
**Universidade Federal de Ouro Preto**
**Instituto de Ciências Exatas e Aplicadas**
**Departamento de Computação e Sistemas**

# Monitoramento e análise de dados demográficos coletados a partir de uma rede social

**Innan Plínio Rangel Amorim**

## TRABALHO DE CONCLUSÃO DE CURSO

ORIENTAÇÃO:
Filipe Nunes Ribeiro

**Novembro, 2022**
**João Monlevade–MG**

**Innan Plínio Rangel Amorim**

# Monitoramento e análise de dados demográficos coletados a partir de uma rede social

Orientador: Filipe Nunes Ribeiro

Monografia apresentada ao curso de Engenharia da Computação do Instituto de Ciências Exatas e Aplicadas, da Universidade Federal de Ouro Preto, como requisito parcial para aprovação na Disciplina "Trabalho de Conclusão de Curso II".

**Universidade Federal de Ouro Preto**

**João Monlevade**

**Novembro de 2022**

MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DE OURO PRETO
REITORIA
INSTITUTO DE CIENCIAS EXATAS E APLICADAS
DEPARTAMENTO DE COMPUTACAO E SISTEMAS

# FOLHA DE APROVAÇÃO

**Innan Plínio Rangel Amorim**

**Monitoramento e análise de dados demográficos coletados a partir de uma rede social**

Monografia apresentada ao Curso de Engenharia de Computação da Universidade Federal de Ouro Preto como requisito parcial para obtenção do título de Bacharel em Engenharia de Computação

Aprovada em 04 de novembro de 2022

Membros da banca

Doutor - Filipe Nunes Ribeiro - Orientador(a) (Universidade Federal de Ouro Preto)
Doutora - Helen de Cássia Sousa da Costa Lima - (Universidade Federal de Ouro Preto)
Doutora - Gilda Aparecida de Assis - (Universidade Federal de Ouro Preto)

Filipe Nunes Ribeiro, orientador do trabalho, aprovou a versão final e autorizou seu depósito na Biblioteca Digital de Trabalhos de Conclusão de Curso da UFOP em 11/11/2022

Documento assinado eletronicamente por **Filipe Nunes Ribeiro**, **PROFESSOR DE MAGISTERIO SUPERIOR**, em 11/11/2022, às 16:59, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

A autenticidade deste documento pode ser conferida no site http://sei.ufop.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0 , informando o código verificador **0426359** e o código CRC **9A8EA87C**.

**Referência:** Caso responda este documento, indicar expressamente o Processo nº 23109.015290/2022-05 SEI nº 0426359

R. Diogo de Vasconcelos, 122, - Bairro Pilar Ouro Preto/MG, CEP 35402-163
Telefone: (31)3808-0819 - www.ufop.br

*A minha amada mãe, Mônica Fátima Garuzzi Rangel (Em Memória).*

# Resumo

O surgimento dos portais de mídia social possibilitaram uma abundância de conteúdo gerado pelo usuário na web. Além disso, a popularidade de sites como Twitter, MySpace e Facebook tem crescido ininterruptamente. Com os Censos necessitando de muito tempo, esforços e recursos em grande quantidade, países em desenvolvimento podem enfrentar dificuldade em sua realização. Uma alternativa é o uso das plataformas de publicidades das redes sociais, que vem se mostrando cada vez mais tecnológicas e precisas, como forma de substituir práticas antigas, como por exemplo a realização do Censo de forma tradicional. Acredita-se que, apesar de esforços em estudos para inferir demografia a partir de dados online, ainda há espaço para avanços. Este trabalho tem como objetivo principal a coleta de dados demográficos a partir da plataforma de publicidade do Facebook, o monitoramento da variação desses dados ao longo do tempo, e a comparação com dados reais. A metodologia utilizada será aplicada no contexto do EUA e Brasil. De forma similar ao censo demográficos providos por fontes oficiais, como o US Census Bureal e o IBGE, com uma periodicidade maior, temos em vista complementar as estatísticas já oferecidas, bem como diminuição com custo e tempo consumido para obtenção de tais dados. Um dos objetivos principais é a comparação desses dados com relatórios das fontes referidas, com a finalidade de identificar o quão confiáveis são os dados extraídos das redes sociais.

**Palavras-chaves**: Redes sociais. Publicidade. Censo demográfico.

# Abstract

The emergence of social media portals has enabled an abundance of user-generated content on the web. Furthermore, the popularity of sites like Twitter, MySpace and Facebook has been growing nonstop. With Censuses requiring a lot of time, effort and resources in large quantities, developing countries may face difficulties in carrying them out. An alternative is to use social media advertising platforms, which have been increasingly technological and accurate, as a way of replacing old practices, such as carrying out the Census in a traditional way. It is believed that, despite efforts in studies to infer demographics from online data, there is still room for progress. The main objective of this work is to collect demographic data from the Facebook advertising platform, monitor the variation of this data over time, and compare it with real data. The methodology used will be applied in the context of the USA and Brazil. Similar to the demographic census provided by official sources, such as the US Census Bureau and the IBGE, with greater frequency, we aim to complement the statistics already offered, as well as reduce the cost and time consumed to obtain such data. One of the main objectives is to compare these data with reports from the aforementioned sources, in order to identify how reliable the data extracted from social networks are.

**Key-words**: Social networks. Advertising. Demographic census.

# Lista de ilustrações

# Lista de tabelas

# Sumário

# 1 Introdução

## 1.1 Contextualização e definição do problema abordado

Um censo ou recenseamento de população pode ser descrito como coleta, agrupamento e publicação de dados demográficos, econômicos e sociais que são referentes a um determinado período de tempo, aos habitantes de um país ou território (NATIONS, 2017). Podem ser utilizados politicamente, para determinar onde os investimentos precisam ser aplicados, porém, apesar de necessários, podem ter um alto custo financeiro e de tempo(CENSUS, 2017).

No Brasil, o primeiro recenseamento foi efetuado em 1808, com a intenção de atender exclusivamente a interesses militares, a respeito de recrutamento para Forças Armadas. Porém, estima-se que os resultados obtidos ficam abaixo do esperado, talvez por uma espécie de mecanismo de defesa contra operações censitárias ou talvez por causa de seus objetivos (IBGE, 2022a).

A quesito de registro histórico, é reconhecido como sendo o primeiro censo realizado no país, o denominado Censo Geral do Império, realizado em 1879. A partir disso, houveram várias mudanças no processo como um todo, assim como vários outros recenseamentos executados durante os anos (com uma certa dubitabilidade em seus resultados, devida a baixa aderência da população), bem como uma mudança em seus interesses, podendo assim então se estabelecer uma periodicidade decenal. Temos o início do recenseamento decenal no ano de 1890, falhando apenas nos anos 1910 e 1930 em que foram suspensos e 1990 em que a operação foi transferida para o ano seguinte (IBGE, 2022a).

Figura 1 mostra o nível populacional aferido em todos os censos já executados para o Brasil.

