorar o desempenho vai além dos atletas de elite, a busca por melhores resultados e a competitividade despertam comportamentos similares tanto entre atletas de níveis variados como na população geral. O uso de Anabolic Androgenic Steroids (AAS) tem se tornado cada vez mais comum entre jovens atletas em ambientes escolares e amadores não participantes de competições (NILSSON S.; ALLEBECK, 2005). Além do que, o uso indevido dessas substâncias vem aumentando dentro de academias, especialmente para aqueles que priorizam a aparência física.

A pressão estética de um corpo ideal sempre foi uma questão relevante na sociedade e, ao longo dos anos, tem gerado consequências significativas para indivíduos ao redor do mundo. A imposição de padrões de beleza inatingíveis é frequentemente promovida pela mídia e pela indústria da moda, criando expectativas irreais em relação à aparência física. Com o advento das redes sociais, a comparação constante com corpos "idealizados" tornou-se ainda mais intensa, resultando em sentimentos de grande insatisfação com o próprio corpo (AZIZ, 2017).

As redes sociais exercem uma influência significativa no uso de AAS, sendo os homens os maiores usuários, apresentando uma proporção mais elevada em comparação às mulheres (HILKENS L.; WOERTMAN, 2021). Distúrbios psicológicos são comuns entre a população que faz o consumo dessas drogas, essas pessoas possuem maiores sintomas de raiva, ansiedade, depressão e baixa auto-estima comparada com os não-usuários (GESTSDOTTIR S.; SIGFUSDOTTIR, 2021). Outro efeito colateral da utilização de AAS são os problemas físicos como doenças cardiovasculares, atrofia testicular, aumento da pressão arterial e outros (TOKISH J. M.; HAWKINS, 2004).

Ao considerar o uso de substâncias proibidas, fica evidente a existência de dois cenários distintos: um em que seu uso é estritamente proibido, podendo resultar no banimento do atleta no esporte, e outro em que é popularizado e, em alguns casos, até mesmo aceito pela sociedade como normal. Como resultado, os atletas enfrentam um dilema significativo, pois separar e lidar com esses dois cenários pode ser extremamente desafiador. Dessa forma, torna-se ainda mais importante promover

movimentos que visem a educação e conscientização sobre o uso desses elementos, a fim de mitigar os problemas decorrentes da normalização de seu uso, bem como os perigos associados a eles.

A Agência Mundial Antidopagem, conhecida internacionalmente como World Anti-Doping Agency (WADA), reconhece a importância da educação como um dos pilares fundamentais em suas estratégias para combater o doping. Isso fica evidente no Código Mundial Antidopagem de 2021, onde a WADA visa fornecer informações precisas e atualizadas sobre os riscos do *doping*, as consequências para a saúde, as proibições e as consequências legais (WADA, 2021). Além disso, a WADA trabalha em estreita colaboração com as agências nacionais antidopagem, como a Autoridade Brasileira de Controle de Dopagem (ABCD), organizações esportivas e governos para disseminar essas informações e promover a conscientização.

Com o objetivo de informar e educar atletas e jovens sobre as consequências do uso indevido de substâncias proibidas, tanto para a saúde quanto para possíveis penalidades, o Comitê Olímpico do Brasil (COB) implementou em outubro de 2018 a área de Educação e Prevenção ao Doping. É preocupante o fato de que muitos jovens ou futuros atletas estão consumindo substâncias proibidas por falta de conhecimento. Portanto, torna-se evidente que esta problemática vai vai além do contexto esportivo, transformando-se em uma preocupação significativa de saúde pública (AITH, 2013).

Além disso, é importante destacar que o tema do *doping* vai além da esfera esportiva e de saúde. Um exemplo marcante disso foi o escândalo que envolveu a Rússia em 2014. Nesse caso, o país foi acusado de participar de um esquema de *doping* sistemático, que não apenas envolvia os atletas, mas também autoridades esportivas e até mesmo serviços de segurança, com o objetivo de encobrir resultados positivos nos exames *antidoping*. Esse escândalo ressaltou a complexidade e a gravidade do problema em questão, revelando como o *doping* pode afetar diversos aspectos da sociedade para além do âmbito esportivo, levantando questões éticas e de integridade no esporte mundial.

Nesse cenário, a análise dos textos compartilhados por usuários do Twitter desempenha um papel significativo no combate ao *doping*, oferecendo resultados

interessantes para uma compreensão abrangente das percepções e debates públicos relacionados ao uso de substâncias proibidas no esporte. Ao examinar esses textos nas mídias sociais, é possível identificar padrões, opiniões divergentes e comportamentos suspeitos ou que promovem o uso de *doping*. Essas informações são fundamentais para aprimorar as estratégias de conscientização, políticas de prevenção e medidas de combate ao doping, visando a preservação da saúde dos usuários e de um ambiente esportivo íntegro.

3.2 Coleta dos Dados

A partir do surgimento das redes sociais como *MySpace* e *Friendster* na década de 2000, esses serviços se multiplicaram e passaram a dominar a Internet, oferecendo diversas funcionalidades e formas de compartilhar o cotidiano. Como resultado, hoje em dia mais de 4 bilhões de pessoas em todo o mundo utilizam essas redes sociais. Entre eles, o Twitter é uma das mais relevantes. Com seus mais de 330 milhões de usuários ativos mensais e uma ampla variedade de tópicos discutidos diariamente, o Twitter se tornou uma das principais fontes de informação e notícias atualizadas em todo o mundo (COSSU; DUGUé; LABATUT, 2015).

Como dito acima, o Twitter é uma valiosa fonte de dados, contando com dados atualizados, diversos e de fácil acesso. O Twitter se torna uma excelente escolha para analisar tendências, sentimentos, comportamento de usuários e muito mais. Desta forma, a empresa fornece uma Interface de Programação de Aplicativos, também conhecida como Application Programming Interface (API), para o compartilhamento de informações e funcionalidades de maneira padronizada e segura.

Ao utilizar a API do Twitter é possível extrair os dados dos últimos sete dias baseando-se em filtros pré-definidos, extrair *tweets* em tempo real seja de forma aleatória, informações do usuário como número de seguidores, quantidade de *tweets* e outros. Essas e outras particularidades são entregues por diferentes *endpoints* fornecidos pela plataforma. Um *endpoint*, é uma URL com métodos de solicitação, parâmetros e possíveis respostas. Cada *endpoint* é projetado para realizar operações

específicas e fornecer acesso a um conjunto particular de dados. Desta forma, quanto maior for a necessidade de adicionar informações sobre os dados coletados, maior será a utilização dos *endpoints* para enriquecer a base de dados. Essas informações são baseadas na documentação da plataforma até junho de 2023, e caso haja alguma alteração na documentação, essa pesquisa assegura a precisão das informações até essa data (TWITTER, 2023b).

Para a extração de *tweets* foi necessário utilizar os recursos de filtragem nos *endpoints* da API. A filtragem é utilizada para capturar *tweets* que possuam determinada palavra-chave, usuário, *hashtags*, localizações geográficas ou até mesmo idioma. Esse parâmetro possibilita garantir que os dados recebidos correspondam aos critérios definidos, simplificando até o processo de análise de dados na plataforma. Neste trabalho, utilizamos um total de quinhentos e nove termos diferentes para filtrar nossa base de dados, sendo quinhentos e dois relacionados a esportes e substâncias proibidas e sete sinônimos de outros termos já presentes na seleção inicial. Portanto, foram coletados 1.962.822 *tweets* entre os dias 21 de agosto de 2022 e 05 de janeiro de 2023.

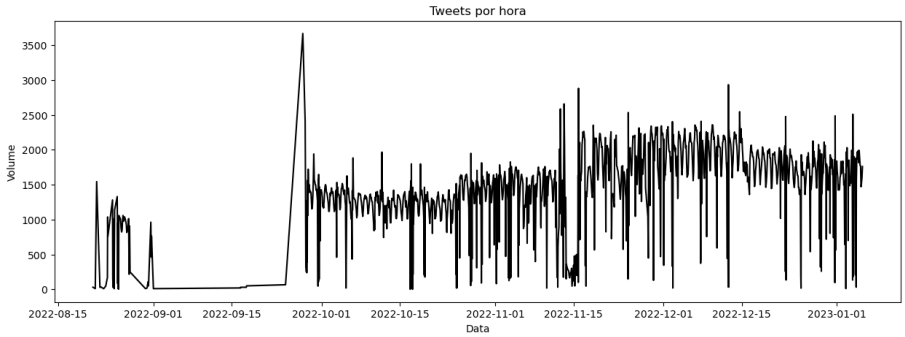

Figura 3.1: Quantidade de *tweets* entre o período 21/08/22 e 05/01/23 dividido por horas

Entretanto, a API do Twitter possui limitações significativas para garantir a estabilidade e a segurança da plataforma, bem como para preservar a integridade de suas estratégias de negócios relacionadas aos dados. Isso se deve ao fato de que os dados são extremamente valiosos e não seria lucrativo para a plataforma disponibilizá-los publicamente. A coleta de dados enfrenta várias restrições, incluindo restrições por taxas, que limitam a quantidade de solicitações que podem ser feitas

em um determinado intervalo de tempo. Outra limitação é a por *tweets*, que define a quantidade máxima de *tweets* que podem ser coletados por minuto, e essa quantidade varia de acordo com o nível de acesso da conta de desenvolvedor. No contexto deste trabalho, foram capturados, em média, 1.431 dados por hora, no entanto, existem exceções, como pode ser observado na figura 3.1.

Desta forma, é possível concluir que a coleta de dados por meio da API do Twitter não é capaz de capturar todas as informações relacionadas aos termos definidos, mas apenas uma amostra representativa. Esse fato deve ser levado em consideração na análise e interpretação dos resultados obtidos, uma vez que podem haver vieses e limitações decorrentes dessa seleção de amostras.

3.3 Modelagem de Dados

A modelagem de dados desempenha um papel fundamental no desenvolvimento de sistemas de banco de dados relacionais. Essa etapa envolve um processo cuidadoso de projetar a estrutura e as relações dos dados que serão armazenados, visando garantir a organização, eficiência e integridade dos dados. O principal objetivo é identificar as entidades, definir seus domínios, estabelecer restrições e identificar os relacionamentos entre elas.

No contexto mencionado, o modelo UML é amplamente empregado como uma ferramenta para representar de forma clara e visual os elementos estruturais identificados, proporcionando uma visão abrangente dos componentes do sistema de banco de dados. Por exemplo, na figura 3.2 e 3.3, é possível observar os resultados do mapeamento relacional representados como um conjunto de relações interconectadas por meio de restrições de chaves estrangeiras, oferecendo uma representação visual compreensível e informativa.

As classes principais são *Tweet*, *TwitterProfile* e *Terms*, enquanto os demais elementos estão diretamente relacionados a pelo menos uma dessas classes principais. No caso do *Tweet* ele engloba características como informações de localização, fonte, língua e a data de postagem dos *tweets*, além isso, possui contadores de *retweet*,

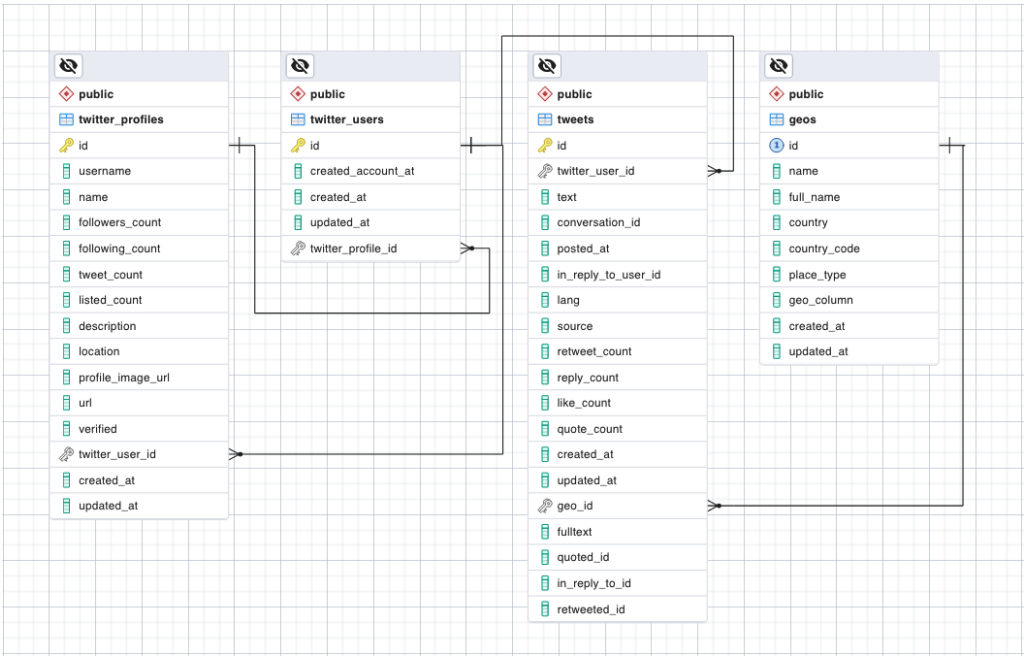

Figura 3.2: Modelagem dos dados relacionados a "Tweets" contidos na base de dados no modelo *UML*

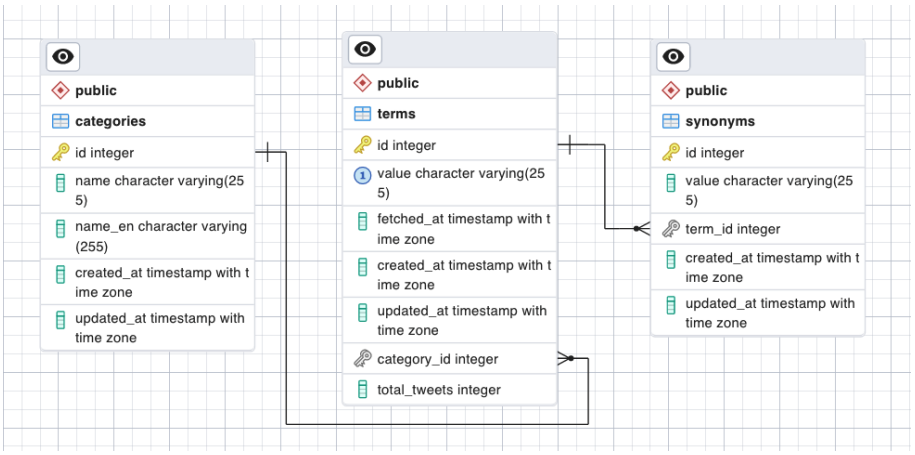

Figura 3.3: Modelagem dos dados relacionados a "Terms" contidos na base de dados no modelo *UML*

respostas, curtidas e comentários feitos aquele *tweet*.

O *TwitterProfile* contém informações do perfil, como nome, usuário, imagem de perfil, localização e descrição, além de indicar a verificação do usuário. Também registra o número de seguidores, perfis seguidos e *tweets* publicados. É importante ressaltar que há uma alta incidência de usuários duplicados devido ao processo de coleta, que atualiza periodicamente os dados, como quantidade de seguidores e *tweets*. Já o *TwitterUser* concentra-se principalmente em informações relacionadas à criação

da conta.

O *Terms* está diretamente relacionado à filtragem de *tweets*, fornecendo informações sobre a categoria na qual cada termo se encaixa, como esportes, eventos, substâncias proibidas e outros. Além do mais, são registradas a quantidade de *tweets* associados a cada termo e a data e hora em que o termo foi buscado. Também temos a classe de *Synonym*, que foi criada para adicionar sinônimos aos termos existentes, com o objetivo de capturá-los de forma mais organizada durante a extração de dados. Essa abordagem permite uma filtragem mais precisa e uma extração de informações mais abrangente, facilitando a análise dos *tweets* relacionados aos diferentes termos e categorias.

Capítulo 4

Análise dos Dados

Neste capítulo, é realizada uma análise quantitativa e qualitativa da base de dados, com maior foco nas substâncias proibidas e nos *tweets* dos usuários.

4.1 Análise Quantitativa

Neste tópico, realizamos uma análise minuciosa das tabelas, examinando predominantemente cada uma delas de forma individual, como já apresentadas na modelagem de dados 3.3. O objetivo principal foi extrair informações relevantes relacionadas às colunas principais, proporcionando uma compreensão mais aprofundada dos dados em questão.

4.1.1 Termos, Categorias e Sinônimos

Todos os termos usados como critérios de filtragem para coletar dados estão divididos em quatro categorias: substâncias proibidas, esportes e eventos, identificados pelos números 1, 2 e 3 respectivamente. Dessa forma, foi possível constatar que a base de dados contém 366 termos com *tweets*, considerando que nem todos esses termos possuem dados.

A categoria com o maior número de *tweets* é a de esportes, totalizando 871.734 *tweets*. Em segundo lugar, temos a categoria de substâncias proibidas com 801.573

tweets. Em terceiro lugar, encontramos a categoria de eventos com 17.728 *tweets* e que possuí apenas o termo: *Jogos Sul-americanos 2022*.

Na categoria de esportes, conforme pode ser visualizado na figura 4.1, o futebol é o termo com o maior número de *tweets*, o que é facilmente compreensível devido à sua enorme popularidade. Em seguida, temos tênis e basquete, também esportes amplamente populares, bem como outros esportes da lista, que também são conhecidos.

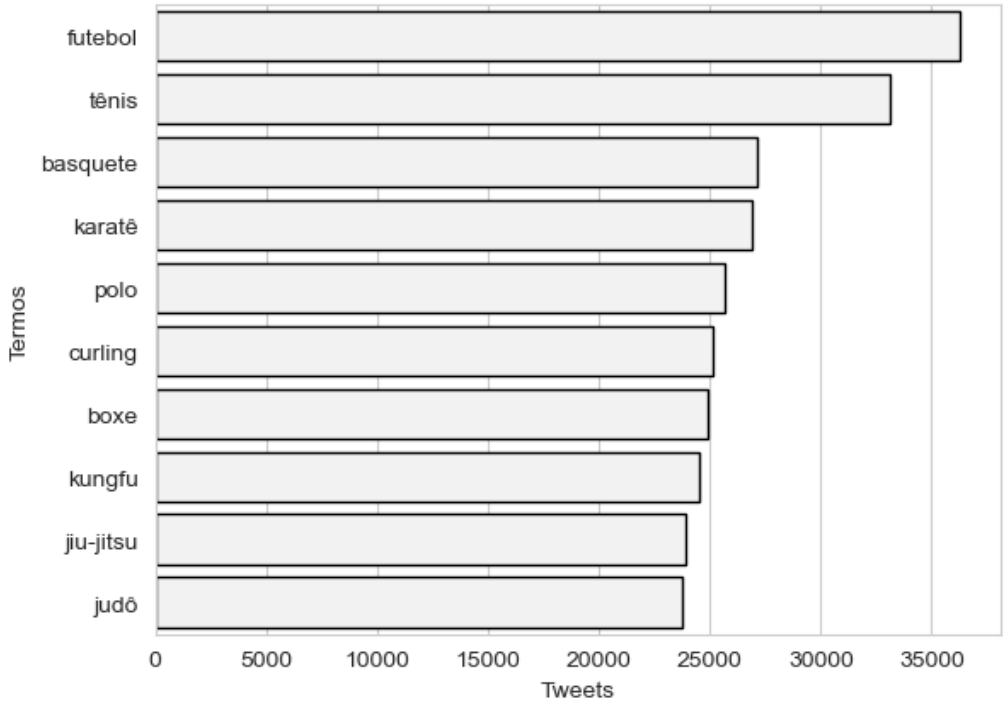

Figura 4.1: *Tweets* relacionados aos dez esportes mais frequentes na base de dados.