Os censos realizados pelo mundo inteiro são uma fonte chave de dados que guiam investimentos governamentais e políticas públicas. Realizados durante séculos, os censos são altamente necessários na sociedade moderna, além de serem cruciais para definição da prioridade de investimentos para educação, infraestrutura e outras políticas públicas do país. Entretanto, apesar de sua importância, o custo e o tempo consumido para obter esses dados são bem altos. Além do fato de não possuírem uma atualização tão constante quanto o desejado, tendo em vista que os censos são decenais.

Apesar de sua importância, o custo e o tempo consumido para obter esses dados são bem altos. Em 2010 esse valor foi de R$1,667 bilhão (IBGE, 2022c) e em 2020 o valor estimado chegava a um total de R$3,4 bilhões (ACKER, 2019b). Nos Estados Unidos esse valor chega a $15 bilhões de dólares (GAO, 2021). A maior parcela desses valores,

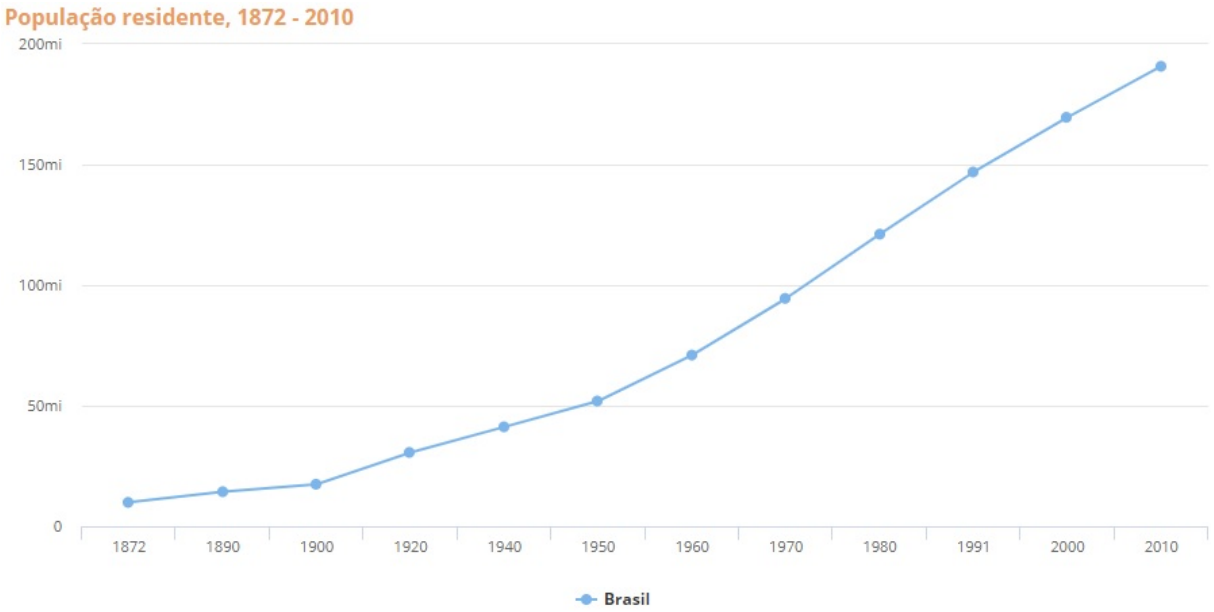



Figura 1 – Censo demográfico no Brasil ao longo dos anos.
Fonte: Reprodução IBGE (2010c)

é destinada a contratação de cerca de 240 mil funcionários temporários (IBGE, 2022b), e essa é uma das características que impedem a redução significativa de orçamentos da pesquisa.

Atualmente, grande parte da população brasileira possui acesso à internet. Em simulações realizadas pelo IBGE, no qual se buscava fazer com que a amostra da população respondesse o questionário através da internet para não haver a necessidade da visita do recenseador, apenas uma pequena parcela da amostra respondeu através dessa opção (ACKER, 2019a).

Em um espaço de tempo de 10 anos, pode ser que aconteçam mudanças significativas nesses dados, ainda mais se considerarmos um país com tamanha extensão territorial como o Brasil, o que faria com que todo um planejamento feito anteriormente se torne inválido. Por outro lado, as redes sociais inferem uma série de informações privadas de seus usuários com base em suas postagens, likes, etc. Tais informações são muito importantes para prover uma rica plataforma de publicidade, a qual representa a principal fonte de lucro destas empresas.

## 1.2 Gap e solução proposta

Visando oferecer uma potencial complementação às estatísticas oficiais, devido ao alto custo de tempo e recursos conforme já foi supracitado, é buscado uma nova forma de inferir demografia a partir de dados disponíveis na internet. Utilizando-se de uma rede social, é possível inferir uma série de informações, dentre elas informações demográficas.

O presente trabalho trata da coleta de dados demográficos a partir da plataforma de publicidade do Facebook, o monitoramento da variação desses dados ao longo do tempo, e a comparação com dados reais. A coleta de dados, feita de forma online, garante um alto poder de monitoramento dos dados coletados, possibilitando análises mais frequentes.

Acredita-se que os dados coletados possam complementar as estatísticas oficiais, oferecendo estimativas oportunas entre os censos. A disponibilização, bem como a continuação da coleta periódica, podem criar novas perspectivas para o censo em períodos de tempo menores. Portanto, essa é uma proposta que objetivará, além do aumento no poder de monitoramento da variação dos números, a diminuição do custo para realização do censo, bem como sua execução em intervalos menores de tempo.

Apesar de esforços em estudos para inferir demografia a partir de dados online, ainda há espaço para avanços. As plataformas de publicidade das redes sociais vem se mostrando cada vez mais tecnológicas e precisas. Uma vez que avanços acontecem, novas técnicas vão surgindo, o que permite por sua vez substituir práticas antigas, como por exemplo a realização do Censo de forma tradicional. Entretanto, o presente trabalho não visa substituir o censo, pelo contrário, a ideia é complementar.

Alguns trabalhos tentaram inferir informações demográficas a partir de dados online, alguns em particular exploraram plataformas de publicidade de redes sociais. O objetivo é explorar os dados da plataforma de publicidade do Facebook, e comparar com dados de fontes oficiais, bem como verificar a análise temporal histórica.

## 1.3 Estado da arte

Um constante crescimento das redes sociais tem sido observado nos últimos anos, e com ele também um crescimento no interesse em obter dados demográficos a partir de um ambiente online, especialmente dados que são difíceis de se obter através de métodos tradicionais(CESARE et al., 2018).

Junto a esse crescimento, pôde-se notar uma exposição maior de informações individuais, que anos atrás seriam muito mais difíceis de conseguir. Quando fala-se disso, refere-se a atributos relacionados a dados demográficos básicos como idade, localização, gênero, dentre outros, como também interesses que podem variar dentre diversos tipos de preferências e características de comportamento.

Alguns trabalhos extraíram informações de plataformas de redes sociais. Disponibilizados através de ferramentas de publicidade, anunciantes tem acesso gratuito a dados agregados sobre os público-alvo, em que utilizam-se desses dados para construir cortes demográficos com pouco ou nenhum custo. A partir de informações disponíveis, estudos conseguiram inferir uma série de informações, como padrões comportamentais (DONG et

al., 2014), faixa etária e gênero (SAP et al., 2014), dentre outros. Trabalhos como estes, que inferem algum tipo de informações a partir de dados coletados das plataformas de publicidade, são expostos com mais detalhes no Capítulo 2.

## 1.4 Objetivos gerais

No presente trabalho, tem-se como objetivo a coleta de dados demográficos a partir da plataforma de publicidade do Facebook, o monitoramento da variação desses dados ao longo do tempo, e a comparação com dados reais.

Para os dados demográficos coletados referente a população dos Estados Unidos, foram coletados 7 tipos de atributos que são: *Inclinação política, afinidade racial, gênero, idade, nível educacional, status de relacionamento e imigrantes*, esses atributos possuem subgrupos (Tabela 1 mostra exemplos de atributos primários e seus subgrupos).

Para os dados demográficos coletados referente a população dos Brasil, foram coletados 5 tipos de atributos que são: *Gênero, idade, nível educacional, status de relacionamento e imigrantes*, esses atributos possuem subgrupos também mostrados na Tabela 1.

Nossa proposta é realizar coletas de dados periódicas, em que a partir desses dados, apresentaremos uma análise da variação sob o tempo. Os resultados obtidos, bem como os dados coletados, serão disponibilizados através de alguma plataforma para que possa ser facilmente utilizada em um trabalho futuro ou a pesquisadores da área.

### 1.4.1 Objetivos específicos

- Realizar a definição dos atributos a serem utilizados como parâmetro para a coleta de dados.
- Realizar a coleta dos dados semanalmente, por períodos maiores para analisar a variação.
- Inferir demografia a partir do conjunto de dados obtidos através das coletas.
- Verificar a variação sob o tempo da série histórica das coletas, na busca de possíveis variações abruptas ou estabilidade.
- Comparar resultados obtidos com fontes oficiais.
- Publicar os resultados e disponibilizar os dados coletados.

## 1.5 Organização do trabalho

Para início do trabalho, foi realizada uma revisão bibliográfica para compreender as metodologias de inferências demográficas a partir de dados coletados online, conforme é mostrado no Capítulo 2. O Capítulo 3 fornece detalhes da metodologia utilizada para a coleta e análise dos dados. A seguir, no Capítulo 4 são descritos os resultados obtidos a partir das análises da variação temporal dos dados e a comparação com fontes oficiais. Por fim, no Capítulo 5, conclusões e trabalhos futuros são apresentados.

# 2 Revisão bibliográfica

Meios de comunicação online, como por exemplo o Facebook, tornaram-se importantes para a comunicação dos indivíduos, mas por outro lado, por fornecerem a capacidade de classificar atributos latentes do usuário como sexo, idade e orientação política, tornaram-se também de grande importância para anúncios, marketing e vendas.

O constante crescimento das redes sociais possibilitou uma exposição maior de informações individuais, que anos atrás seriam muito mais difíceis de conseguir. Segundo Dong et al. (2014), é possível descobrir padrões de comportamento que os usuários utilizam para manter suas conexões. Segundo Qiu et al. (2018), a predição de influência social para cada usuário é critica para uma variedade de aplicações como recomendações online e publicidade. Os autores exploram ferramentas de rede neural profunda("Deep Neural Network") para aprender a representação de recursos latentes dos usuários para prever a influência social de diversos usuários em redes sociais como Facebook e Twitter. A influência social neste estudo é entendida como o "Fenômeno em que as emoções, opiniões ou comportamentos de uma pessoa são afetados por outras". Os experimentos demonstraram que o modelo proposto supera significativamente as abordagens tradicionais.

Alguns trabalhos utilizam várias informações oferecidas pelas redes sociais online para inferir dados demográficos. Estudos aproveitaram dessas informações disponíveis para reconhecer padrões comportamentais a partir da idade do usuário (DONG et al., 2014), determinar com alta precisão a faixa etária e gênero dos usuários a partir de textos públicos (SAP et al., 2014), inferir atributos como idioma nativo (ARGAMON et al., 2009), origem (RAO et al., 2010) e localização (JONES et al., 2007). Também foi utilizada para rastrear o interesse em causadores do tabagismo, obesidade e diabetes, em populações que apresentam essas condições (ARAUJO et al., 2017), e situações em que os anúncios discriminavam usuários de grupos sensíveis de receberem seus anúncios (SPEICHER et al., 2018).

No artigo publicado por Rao et al. (2010), os autores investigam e avaliam o desempenho em três atributos biográficos do usuário - gênero, idade e origem regional - e um atributo de personalização - orientação política. Para cada atributo, um conjunto de "sementes" de usuários relacionados é coletado e, a partir dessas sementes, outros candidatos em potencial são explorados de maneira ampla por meio da rede de seguidores do Twitter.

No atributo gênero, os autores detectaram se o usuário era um homem ou uma mulher, que foram determinados a partir do conteúdo e comportamento de suas postagens. Para o atributo idade, o trabalho foi feito manualmente, tendo cada usuário pertencente

a um grupo, os abaixo de 30 anos e os acima de 30 anos. O atributo origem regional os usuários foram dividimos manualmente entre norte e sul, e também utilizaram do GeoLocation do Twitter API, porém sem muito sucesso na cobertura. Por fim, para o atributo origem política, os autores investigaram os usuários do Twitter nos Estados Unidos, buscando distinguir os usuários entre republicano/conservador e democrata/liberal.

Atualmente, as redes sociais são fonte poderosa de compartilhamento de notícias em tempo real. Os usuários são bombardeados diariamente com um número desconcertante de opções de mídia de notícias. Um grande desafio da atualidade, é reconhecer uma fonte de dados como confiável ou não. No artigo publicado por Ribeiro et al. (2018), os autores utilizam as interfaces dos anunciantes, onde a partir delas buscam informações detalhadas sobre a demografia do público, para, por exemplo, estimar a inclinação ideológica (liberal ou conservadora) de uma fonte de notícias. Também foi possível estimar vieses demográficos como raça, gênero, idade, renda, entre outros. Com isso, foi possível desenvolver uma ferramenta escalável chamada "Media Bias Monitor", em que torna transparente para os usuários os preconceitos na demografia da audiência de milhares de veículos de notícia no Facebook.

Apesar da grande importância da realização dos censos por todo o mundo, eles representam uma fonte de dados muito caras, e possuem uma frequência de coleta muito baixa. Diversas pesquisas tem sido direcionadas para o uso de dados de mídias sociais e suas plataformas de propagadas para complementarem as fontes oficiais. Entretanto, os dados coletados de forma online, não representam um retrato fiel a realidade. Através da plataforma de publicidade do Facebook, Ribeiro, Benevenuto e Zagheni (2020) apresentaram uma compilação de informações buscando a equivalência de um censo com apenas usuários do Facebook. Foi possível identificar casos em que as estatísticas oficiais podem estar subestimando contagens populacionais, como por exemplo a imigração. Comparando os dados coletados de forma online, com relatórios oficiais fornecido por órgãos governamentais, foi possível encontrar diversas correlações em que, em termos percentuais, as estimativas eram próximas.