Substâncias proibidas são também uma categoria de extrema importância. Ao observar o gráfico 4.2, podemos identificar 4 substâncias relacionadas à *Cannabis Sativas*. O *CBD*, também conhecido como *Canabidiol*, é dos compostos químicos encontrados na *Cannabis*. Além disso, temos *marijuana* e *maconha*, que são os termos usado para se referir ao uso recreativo da droga e a *cannabis* em si que é o nome científico da planta.

Outras estimulantes que também apareceram no gráfico, Selective Androgen Receptor Modulator (SARM), ou modulador seletivo de receptor de androgênio, em português. Essa substância é um agente anabolizante, que vem sendo utilizado

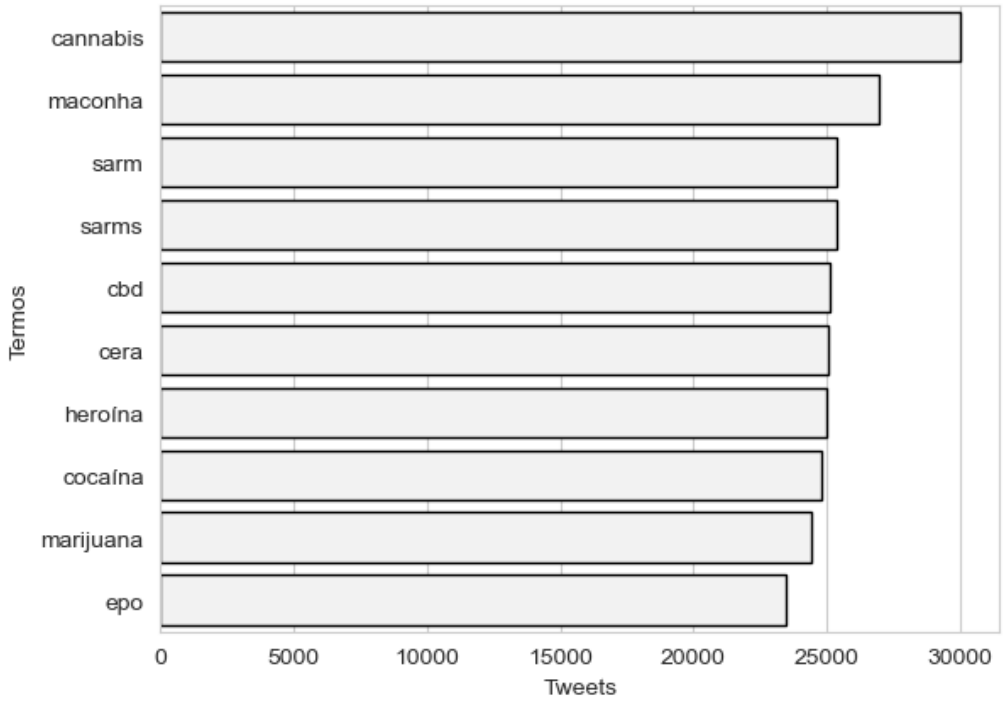

Figura 4.2: *Tweets* relacionados às dez substâncias proibidas mais frequentes na base de dados.

como uma melhor alternativa para o ganho de massa, mas que ainda está em fase de estudos sobre os efeitos a longo prazo (THEVIS; SCHäNZER, 2018).

CERA é outra substância encontrada na análise, a qual é sintetizada a partir da Eritropoietinas (EPO), também presente na base de dados. Essas substâncias influenciam na capacidade aeróbica do atleta e aumentam sua resistência. No entanto, a palavra *cera* é muito comum no português e espanhol, possuindo inclusive o mesmo significado em ambos os idiomas. Portanto, é possível ressaltar que, embora tenha sido identificada uma alta quantidade de *tweets* relacionados a essa palavra, é muito provável que a maioria deles não esteja referindo-se à substância em foco.

Por fim, as palavras *heroína* e *cocaína* também apresentam quantidades consideráveis de *tweets* relacionados. São conhecidas por serem drogas recreativas devido aos seus efeitos eufóricos. No entanto, a palavra *heroína* entra no mesmo caso já citado anteriormente. Tanto em português como em espanhol, esta palavra representa o feminino de herói, portanto é possível considerar que boa parte dos dados sejam referentes também a esse sentido da palavra.

4.1.2 Tweets

Dos 1.962.822 *tweets* coletados inicialmente, armazenando uma média de 17.217 *tweets* por dia, tendo variado nos primeiros meses por motivos da experimentação de coleta, como apresentado na figura 4.3.

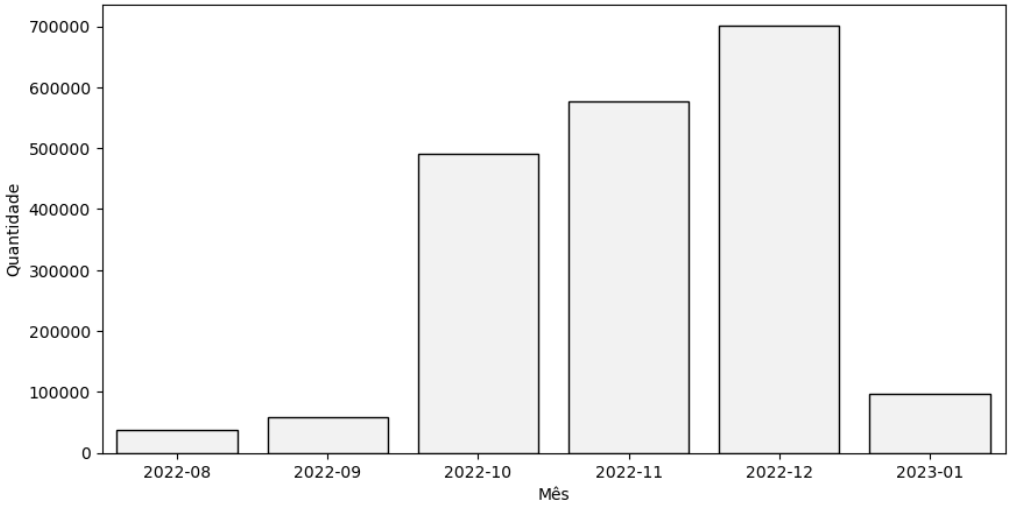

Figura 4.3: Quantidade de *tweets* coletado entre o período de agosto/2022 a janeiro/2023 dividido por meses.

Como apresentado na tabela 4.1, aproximadamente 42% dos *tweets* são *retweets*. Essa informação foi obtida analisando os textos que se iniciam com *RT*, que é uma característica da API da plataforma, de representar quando esses textos não são os originais. *Retweets*, são *tweets* que foram compartilhados por outras pessoas.

Métricas	Quantidade
Tweets	1.962.822
Retweets	822.520
Respostas	487.300
Citações	26.267
Línguas	71
Geolocalizações	25.217

Tabela 4.1: Quantidade de cada tipo de *tweets* encontrados na base de dados

Outros 25% dos *tweets* são respostas a outros usuários. Esse conhecimento é com base na coluna *in_reply_to_id*, que indica que o texto em questão foi uma resposta a outro *tweet* que possui determinado ID.

Além disso, observa-se mais 1,3% de *tweets* que citam outro *tweet*. Citações é

quando um usuário realiza um *retweet* adicionado de um comentário ou uma resposta. Esta dado é fornecido de acordo com a coluna *quoted_ id*, que informa o ID do *tweet* citado.

Os *retweets* transmitem a mensagem de que a pessoa que os compartilhou acredita, gosta ou concorda com a informação original do *tweet*, enquanto as citações e respostas podem expressar uma ideia a favor ou contrária ao conteúdo do *tweet* em questão.

Todas essas métricas mencionadas representam uma ampla disseminação das informações. Dos quase dois milhões de dados publicados para os seguidores, cerca de 68% alcançam um público ainda maior, graças a diversas formas de propagação da informação.

O Twitter detecta de forma automática a língua na qual aquele texto se encontra. Na base analisada, foram identificadas 71 línguas diferentes, sendo duas delas para sinalizar línguas não identificadas. As línguas mais frequentes são o espanhol, seguido pelo português, conforme ilustrado na figura 4.4. Isso ocorre porque grande parte dos termos estão em português e algumas dessas palavras também estão inseridas no espanhol.

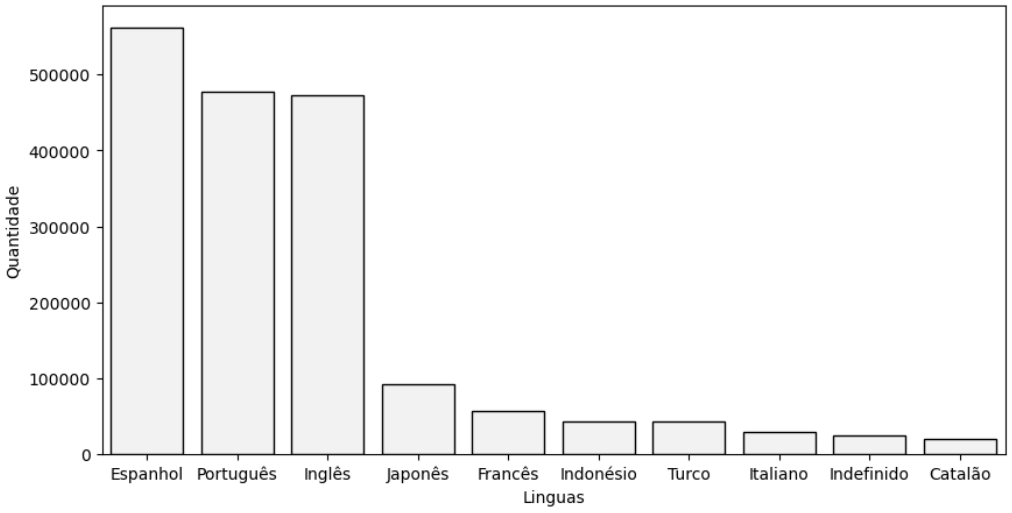

Figura 4.4: Dez idiomas mais frequentes na base de dados.

Dentre toda e extensão dos *tweets*, apenas 1,26% possuem identificações de localização. Isso significa que uma parcela relativamente pequena dos *tweets* está associada a dados geográficos específicos. Os usuários têm a opção de marcar seus

tweets com informações de localização, seja fornecendo uma localização específica ou selecionando locais próximos a eles.

4.1.3 Usuários

Os usuários desempenham um papel fundamental no estudo deste caso. Inicialmente, a base de dados continha dois usuários cuja criação da conta foi relatada como ocorrida há 53 anos, o que é claramente um erro, considerando que o Twitter possui apenas 17 anos de existência. A idade média das contas é de aproximadamente 8 anos e meio, impulsionada por uma grande criação de contas quando a rede social se popularizou, entre 2011 e 2013. Após esse período, o comportamento permaneceu relativamente estável até 2019, como evidenciado na figura 4.5.

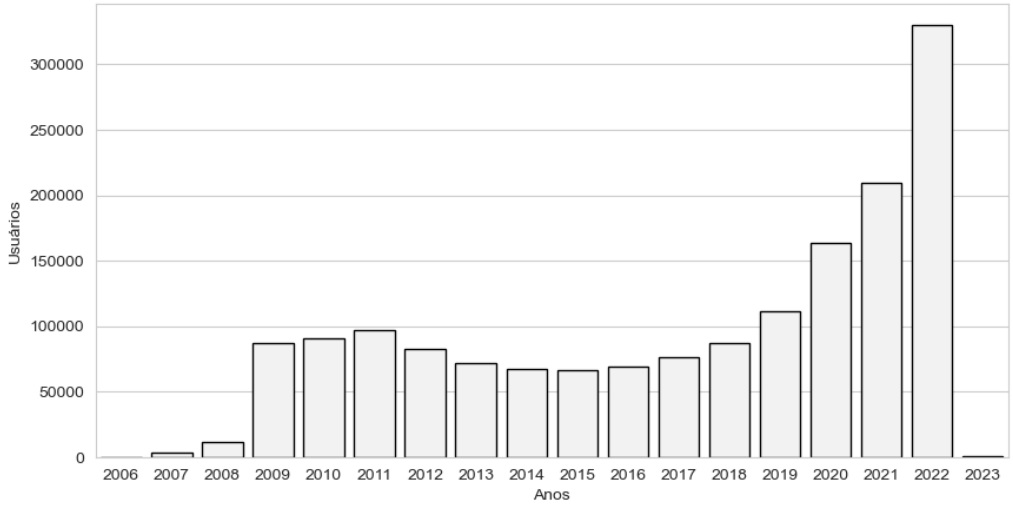

Figura 4.5: Quantidade de criação de novas contas de usuários por ano.

A partir de 2019, houve uma mudança significativa nesse padrão, como pode ser visto também no gráfico acumulado 4.6. Após 2018, o comportamento normal mudou para um ritmo mais acelerado, indicando uma mudança notável em como a plataforma foi adotada e utilizada pelos usuários a partir desse ponto.

4.1.4 Perfis do Twitter

Na base de perfis do Twitter, há uma grande quantidade de usuários duplicados, devido ao processo de coleta que atualiza os dados dos usuários, como a quantidade de

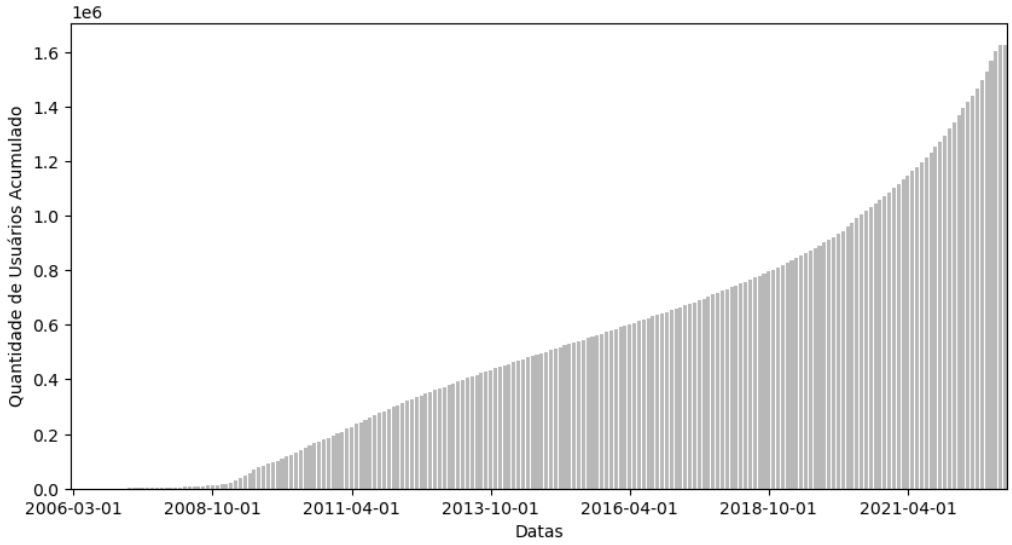

Figura 4.6: Quantidade de usuários acumulados por dia.

seguidores, *tweets* e outros atributos. Assim, temos 1.615.578 *usernames* diferentes, mas ainda é possível encontrar o mesmo usuário duplicado caso ele tenha sido coletado novamente após uma mudança de *username*, por exemplo.

O total de usuários únicos nesta base é de 1.607.669, e dentre esses usuários, 42.853 são contas verificadas pela empresa. As contas verificadas representam perfis autênticos, notáveis de interesse público, que o Twitter verificou de forma independente com base em determinados requisitos (TWITTER, 2023a). Esses dados são relevantes, especialmente considerando que foram coletados antes do lançamento do *Twitter Blue*, que oferece aos usuários a opção de assinar um plano que concede o selo azul da plataforma.

São coletados diversos tipos de usuários, devido à ampla variedade de termos e categorias utilizados. A base de dados inclui cantores, políticos, jogadores e outros perfis diversos, como evidenciado no gráfico 4.7 que apresenta os dez usuários com mais seguidores na base de dados. Liderando a lista, está Barack Obama com impressionantes 133.523.433 seguidores.

Outra análise interessante em relação à diversidade coletada é referente aos usuários com o maior número de *tweets* dentro da plataforma. A maioria desses usuários é representada por contas japonesas que se dedicam à divulgação de seus

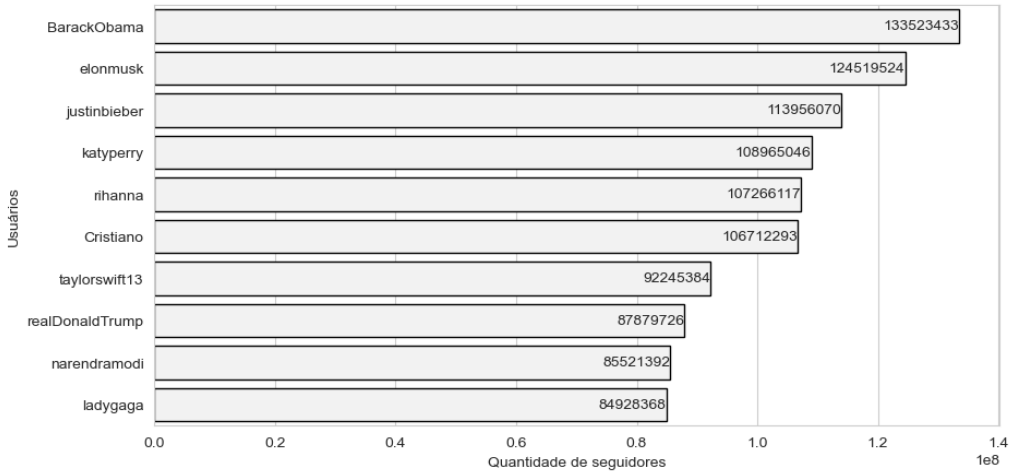

Figura 4.7: Dez usuários com mais seguidores da base de dados.

produtos e/ou lojas, como claramente demonstrado na figura 4.8. Contas notáveis como McDonald's, Subway e KFC são superadas por Lawson, uma popular loja de conveniência japonesa, que lidera com folga o número de *tweets* dentro do Twitter.