As redes sociais possuem como principal fonte de receita, a publicidade online. Plataformas como por exemplo o Facebook, permitem que anunciantes selecionem o alvo para a publicidade em questão. Características como idade, sexo, país de origem, nível educacional, entre outros podem ser definidos no momento da compra de uma publicidade digital, com a intenção de direcionar aquele anúncio para potenciais clientes. Ao selecionar o publico alvo, a plataforma de publicidade fornece uma estimativa de usuários que serão atingidos com os critérios definidos pelo anunciante. Tendo isso em vista, Zagheni, Weber e Gummadi (2017) utilizaram dessa ferramenta para mostrar a viabilidade de estimar a quantidade de imigrantes dentro e entre os países. Segundo os autores, com a rápida expansão global do uso das mídias sociais e a indústria da publicidade digital, uma gama

de novas oportunidades surgem para o monitoramento de dados demográficos a partir de dados online.

Por outro lado, plataformas de publicidade direcionadas digital, como o Facebook, foram criticadas por permitir que os anunciantes discriminem usuários pertencentes a grupos sensíveis, ou seja, excluam usuários pertencentes a uma determinada raça ou gênero de receber seus anúncios (SPEICHER et al., 2018). Com isso, para tentar amenizar o problema, o Facebook tentou proibir o uso de alguns desses atributos. Porém, segundo os autores, medidas como esta estão longe de ser suficiente e que o problema é muito mais pernicioso. Foi mostrado que, é possível para o anunciante criar anúncios altamente discriminatórios sem usar atributos confidenciais. Diante de uma análise completa do sistema de direcionamento de publicidade, foi possível observar mudanças necessárias no sistema para que não hajam brechas para esse tipo de situação.

Similarmente a realização do censo demográfico, em anos ou épocas próximas a eleições, é comum a realização de pesquisas eleitorais. Essas pesquisas mostram a intenção de voto em candidatos e podem ser usadas para monitorar as preferências eleitorais de determinadas populações durante a campanha. Assim como os censos, as pesquisas possuem um alto custo de tempo e recursos financeiros, principalmente as realizadas através de entrevistas presenciais. Em um trabalho recente, Ribeiro, Kansaon e Benevenuto (2019) propuseram uma abordagem que utiliza dados coletados da plataforma de publicidade do Facebook para inferir a demografia da audiência de candidatos as eleições. Os autores utilizaram a Attribute-Based Targeting (Direcionamento baseado em atributos) para selecionar "interesses" relacionados aos candidatos, e então comparar os dados com as pesquisas eleitorais realizadas no mesmo período. Os resultados sugerem que, a utilização dos dados extraídos representa uma boa indicação da variação nas intenções de voto. Também foi possível mostrar a ocorrência de variações nos aspectos demográficos dos apoiadores (sendo mais preciso para candidatos mais populares), bem como variações provocadas por eventos como protestos, por exemplo.

Considerando o Brasil como alvo do estudo, Júnior e Ribeiro (2020) realizaram um trabalho em que foi medido a similaridade cultural entre os estados brasileiros, analisando as preferências de comida e bebida como indicador e comparando essas informações com dados de migração interestadual. Também utilizando a plataforma de publicidade do Facebook, os autores buscaram estimativas do número total de usuários que atendem a um conjunto de atributos fornecidos, a fim de encontrar a audiência de alimentos e bebidas. Com a comparação dos dados obtidos de forma online com dados oficiais obtidos a partir do IBGE foi possível encontrar evidências do quanto a migração interestadual influencia no desenvolvimento cultural de um grupo. O estudo reforça a possibilidade da utilização de dados provenientes da plataforma de propaganda de publicidade de mídias sociais para pesquisas em diversas áreas.

# 3 Metodologia

Com a evolução significativa das redes sociais virtuais nos últimos anos, foram desenvolvidas diversas técnicas para inferir demografia através da internet. Com o acesso a informações pessoais e atividades de milhões de pessoas ao redor do mundo, as plataformas de publicidade conseguiram direcionar as publicidades para nichos específicos de usuários considerando certos tipos de atributos como nome, aspectos demográficos, comportamentos, entre outros. Este trabalho utiliza dados coletados a partir da plataforma de publicidade do Facebook.

As ferramentas de gerenciamento de anúncios providenciam três maneiras para definir a audiência de determinado anúncio, *Personally Identifiable Information (PII) targeting*, *Look-alike audience targeting* e *Attribute-based targeting*. A maneira escolhida para este trabalho é a *Attribute-based targeting* (Direcionamento baseada em atributos).

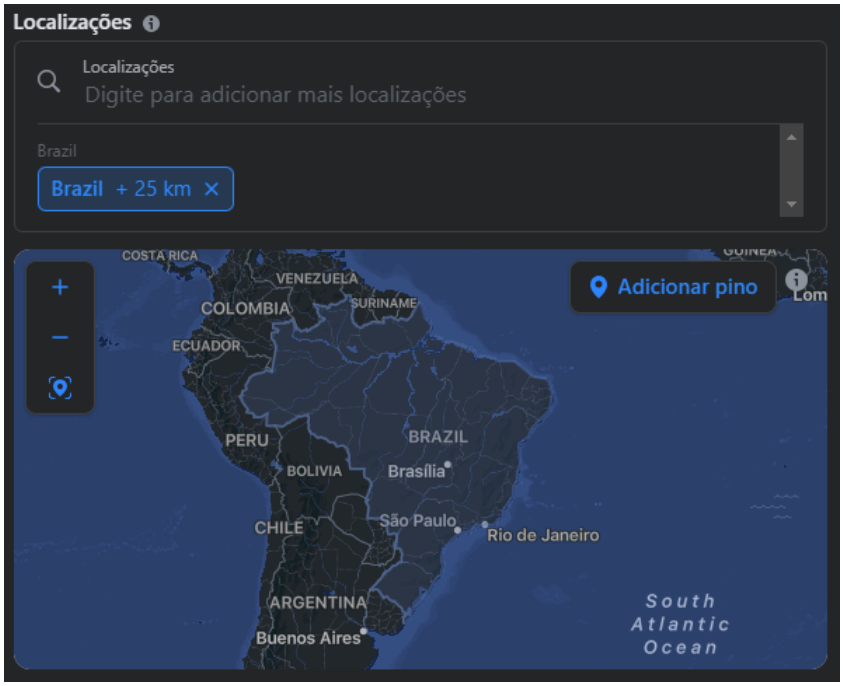
No direcionamento baseada em atributos: o anunciante pode definir o público-alvo com base em uma série de atributos que variam desde informações demográficos, interesses do usuário até o dispositivo que ele usa ao acessar o Facebook. A seleção desse conjunto de informações, é chamado de "fórmula", em que ela é utilizada para direcionar o anúncio para usuários com as caraterísticas selecionadas pelo anunciante. Dentre as opções de seleções, são encontrados atributos como estes:

- Dados demográficos básicos que incluem atributos mais gerais, por exemplo idade.
- Interesses que são os grupos nos quais o usuário demonstra interesse no Facebook, que podem ser desde política ou religião até formas de arte.
- Comportamentos que visam saber qual o estilo do indivíduo levando em consideração o tipo de dispositivo e plataformas utilizadas para acessar o Facebook (plataformas móveis, navegador, etc);
- Dados demográficos avançados que vão incluir dados que não só os dados básicos, mas também universidade em que o indivíduo frequentou até qual o seu nível de renda.