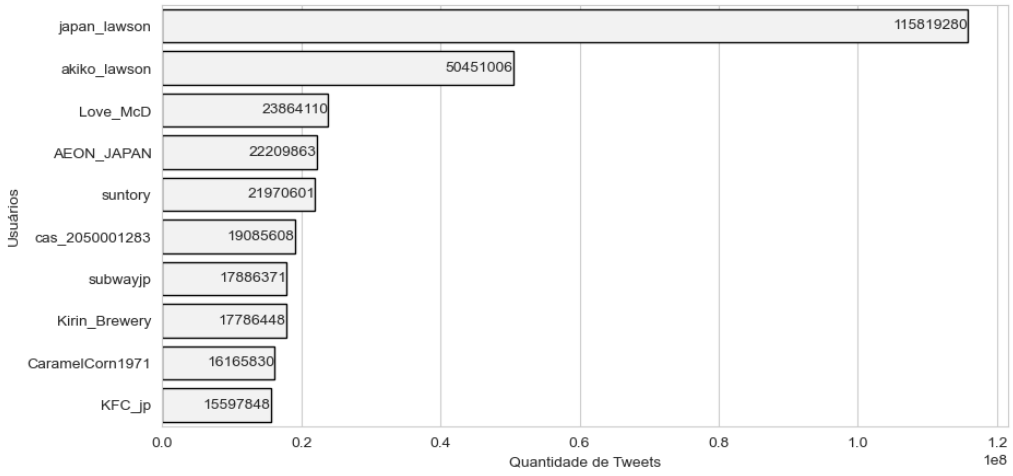

Figura 4.8: Dez usuários com mais *tweets* no Twitter.

4.1.5 Geolocalização

Em geolocalização, existem várias informações sobre a localização dos *tweets*, embora muitas delas sejam inseridas manualmente ou com base no idioma do dispositivo usado para realizar o *tweet*. Isso resulta em diferentes cidades e estados, mas a melhor maneira de quantificar essas informações é considerando o país de origem dos *tweets*.

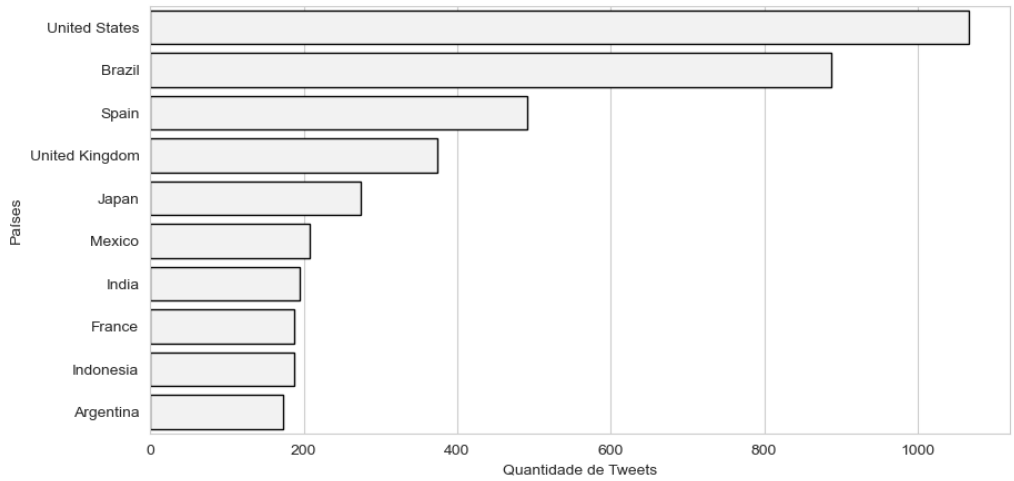

Figura 4.9: Os dez países com mais marcações de localização.

Dentre os dez países com mais marcações de localização nos *tweets*, encontramos os Estados Unidos em primeiro lugar, seguidos pelo Brasil em segundo e a Espanha em terceiro, conforme mostrado na figura 4.9. Esses resultados assemelham-se às análises dos idiomas mais frequentes, onde as três línguas com mais publicações são o Espanhol, Português e Inglês.

4.2 Análises Relacionais

Como mencionado na seção 3.3, a base de dados apresenta diversos relacionamentos. Nesta seção, concentraremos nossa análise nos relacionamentos que produzem resultados mais significativos para o desenvolvimento deste trabalho.

4.2.1 *Tweets* por usuários

Com base no figura 4.10 e na análise realizada, observa-se que aproximadamente 99% dos usuários possuem menos de 10 *tweets* coletados. Essa distribuição pode ser atribuída a diversas razões. Em primeiro lugar, muitos desses usuários podem ser meros espectadores ou entusiastas casuais, que ocasionalmente comentam ou mencionam os temas em seus perfis, mas não se engajam de forma frequente nessas discussões.

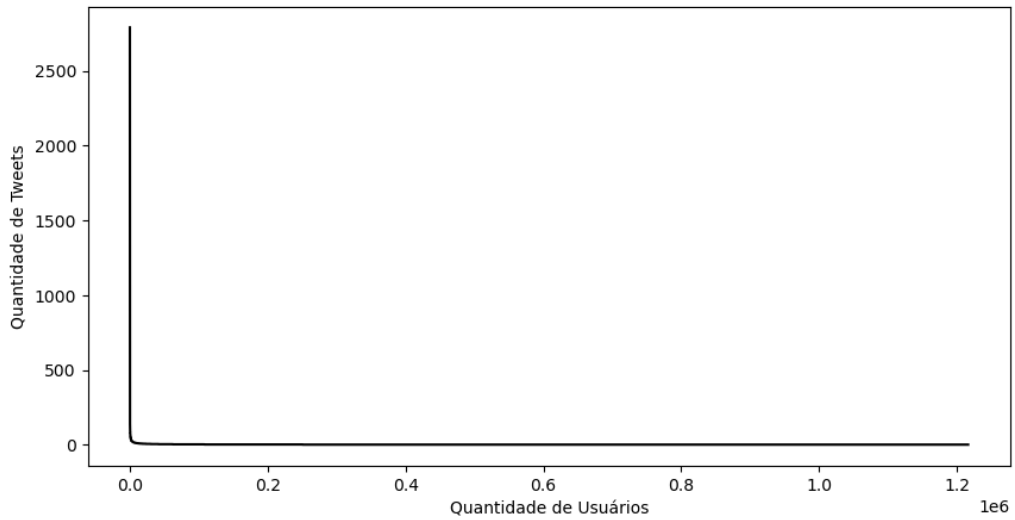

Figura 4.10: Quantidade de *Tweets* por quantidade de usuários

Além disso, a coleta de dados pode ter ocorrido em um período específico, no qual os usuários estavam menos ativos em relação a esses assuntos. Outro fator importante é que os termos relacionados a esportes e substâncias proibidas são bastante nichados, ou seja, referem-se a temas específicos e especializados. Isso significa que apenas um grupo restrito de usuários pode estar envolvido nesses assuntos de maneira mais frequente, enquanto a maioria dos usuários pode ocasionalmente mencionar esses termos em seus *tweets*.

4.2.2 Usuários com mais *Tweets* na base de dados

Tweets	Username
2789	OnlineshoppkE
1231	MMPconnect
1065	AusverkaufE
916	truth_hemp
847	kokushokucom
842	seedsmanrep
764	iembot_fgf
700	AkuCardi
634	MeridianoTV
606	futebol_bot

Tabela 4.2: Quantidade de *tweets* coletados por usuários

Uma análise relevante é entender quais usuários foram coletados mais vezes por

meio de seus *tweets*, para isso usamos as informações da tabela *Tweet* e também da *Twitter_Profiles* para conseguir ligar os *ID's* aos seus respectivos *Usernames*. Visualizando a tabela 4.2, podemos ver a discrepância entre a quantidade de *tweets* do primeiro colocado e do último. O usuário que aparece com maior quantidades de publicações é uma loja que possuía pouquíssimos seguidores, mas, e quase um milhão de *tweets* publicados. Este usuário juntamento com o *@AkuCardi* e *@seedsmanrep* estão suspensos pela plataforma na data desta pesquisa *24/07/2023*.

O segundo usuário destacado é um *blog* que retrata conteúdos médicos sobre a *Canabbis*. Além do mais, a conta *@truth_hemp*, que também aborda essa substância e comercializa produtos relacionados ao *CBD*. Esses exemplos evidenciam a predominância do termo *Cannabis* em nossas análises, demonstrando o interesse e engajamento dos usuários com esse tema específico.

A terceira conta é um *bot* responsável pela venda de diversos produtos. Junto a essa, também temos o *@iembot_fgf*, um outro *bot*, mas desta vez com foco em fornecer informações meteorológicas. Uma outra conta de destaque é a *@kokushokucom*, voltada para abordar temas relacionados ao cerealismo e veganismo. Considerando que veganos não consomem proteína animal, é plausível que a maior parte de suas publicações na nossa base de dados esteja relacionada a essa temática.

Por fim, os dois últimos usuários destacados são voltados para o mundo do esporte. Temos o perfil *@MeridianoTV*, dedicado a fornecer notícias e informações sobre diversos eventos esportivos. E, por fim, o usuário *@futebol_bot*, cujo aparentemente o foco é em apenas futebol, com atualizações sobre times e jogadores. Estes usuários trazem sentindo a sua alta frequência de publicações coletadas pois possuem uma alta probabilidade de serem pegos por termos que envolvem esportes que publicam em suas redes.

4.2.3 Dados Geoespaciais e Substâncias Proibidas

Conforme mencionado anteriormente, a base de dados abrange *tweets* provenientes de diversas localidades. Nesta etapa, realizamos uma análise para visualizar a distribuição geográfica dos *tweets* relacionados a substâncias proibidas, tanto dentro

do país como em escala global.

A fim de realizar uma análise mais aprofundada, foi necessário estabelecer relações entre cinco tabelas distintas: *tweets, geos, categories, terms e synonyms*. A filtragem dos *tweets* baseou-se no uso de *categories, terms e synonyms*, para identificar as substâncias proibidas. Por outro lado, a tabela *geos* foi empregada para selecionar o ID de localização dos *tweets* no *Twitter*.

Com base nisso, foi realizada uma breve limpeza nos dados para obter respostas mais objetivas. Em seguida, uma busca de texto completo foi conduzida para abranger todos os termos e sinônimos relacionados à categoria de substâncias proibidas, bem como seus respectivos identificadores de localização. Dito isso foram obtidos dois gráficos.

Na Figura 4.11, podemos observar que o tema das substâncias proibidas está distribuído por diversos locais, com concentração significativa na América do Sul, Europa e América do Norte, especialmente nas proximidades do México e Estados Unidos. Além disso, também são visíveis pontos na Índa, Japão, Indonésia e em outras regiões.

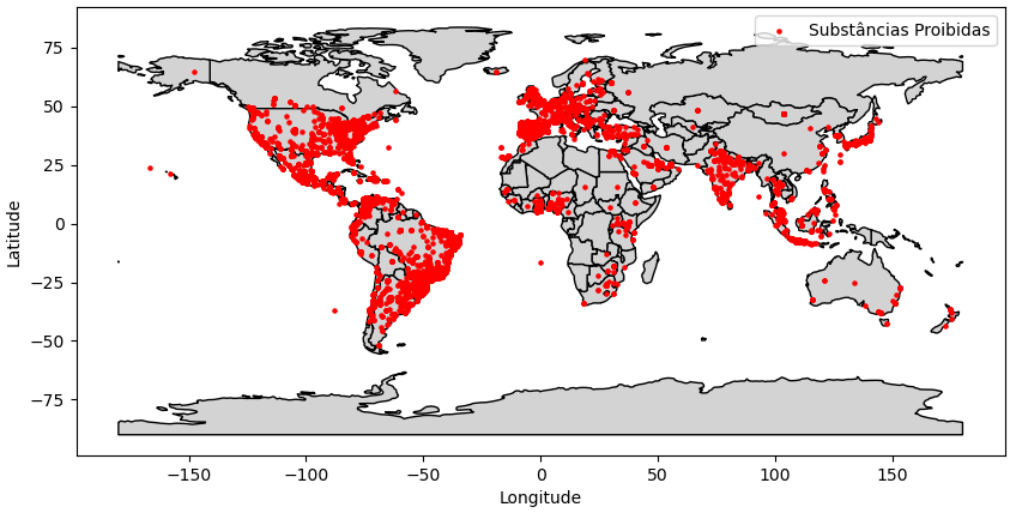

Figura 4.11: Mapa com pontos representando *tweets* sobre substâncias proibidas pelo mundo.

Adicionalmente, conduzimos uma análise mais específica em relação à geolocalização, focalizada dentro do país. Nessa abordagem, foram adicionados mais dois

filtros: a língua dos *tweets*, que deve ser em português, e o país, que deve ser o Brasil. Dessa maneira, conseguimos concentrar esta análise exclusivamente nesse território.

Observando a figura 4.12, é evidente uma concentração maior de *tweets* ao longo do litoral sul e sudeste do país, com diversos pontos espalhados por outras regiões. Essa distribuição pode ser influenciada pela densidade populacional mais significativa nas áreas costeiras e economicamente desenvolvidas.

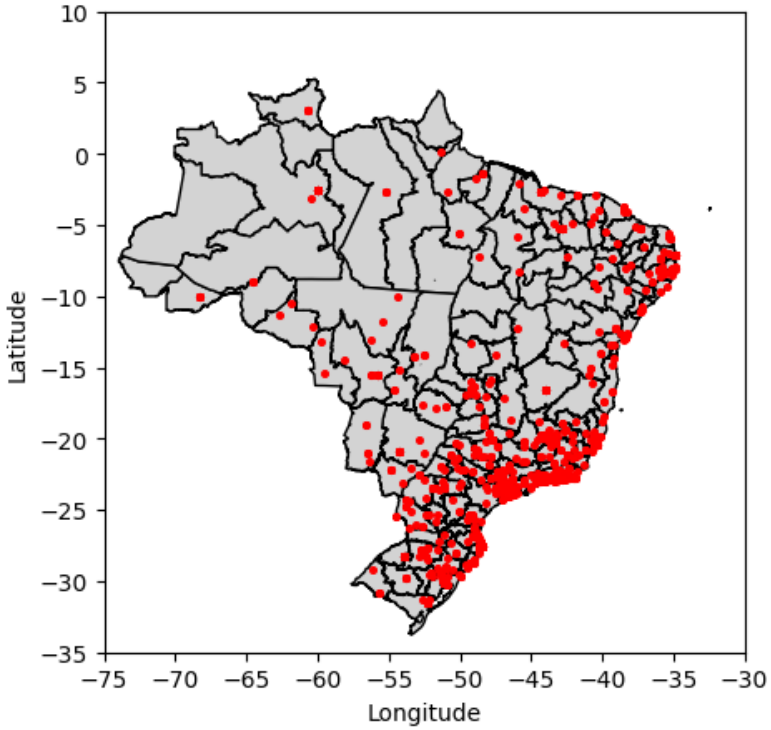

Figura 4.12: Mapa com pontos representando *tweets* sobre substâncias proibidas no Brasil.

Portanto, essa análise pode não ser tão relevante devido à distribuição populacional não homogênea do Brasil. Fatores como as condições ambientais impedem que determinadas áreas sejam habitadas, como em regiões florestais ou arenosas. Além disso as questões socioeconômicas influenciam muito a migração de pessoas para áreas mais específicas do país, resultando nesta desigualdades demográficas. Consequentemente, a concentração de *tweets* pode ser resultado dessas dinâmicas populacionais, em vez de indicar uma relação direta com o tema das substâncias proibidas.

4.3 Análises Qualitativa dos Agentes Dopantes

Dentre as centenas de substâncias proibidas contidas na base de dados para a extração dos *tweets*, apenas algumas delas são populares e de fato utilizadas pelo público geral em busca dos efeitos estéticos e desempenho físico. No Brasil, as substâncias anabolizantes mais frequentemente citadas e utilizadas são: Durateston, Deca-durabolim, Deposteron e Oxandrolona (IRIART; ANDRADE, 2002)(SANTOS et al., 2006). Portanto, é por meio dessas substâncias que as análises serão direcionadas, com o objetivo de compreender em quais contextos elas são mais frequentemente mencionadas e utilizadas.

Para as análises que serão realizadas a seguir, foram necessários alguns procedimentos de tratamento nos dados para que pudéssemos explorar as características reais de cada *tweet*. Assim, filtramos a base de dados *Tweets* apenas para incluir publicações em português que mencionavam a categoria completa de *Substâncias Proibidas*, bem como seus sinônimos, todos eles presentes na base de dados.

Em seguida, realizou-se a remoção das *stopwords* presentes nas publicações. *Stopwords* são palavras comuns que são removidas durante o processamento de texto em análises de linguagem natural, como mineração de texto. Essas palavras são consideradas irrelevantes para a análise, pois aparecem com muita frequência nos textos e não carregam informações significativas sobre o conteúdo. Alguns exemplos de *stopwords* incluem: *"o", "a", "um", "uma", "em", "para", "com", "por", "e", "ou"* e outros.

A primeira substância analisada foi *Deca-Durabolin*, conhecida por alguns sinônimos como *Deca, Nandrolona* ou apenas *Durabolin*. Na figura 4.13, podemos observar que outras substâncias dopantes são frequentemente mencionadas em conjunto com essa substância, incluindo *Durateston*, *Oxandrolona*, *Testosterona*, *Enantato* e outras. Ao analisar esses *tweets*, foi possível identificar usuários enaltecendo seus ganhos ou buscando por essas substâncias.

Durante o período da coleta, que coincidiu com o final do ano, observou-se que muitos *tweets* mencionavam expressões como "presentes de Natal" e "Papai Noel",

Figura 4.13: Nuvem de palavras relacionadas as substâncias *Deca-Durabolin* e *Nandrolona*.

resultando em uma frequência significativa da palavra "papai" nas análises. Quanto à alta frequência da palavra "garoto", identificou-se que essa palavra, juntamente com "Papai Noel", faz parte de uma letra de música que retrata o pedido de certas substâncias anabolizantes para o Natal, devido a terem sido "bons garotos". Abaixo apresentamos alguns exemplos de outras publicações:

"*meu triceps tá finalmente crescendo obrigado papai do céu (deca durabolin, durateston, oxandrolona, insulina e gh)*"

"*Alguém sabe quem vende Deca e durateston?...* "

"*12 anos de musculação? Ce ta precisando de nebido, durateston, deca, depostero, trembolona, metenolona, oxandrolona, testosterona etc.*"

Além dos *tweets* relacionados ao uso de agentes dopantes e à procura ilegal dessas substâncias, deparamo-nos com situações em que alguns usuários assumem o papel de julgar a aparência física de outros, sugerindo o uso de substâncias para obter resultados estéticos supostamente melhores.