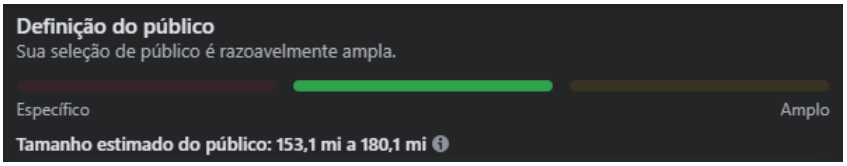
Nos últimos anos, as plataformas de publicidade em diversas redes sociais mostraram uma grande evolução, com isso surgiram novas formas de se definir a audiência a qual um determinado anúncio será direcionado, como citado anteriormente. No presente trabalho, foi utilizado o framework de um trabalho já desenvolvido (RIBEIRO et al., 2018), na qual foi utilizada a *Attribute-based targeting*, em que o anunciante escolhe entre uma variedade

de atributos demográficos, interesses e até atividades de lazer, para definir a audiência alvo.

Na Figura 2, é demonstrada a utilização de uma fórmula simples de direcionamento, que seleciona todos usuários do Facebook que moram no Brasil. Dessa forma obtemos um publico estimado entre 153,1 e 180,1 milhões de usuários de ambos gêneros e com idade acima de 13 anos (idade mínima permitida pelo Facebook). Através da interface de seleção de direcionamento detalhado, é possível incluir uma nova camada de seleção de atributos com intenção de nos permitir estratificar os dados de uma forma abrangente.




(a) Localização

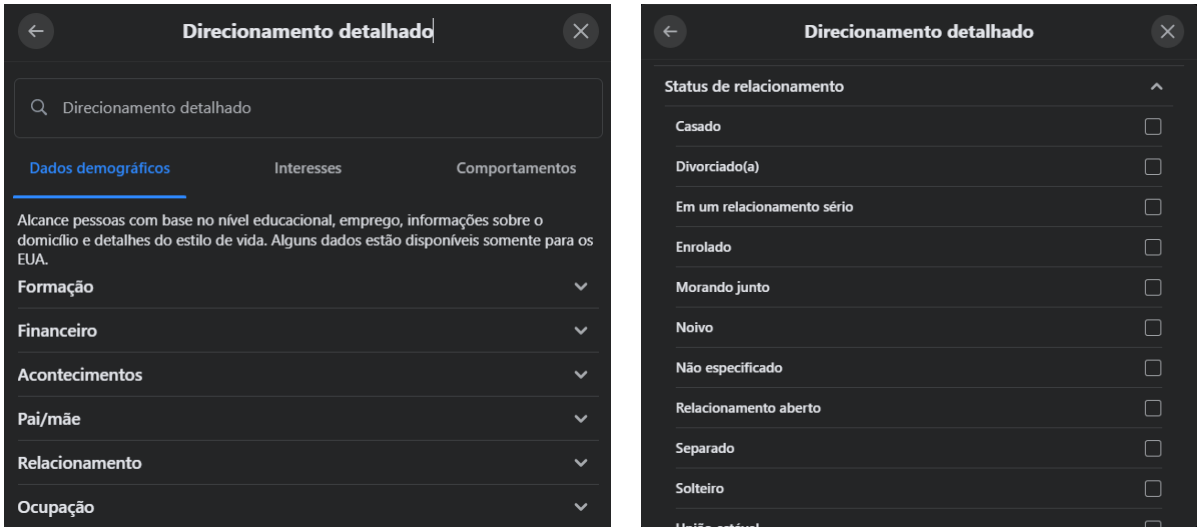



(b) Alcance

Figura 2 – Exemplo de publicidade direcionada para usuários que moram no Brasil para a plataforma de publicidade do Facebook.

A Figura 3 representa um exemplo de como escolher uma audiência alvo através da interface de direcionamento detalhado da plataforma de publicidade do Facebook. É possível assim, refinar a coleta de dados para obter a quantidade de usuários que além de morarem no Brasil também são solteiros ou casados por exemplo, obtendo então, as seguintes quantias respectivas: 20 milhões e 28 milhões aproximadamente.

De forma similar, pode-se criar deferentes formulas para direcionar publicidades para diversos grupos específicos de usuários, a critério do anunciante. Com isso, é possível




(a) Direcionamento detalhado (b) Seleção de Interesses

Figura 3 – Exemplo de seleção de interesses para direcionamento da publicidade no Facebook.

estratificar ainda mais alguns atributos, e criar subgrupos a partir deles, como por exemplo com o gênero (Masculino e Feminino). A partir da divisão de um atributo em subgrupos, pode ser calculada a proporção de cada subgrupo para com o todo.

Realizando uma simulação de anúncio, e direcionando apenas para homens (H) e em sequência para mulheres (M) que moram no Brasil, obtemos os respectivos valores: 63,4 milhões e 70,7 milhões. Portanto, para calcular a proporção de algum subgrupo (P), basta dividir o respectivo valor pela soma de todos subgrupos. Aplicando essa conta ao exemplo do subgrupo gênero, pode-se dizer que as proporções para homens e mulheres são respectivamente 47% e 53%. Isso é possível ser feito de forma similar para os outros atributos coletados.

$$P(homem) = \frac{H}{H+M} = \frac{63,4}{63,4+70,7} = 47\%$$

$$P(mulher) = \frac{M}{H+M} = \frac{70,7}{63,4+70,7} = 53\%$$

Segundo o próprio Facebook, as estimativas de público-alvo são obtidas a partir de vários fatores, como comportamento, demografia, localização, entre outros. Esse processo é desenhado de forma a estimar quantas pessoas, dadas uma série de características, são elegíveis a visualizar uma publicidade que um negócio pode rodar. Apesar de esforços para que as estimativas sejam precisas, elas não são desenhadas para corresponder com os resultados oficiais dos censos.

Tendo isso em mente, é possível identificar alguns desafios no momento da realização deste trabalho: um dos principais pontos observados, foi a diferente nomenclatura para

os dados coletados a partir do Facebook e os dados dos censos. Na plataforma online, o usuário possui mais opções de escolha de características do que em um censo real. Portanto, para alguns atributos analisados, foi necessário manipular os dados para que os mesmos correspondessem ao censo oficial.

Outro ponto importante, é o fato do Facebook possuir um mecanismo que protege a identidade dos seus usuários. Em outras palavras, caso os atributos selecionados no momento da pesquisa, correspondam a uma subpopulação menor do que mil usuários, o valor retornado será sempre 1000. Apesar desse mecanismo de proteção à privacidade acarretar em uma limitação na busca de dados para pequenas cidades, o presente estudo não foi impactado por tratar apenas do nível nacional.

Por fim, não é possível realizar uma análise sobre a quantidade de características falsas informadas pelos usuários em seus próprios perfis, como por exemplo idade ou nível educacional. Por esse motivo é impossível estimar esses valores na coleta dos dados. O próprio Facebook busca maneiras de combater informações falsas em sua plataforma, na tentativa de realizar estimativas precisas. Isso é feito almejando uma maior confiabilidade em sua plataforma de publicidade.

| Lista de Atributos | | |
|---|---|---|
| Status de Relacionamento | Afinidade Racial | Gênero |
| Solteiro, Em Relacionamento, Noivo, Casado, União Civil, Parceria Doméstica, Rel. Aberto Complicado, Separado, Divorciado Viúvo, Não Especificado | Africano-Ameriacano, Asiático-Americano, Caucasiano, Hispânico | Masculino Feminino |
| Escolaridades | Imigrantes | Inclinação Política |
| Fundamental completo e médio incompleto, Médio completo e superior incompleto, Superior completo, Não determinado | Mexico, Europa, Asia, America Latina, Africa | Conservador, Liberal, Moderado, Muito conservador, Muito Liberal |
| Idade | | |
| 13 - 17; 18 - 24; 25 - 34; 34 - 44; 45 - 54; 55 - 64; 65+ | | |

Tabela 1 – Exemplo de subgrupos para atributos primários.

Para a automatização da coleta de dados foi utilizada a API de Marketing do Facebook, disponibilizada na linguagem *Python*.

Foi necessário um estudo do pacote *facebook_business*, o qual fornece uma interface entre a aplicação e a API de marketing do Facebook. Para isso, vários testes foram realizados até ser possível o entendimento de como são feitas as requisições para obtenção

dos dados que desejamos.

Com isso, foi realizado a definição dos interesses que compuseram as fórmulas para coletar cada um dos atributos e seus subgrupos. Esses atributos e seus subgrupos são mostrados na Tabela 1, e serão utilizados no Capitulo 4 para a demonstração dos resultados.

Com o escopo da coleta de dados definido, conforme foi mostrado no Capítulo 1, foram iniciadas as coletas de dados. As coletas foram realizadas uma vez por semana, nos períodos indicados a seguir:

- Estados Unidos:
  - Semanalmente de 05/2019 até 11/2019
  - Semanalmente de 03/2020 até 07/2020
  - Semanalmente de 10/2020 até 11/2021 (Sem o atributo ”inclinação política”)
- Brasil:
  - Uma coleta em 10/2019 - Coleta teste.
  - Semanalmente de 06/2020 até 11/2020
  - Semanalmente de 01/2021 até 11/2021

## 3.1 Análise dos dados

Nessa seção será detalhado como foram conduzidas as análises realizadas a partir dos dados obtidos através das coletas. Como é de objetivo do trabalho, foi realizado a verificação da variação temporal de cada um dos atributos coletados na busca de variações abruptas. O primeiro passo foi gerar graficamente a série histórica da fórmula mais simples utilizada nas coletas de dados, da qual simplesmente buscava saber o número de usuários do Facebook, de qualquer idade e gênero, tanto para o Brasil quanto para o EUA.

Com a visualização facilitada, a ideia é identificar casos em que houve variação abrupta em algum subgrupo. Portanto, identificar casos em que atributos tiveram variações, em que na realidade, essas variações não ocorrem. O mesmo procedimento foi realizado para os demais atributos, ainda com a intenção de facilitar a visualização e identificação da variação sob o tempo.

A partir disso, podemos comparar esses resultados com fontes oficiais, como por exemplo o IBGE. Uma das análises realizadas, conforme será demonstrada no Capítulo 4, exibe recortes da população estimada tanto pelo Facebook, quanto pelo IBGE (IBGE, 2018), para os anos de 2020, 2021 e 2022, na intenção de comparar o crescimento populacional estimado em ambos os casos.

Também utilizando-se de um recorte, datado de novembro de 2022 da coleta do Facebook, e comparando com o censo de 2010 do IBGE (IBGE, 2010b), foi construído um gráfico de barras agrupadas para exibir a população agrupada por idade e comparar ambas as estimativas. Novamente, com a visualização gráfica é possível identificar situações em que o Facebook pode estar subestimando ou superestimando as estimativas. O mesmo procedimento foi realizado para a análise da distribuição educacional.

De forma similar, as análises dos demais atributos seguiram a mesma linha. Em que, a partir de um recorte, foram feitas as comparações das informações obtidas de forma online com fontes oficiais na busca de correlações e disparidades. Para o atributo Afinidade Racial, foi utilizada uma tabela da American Community Survey, a ACS-S0201 (CENSUS, 2019), disponibilizada pelo próprio U.S. Census Bureal. Para o atributo Status de Relacionamento, foi utilizada a amostra de nupcialidade do censo de 2010 (IBGE, 2010d), disponibilizada pelo IBGE.

Por fim, utilizando-se de dados disponibilizados no Portal da Imigração, através do relatório anual de 2020 da OBMigra (OLIVEIRA T; CAVALCANTI, OBMigra, 2020), do relatório conjuntural do primeiro quadrimestre de 2020 (SIMÕES A; HALLAK NETO, 2020), e do relatório de dados consolidados de 2022 (OLIVEIRA T; CAVALCANTI, 2022), foram realizadas as análises da variação temporal do atributo Expatriados na busca de correlações.