Outra substância popular a ser analisada é a *Durateston*. Através da figura 4.14, podemos identificar outras substâncias proibidas, tais como *Winstrol*, também

conhecido como *Stanozol*. Temos a *Oxandrolona*, *Hemogenin*, *Dianabol* e ainda o apelido de *Trembolona*, conhecido como *Trembo*. Além disso, encontramos outras palavras relacionadas a este contexto de esteróides anabolizantes, como *ciclo, hormônio, dianabol, ampola, testosterona* e também ao contexto de treino e academias como *shape, creatina e academia*.

Figura 4.14: Nuvem de palavras relacionadas a substância *Durateston*.

Ao buscar por algumas dessas palavras relacionadas a *Durateston*, foi possível identificar alguns *tweets* relevantes de serem analizados.

"*pra que endocrino ? procura uma farmacia e mete durateston pra cima*"

"*Aproveita q o durateston vende por 15 pila na farmacia e n precisa de receita*"

"*Compro durateston barato sem receita* "

"*Envio Sedex ou loggi Zolpidem Stanozolol Durateston Anfepramona Oxandrolona Canabidiol ... TDAH, ansiedade, emagrecer,treino Chama no whatsapp.*"

Aparentemente, existe uma falsa sensação de segurança associada ao uso independente dessas substâncias, juntamente com uma facilidade de aquisição das mesmas, como é possível ver também no último *tweet* apresentando. A venda ilegal é outra

preocupação evidenciada nesta análise é uma questão que requer atenção especial e a implementação de medidas para coibir a comercialização dessas substâncias.

Outro *tweet* pertinente a esta análise destaca um efeito colateral já mencionado anteriormente neste estudo. O usuário relatou estar enfrentando episódios de agressividade e dificuldades para conciliar o sono.

> "*Vou dar uma parada com os meus ciclos com durateston. Não consigo dormir direito e venho estado muito agressiva.*"

"Oxandrolona" é uma substância bastante conhecida e popular entre as mulheres e como observado na figura 4.15, contém o termo "mulher" em evidencia, diferente das outras análises. Nessa investigação, identificamos outras substâncias, algumas semelhantes às mencionadas anteriormente e outras distintas. Na nuvem de palavras, destacaram-se agentes anabolizantes, como *Dianabol* e *Durabolin*, assim como medicamentos relacionados ao controle de peso, tais como *Saxenda*, *Sibutramina* e *Topiramato*. Além disso, também foram encontrados medicamentos para transtornos do sono e antidepressivos, como *Ritalina* e *Sertralina*, respectivamente.

Figura 4.15: Nuvem de palavras relacionadas a substância *Oxandrolona*.

A análise revelou que a ocorrência desses medicamentos é resultado de um grande número de *tweets* relacionados à venda dessas drogas, sendo também evidente a presença da palavra "Whatsapp" associada a essas publicações, como forma para contato, como exemplificado abaixo:

"*Durateston, Deposteron, Oxandrolona, Ritalina, Venvanse, Stavigile, Concerta, Metilfenidato, Lisdexanfetamina, Codeína, Codein, Modafinil, Clonazepam, Diazepam, Alprazolam, Zolpidem, Rivotril, Midazolam, Citotec, Ozempic, Saxenda, Victoza SEM RECEITA https://t.co/*****.*[1]"

Outras palavras que também apresentaram alta frequência foram *video, tiktok e sinal*. Essas três ocorrências se mostraram recorrentes em diversos contextos. Foram identificados *tweets* nos quais os usuários expressavam suas queixas em relação às postagens contidas na plataforma de vídeo *TikTok*. Abaixo, estão alguns exemplos dessas ocorrências:

"*o tiktok romantizando o uso de oxandrolona estou em pânico.*"

"*A moda do tiktok agora é exaltar oxandrolona como se fosse milagre.*"

"*Tá aparecendo muito video de oxandrolona pra mim no tiktok, acho q é um sinal.*"

Além disso, foram encontrados outros *tweets* relacionados à "Oxandrolona". Um deles expressava a insatisfação de um usuário com a dificuldade de aumentar sua massa muscular de forma natural. O outro *tweet* apresentava uma indicação completa de uso para mulheres, contendo informações detalhadas sobre a dosagem em miligramas, duração do ciclo em semanas e próximos passos. Essa atitude é extremamente irresponsável e perigosa, principalmente considerando que essas informações estavam sendo divulgadas em uma rede social como o Twitter.

"*Mulheres, doses de w-z*[2]*mg de Oxandrolona por dia durante y-z*[3] *semanas (não passar disso), pós o ciclo, Novaldex, Clomid e HCG pra mais eficiência da TPC.*"

"*Mais um dia com essa vontade de tomar oxandrolona muitoooo difícil crescer natural vei*"

[1]Link contendo o endereço de Whatsapp do vendedor ocultado por motivos de segurança
[2]Devido à natureza sensível das informações, as quantidades em miligramas foram omitidas
[3]Para preservar a privacidade, as informações referentes ao número de semanas foram omitidas

As palavras *faculdade, silicone* e *mokba* apresentaram uma quantidade considerável de ocorrências na base de dados e estão relacionadas a um discurso de ódio direcionado a uma mulher específica. Nesse contexto, os usuários questionavam o sucesso dela, sugerindo que ele teria vindo por outros meios, que a não faculdade.

"*Foi a faculdade sim, não foi o silicone e a oxandrolona malhando na Mokba e o ex namorado famoso emprestando fama pro canal dela que fez um monte de nerd **********[4] deixar ela famosa e sair do buraco que ela morava no Ceasa*"

Por fim, temos outra substância popular, a *Deposteron*. No entanto, suas análises revelaram um cenário distinto em comparação às outras. Na Figura 4.16, as palavras com maior frequência são *preço*, *aumento*, *real* e *farmácia*. Isso ocorre devido ao aumento significativo no preço da *Deposteron* no último ano. Segundo informações obtidas (GLOBO, 2023), tal aumento foi atribuído à defasagem de preços dessa substância.

Além disso, a reportagem e a análise dos dados apresentou que o público transgênero foi bastante afetado por essa alta de valores, devido à importância da hormonização para esta comunidade. A hormonização é parte do percurso fundamental para a saúde e o bem-estar dos homens trans, pois garante ganhos sociais diversos, como a melhoria da qualidade de vida e a promoção da autoestima (SOUSA; IRIART, 2018).

A diferença entre terapia hormonal, hormonização e o uso deliberado desses agentes para fins estéticos é significativa. A terapia hormonal é um tratamento médico regulamentado, corrigindo desequilíbrios hormonais e melhorando a saúde de pacientes. A hormonização é realizada com acompanhamento médico por indivíduos transgêneros, visando adequar os níveis hormonais ao gênero com o qual se identificam. Em contrapartida, o uso de hormônios para fins estéticos, como o ganho de massa muscular, é uma prática não terapêutica e arriscada à saúde, quando feita sem supervisão médica.

[4]Termo obsceno

Figura 4.16: Nuvem de palavras relacionadas a substância *Deposteron*.

Abaixo é possível observar as reações dos usuários mediantes aos altos preços da substância e outras formas de conseguir adquirir a droga:

"*Chorei ao ver o preço da deposteron hj tá cada vez mais difícil ser trans no br*"

"*Antes a deposteron tava em falta, agora que tá 250,00 tem em todas as farmácias. Incrível.*"

"*Baixem o preço do Deposteron pelo amor de deus trans.*"

"*O que estão usando no lugar da Deposteron com a alta dos preços?*"

"*Médico: aplique deposteron e treine todos os dias segunda/sexta, mas é barato o remédio, no máximo 80 reais Realidade: farmácias cobrando 200+*"

"*Tb tô nessa. 2 semana já q tô x^5mg semanal de enantato de trembo + y^6mg mensal de Deposteron, mas ainda não me deu raiva não, só tô down deprê* "

"*galera trans que precisar de t, só chamar na dm.. envio pra todo lugar do brasil* "

[5]Devido à natureza sensível das informações, as quantidades em miligramas foram omitidas
[6]Devido à natureza sensível das informações, as quantidades em miligramas foram omitidas

"*pela primeira vez em 3 anos e 9 meses irei comprar deposteron direto na farmacia*"

Após essas análises, foram identificadas semelhanças em relação a outras substâncias já analisadas, tais como a venda não autorizada e a ocorrência de alguns efeitos colaterais comuns. No entanto, um aspecto que se destacou foi o contexto do aumento de preços, que se mostrou mais presente nesta substância em particular.

Capítulo 5

Conclusão

Neste capítulo, serão apresentadas as considerações finais acerca deste trabalho, bem como as limitações associadas a ele e as perspectivas para possíveis trabalhos futuros.

5.1 Considerações finais

Com base nas análises realizadas, foi possível identificar uma alta quantidade de *tweets* com teor cômico, nos quais os usuários demonstravam interesse em usar ou faziam referências a músicas que incentivam o uso de substâncias dopantes. Essa atitude tende a normalizar o uso desses agentes, muitas vezes de forma irresponsável, independentemente de ser uma verdade compartilhada ou não.

Uma das constatações comprovadas por meio das análises foi a venda ilegal não apenas de esteróides anabolizantes, mas também de medicamentos para emagrecimento, depressão, problemas com o sono e outros. Os *tweets* relacionados a essas práticas são feitos de maneira explícita, expondo números de telefone e todas as substâncias vendidas pelos usuários. É preocupante notar que esses vendedores parecem não temer as consequências legais, uma vez que agem abertamente na plataforma.