## 3.2 Fluxograma

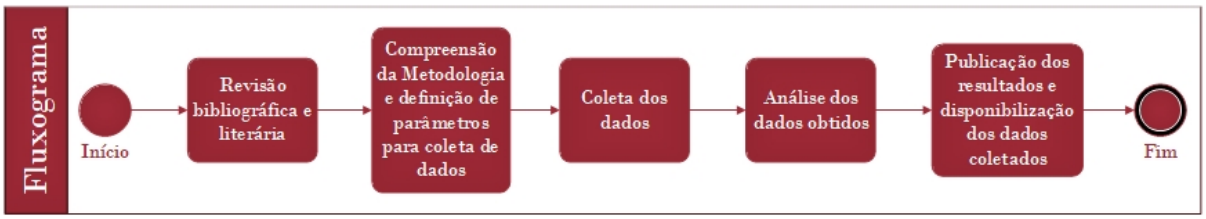
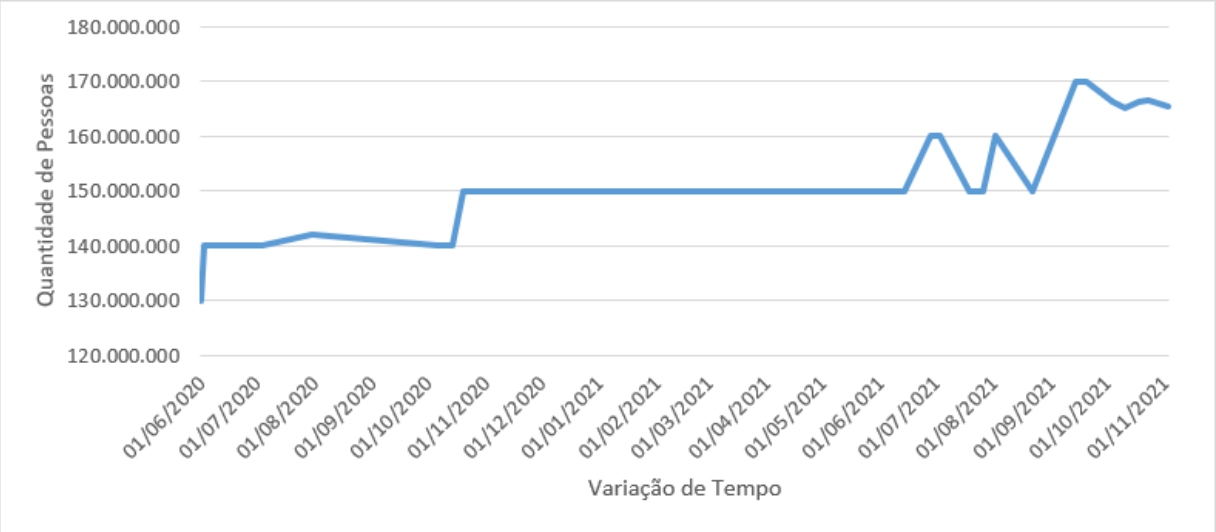



Figura 4 – Fluxograma.

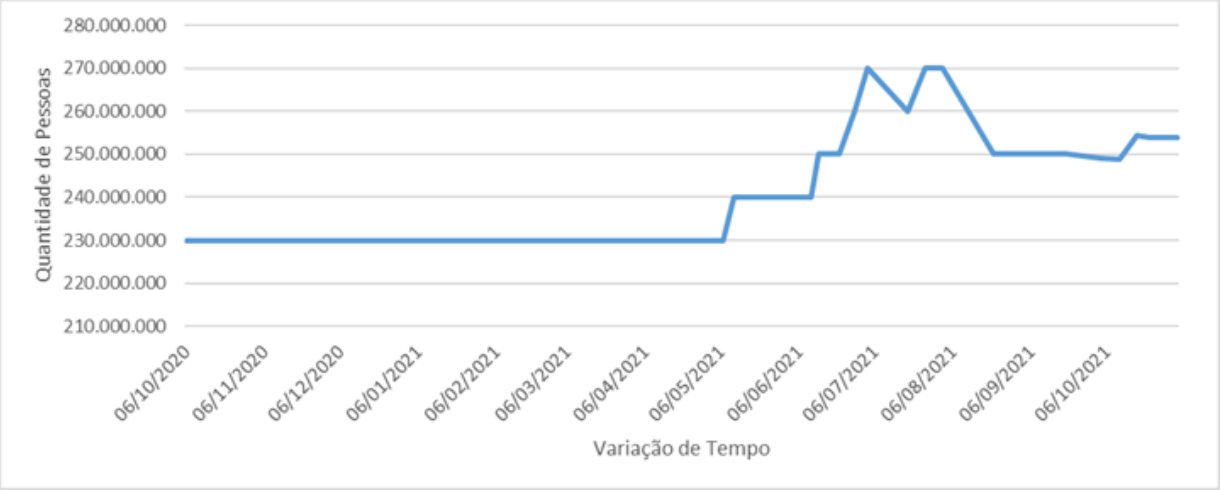
# 4 Resultados

Nesta seção, tem-se como objetivo a visualização dos dados coletados através da plataforma de publicidade do Facebook, e a comparação com os dados de fontes oficiais. Para as análises realizadas, foram utilizadas as bases de dados mais recentes disponibilizados pelo Census Bureal e IBGE, órgãos responsáveis pela realização do Censo demográfico no EUA e Brasil, respectivamente.

Com a consolidação dos dados coletados é possível, a partir da visualização gráfica, verificar e analisar a variação no número de usuários do Facebook como por exemplo é mostrado na Figura 5.




(a) Brasil




(b) EUA

Figura 5 – Variação da população, segundo coleta de dados no Facebook.

Em ambas figuras, nota-se que as maiores variações ocorreram do meio para o final

de 2021, em que para o EUA, por exemplo, entre maio e setembro a quantidade de usuário variou várias vezes. Saindo de 230 milhões até um pico de 270 milhões, esse é um tipo de variação que na realidade não ocorre, exceto em situações atípicas (como por exemplo um país recebendo refugiados de guerra).

Apesar de ser uma variação que chama atenção, o Facebook não provê uma explicação para essas estimativas, o que torna a confiabilidade do dado um pouco mais fraca. Como não é possível saber quando a estimativa estará mais próxima da realidade, dadas as promoções (por causa da grande quantidade de variações), não é possível dizer que a estimativa é precisa.

Alguns dados oficiais, como por exemplo o tamanho da população estimada, é divulgado no do IBGE, juntamente com os dados do último Censo realizado (IBGE, 2010b). Mesmo com algumas variações negativas, é possível perceber que, de forma geral, a quantidade de usuários do Facebook no Brasil aumentou com o passar do tempo. De forma similar, segundo projeções de estimativa da população brasileira, feito pelo IBGE (IBGE, 2018), de 2020 até 2022 também foi possível visualizar um aumento na quantidade de pessoas. É possível observarmos lado a lado o comparativo dessas duas variações, conforme mostrado na Figura 6.