Outra questão relevante a ser abordada é a presença de julgamentos e discursos de

ódio encontrados durante a análise. Identificamos usuários insatisfeitos com o corpo de outras pessoas, sugerindo que elas utilizassem compostos proibidos para alcançar resultados desejados. Além disso, foram encontrados discursos odiosos em relação a mulheres e pessoas transgênero, o que nos traz um alerta para essas manifestações de preconceito.

Uma constatação relevante foi a grande quantidade de usuários exibindo seus resultados após o uso de substâncias dopantes e a facilidade com que obtêm essas substâncias sem receitas médicas. Essa atitude é completamente irresponsável, uma vez que outros usuários relataram efeitos colaterais preocupantes relacionados ao uso dessas substâncias. Esses comportamentos destacam a necessidade urgente de medidas para combater a venda ilegal e a romantização do uso desses agentes dopantes.

5.2 Limitações e trabalhos futuros

Uma das limitações encontradas neste trabalho foi em relação à API do Twitter, que impunha restrições à quantidade de *tweets* que poderiam ser coletados. Entretanto, com as atualizações recentes na plataforma, diversas formas de coletas foram reduzidas para os usuários não pagantes, inclusive a documentação usada neste trabalho para a utilização da API não está mais disponível. Esse cenário pode impactar futuras pesquisas que dependam da coleta de dados do Twitter, uma vez que o acesso às informações pode ser mais restrito e dificultar a análise de grandes volumes de dados.

Este trabalho foi bem-sucedido em implementar análises que proporcionaram uma compreensão sobre como a sociedade lida atualmente com algumas substâncias dopantes. No entanto, há espaço para aprimoramentos e desenvolvimentos futuros. Uma das melhorias possíveis seria a expansão da análise para incluir mais substâncias, uma vez que neste trabalho foram abordadas apenas quatro delas. Além disso, seria relevante investigar outros termos relacionados ao uso de substâncias dopantes para obter uma visão mais abrangente do panorama das discussões nas redes sociais.

Para uma análise mais aprofundada seria interessante uma quantidade maior de dados, então como trabalhos futuros é favorável continuar coletando esses dados e acompanhando a evolução do uso dos agentes dopantes ao longo do tempo, podendo fornecer padrões sazonais ou a mudança de comportamento relacionadas a essas práticas.

Outra sugestão para aprimorar este trabalho é realizar análises dos dados em outros países e em diferentes línguas, visando compreender como esse contexto é abordado globalmente e identificar possíveis diferenças nas políticas de controle e acesso a substâncias proibidas, bem como padrões de uso e a prevalência dessas práticas em diversos contextos culturais e socioeconômicos. Além disso, uma análise mais regional no país também pode ser enriquecedora, uma vez que a população se concentra naturalmente ao longo da costa do país. Investigar possíveis diferenças no uso de substâncias em diferentes regiões do mesmo país pode trazer resultados importantes sobre os fatores que influenciam essas práticas.

Utilizar essas análises e outras informações coletadas para embasar intervenções educacionais pode ser uma estratégia eficaz na conscientização de jovens adultos sobre os riscos e consequências associadas ao uso desses componentes. Ao fornecer dados concretos sobre os efeitos colaterais e impactos negativos à saúde, é possível desmistificar a ideia de que o uso dessas substâncias é inofensivo ou uma solução rápida para atingir determinados objetivos estéticos ou de desempenho. Com informações claras e objetivas, é possível desencorajar a normalização desse comportamento e incentivar escolhas mais saudáveis e responsáveis para o cuidado do corpo e da saúde.

Diante da evidente presença do mercado ilegal de substâncias proibidas, torna-se imprescindível investigar esses casos para que medidas legais adequadas sejam tomadas. Essa investigação não apenas visa coibir a comercialização dessas substâncias, mas também promover a segurança e o bem-estar da população. Nesse sentido, é fundamental que as próprias redes sociais assumam um papel proativo, assim como fazem ao combater discursos de ódio. Monitorar e combater a disseminação de substâncias ou conteúdos ilegais é uma responsabilidade compartilhada, que requer a colaboração tanto das plataformas digitais quanto das autoridades competentes.

Referências

AITH, F. M. A. Regulação antidoping e saúde pública: limites à exposição humana ao risco sanitário e a glória desportiva. *Revista de Saúde Pública*, v. 47(5), p. 1015–1018, 2013.

AMES, J.; SOUZA, D. Falsificação de medicamentos no brasil. *Revista de Saúde Pública*, v. 46(1), p. 154–159, 2012.

AZIZ, J. Social media and body issues in young adults:an empirical study on the influence of instagram use on body image andfatphobia in catalan university students. In: . [S.l.: s.n.], 2017.

BIRD, S. R. et al. Doping in sport and exercise: Anabolic, ergogenic, health and clinical issues. *Annals of Clinical Biochemistry*, The Association for clinical Biochemistry Laboratory Medicine, v. 53, n. 2, p. 196–221, 2016.

COSSU, J.-V.; DUGUé, N.; LABATUT, V. Detecting real-world influence through twitter. *2º European network intelligence conference.*, Institute of electrical and electronics engineers, v. 23, n. 9, p. 1355–1390, 2015.

GESTSDOTTIR S., K. H. S. H.; SIGFUSDOTTIR, I. D. Prevalence, mental health and substance use of anabolic steroid users: a population-based study on young individuals. *Scandinavian journal of public health*, v. 49, n. 5, p. 555–562, 2021.

GLOBO, T. *Aumento de mais de 400% no preço de remédio com hormônio prejudica tratamento de homens trans*. 2023. <https://encurtador.com.br/myAFP>. Acessado em 28/07/2023.

GOLDSCHMIDT, R.; PASSOS, E. *Data Mining - um Guia Prático*. [S.l.]: Elsevier, 2005.

HAN, J.; KAMBER, M. *Data Mining: Concepts and Techniques*. [S.l.]: Morgan Kaufmann, 2011.

HILKENS L., C. M.; WOERTMAN, L. Social media, body image and resistance training: Creating the perfect 'me' with dietary supplements, anabolic steroids and sarm's. *Sports Med - Open*, v. 7, n. 81, 2021.

IRIART, J. A. B.; ANDRADE, T. Musculação, uso de esteróides anabolizantes e percepção de risco entre jovens fisiculturistas de um bairro popular de salvador, bahia, brasil. *Cadernos De Saúde Pública*, v. 18(5), p. 1379–1387, 2002.

MANNILA, H. *Data mining: machine learning, statistics, and databases*. [S.l.]: Proceedings of 8th International Conference on Scientific and Statistical Data Base Management, 1996. 2-9 p.

MEDICINA, C. F. de. *RESOLUÇÃO CFM nº 2.333/2023*. 2023. <https://sistemas.cfm.org.br/normas/visualizar/resolucoes/BR/2023/2333>. Acessado em 28/07/2023.

MORAES1, T. P. B. de. Anabolizantes nas buscas da web: Um estudo sobre o interesse sazonal por esteroides anabolizantes no brasil. *Revista Jurídica Luso-Brasileira*, CIDP, v. 1, p. 1979–2007, 2015.

NILSSON S., S. F. M. B. B. A.; ALLEBECK, P. Attitudes and behaviors with regards to androgenic anabolic steroids among male adolescents in a county of sweden. *Subst Use Misuse*, v. 40, n. 1, p. 1–12, 2005.

REZENDE. JB PUGLIESI., E. M. S.; PAULA, M. Mineração de dados. *Sistemas inteligentes: fundamentos e aplicações*, v. 1, p. 307–335, 2003.

SANTOS, A. et al. Anabolizantes: conceitos segundo praticantes de musculação em aracaju (se). *Revista de Saúde Pública*, v. 11(2), p. 371–380, 2006.

SJöQVIST, F.; GARLE, M.; RANE, A. Use of doping agents, particularly anabolic steroids, in sports and society. *Molecular and Cellular Endocrinology*, Lancet, v. 371,9627, p. 1872–1882, 2008.

SOUSA, D.; IRIART, J. "viver dignamente": necessidades e demandas de saúde de homens trans em salvador, bahia, brasil. *Cadernos de Saúde Pública [online]*, v. 34, p. 1678–4464, 2018.

TAN, P.-N. *Introduction to Data Mining*. [S.l.]: Pearson, 2013.

THEVIS, M.; SCHäNZER, W. Detection of sarms in doping control analysis. *Molecular and Cellular Endocrinology*, Elsevier, v. 464, p. 34–45, 2018.

TOKISH J. M., K. M. S.; HAWKINS, R. J. Ergogenic aids: a review of basic science, performance, side effects, and status in sports. *The American journal of sports medicine*, v. 32, n. 6, p. 1543–1553, 2004.

TRAJANO F, C. *Doping no Esporte: A Bioética como Ponte para o Esporte Seguro*. 2023.

TWITTER. *Como receber o selo azul no Twitter*. 2023. <https://help.twitter.com/pt/managing-your-account/about-twitter-verified-accounts>. Acessado em 23/07/2023.

TWITTER. *Twitter API Documentation*. 2023. <https://developer.twitter.com/en/products/twitter-api/academic-research/product-details>. Acessado em 16/06/2023.

WADA. *Código Mundial Antidopagem*. 2021. <https://www.wada-ama.org/sites/default/files/resources/files/wada_2021_code_november_2019_v._wada_2021_code_june_2020_final_-_english.pdf>. Acessado em 03/07/2023.

WOODIE, A. *Big Growth Forecasted for Big Data*. 2022. <https://www.datanami.com/2022/01/11/big-growth-forecasted-for-big-data/>. Acessado em 15/07/2023.