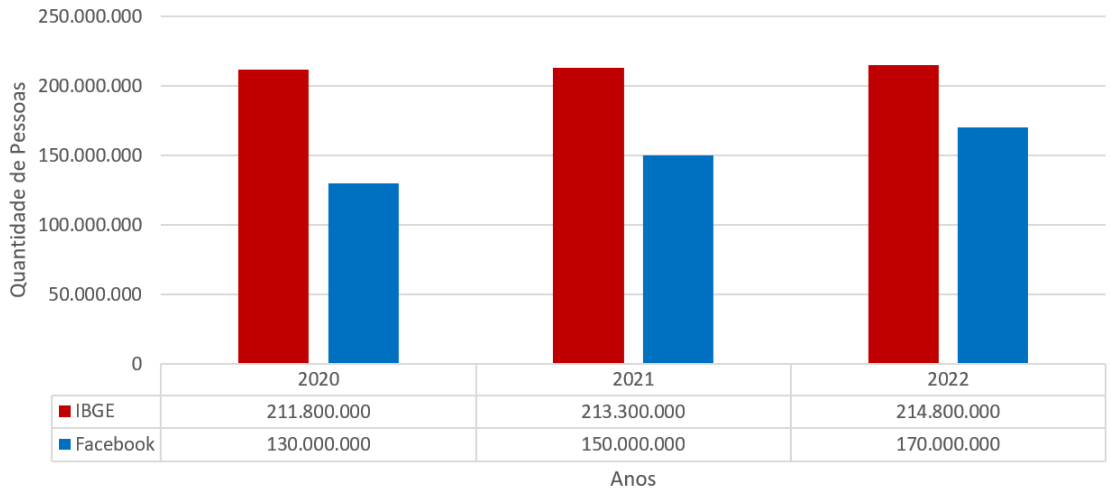



Figura 6 – População total ao longo dos anos: IBGE versus Facebook

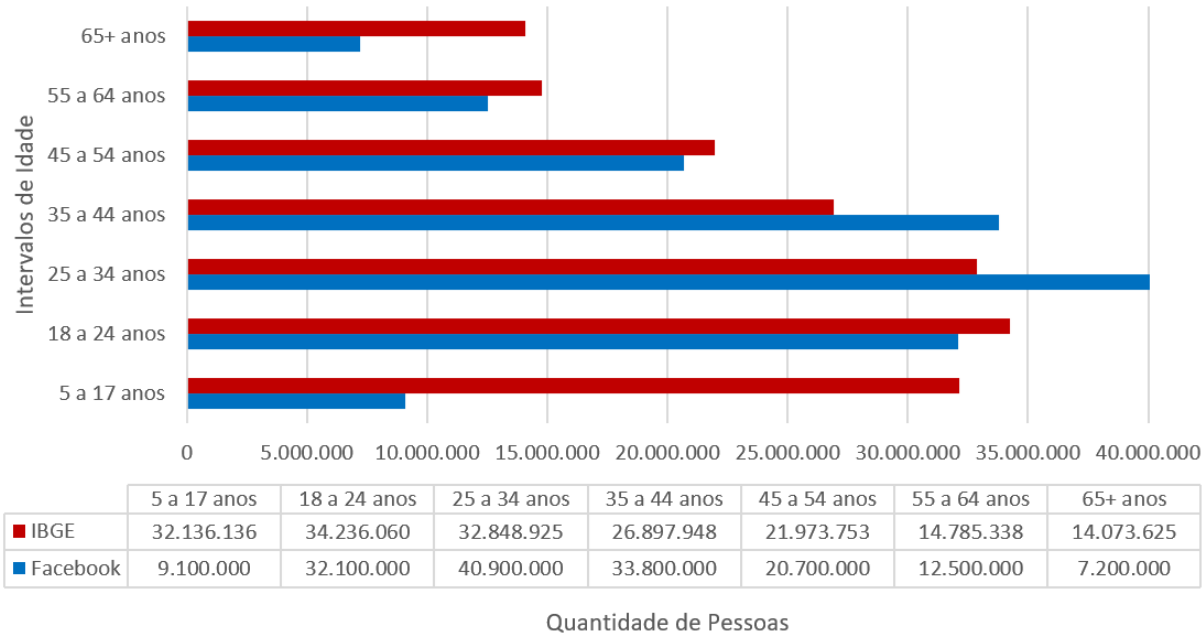
A partir da Figura 6, é possível observar que o crescimento de usuários do Facebook teve um aumento em uma proporção muito maior do que as estimativas para o censo oficial do IBGE. É possível verificar um crescimento de 40 milhões de usuários, contra 3 milhões das estimativas oficiais. Caso a população do Facebook continue crescendo nessa proporção, nos próximos anos ela superaria as estimativas oficiais, o que de certa forma gera uma inconformidade que descredibiliza a confiabilidade dos dados. Esse crescimento pode ter como um dos fatores influenciadores, a pandemia do Covid-19.

Além de uma análise apenas olhando a quantidade total de pessoas, também é

possível segregar ainda mais a pesquisa, com a divisão do tamanho da população por grupos de idade, como mostrado na Figura 7, por exemplo. O Facebook possui aproximadamente 170 milhões de usuários que residem no Brasil, em 2022. A partir da observação da comparação dos dados, é possível notar uma discrepância alta na população com idade abaixo de 17 anos, com o Facebook possuindo uma amostra significativamente menor do que indicado pelos dados oficiais. Essa discrepância ocorre pois o Facebook não permite o registro de pessoas abaixo da idade mínima permitida pela política da rede social, de 13 anos de idade.

A mesma situação pode ser observada para o grupo com idade superior a 65 anos. Mesmo com o aumento de usuários durante os últimos anos, pessoas mais velhas tendem a utilizar menos plataformas de redes sociais, do que pessoas mais novas (GIL-CLAVEL; ZAGHENI, 2019), como é possível evidenciarmos também a partir da visualização dos dados coletados.

Por outro lado, para grupos idade entre 25 e 34 anos, o Facebook possui uma população maior do que a indicada pelo censo oficial. Novamente, esse resultado pode ser um reflexo da criação de múltiplas contas por alguns usuários. Uma das hipóteses é a possibilidade de impacto das contas falsas, conforme indicado que pode ocorrer nas redes sociais, conforme mostrado pelo trabalho desenvolvido por Leite e Salles (2019).

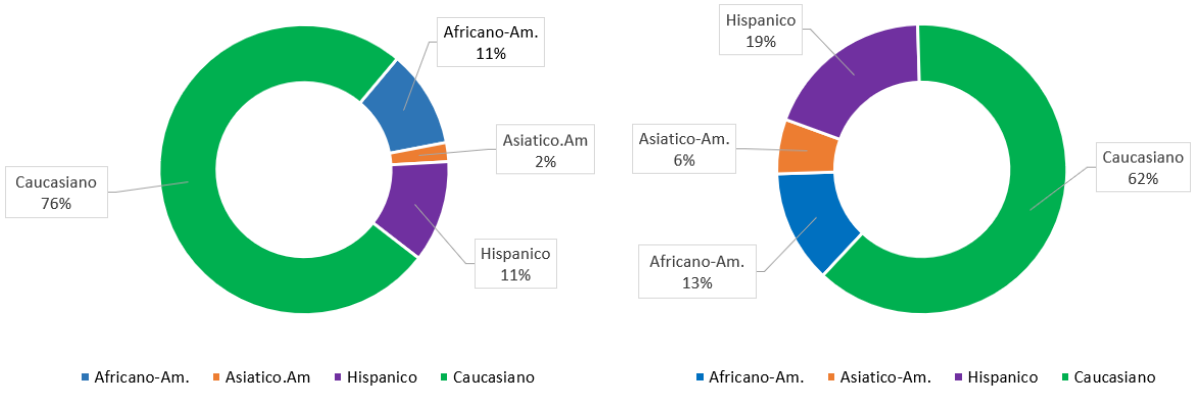



| | 5 a 17 anos | 18 a 24 anos | 25 a 34 anos | 35 a 44 anos | 45 a 54 anos | 55 a 64 anos | 65+ anos |
|---|---|---|---|---|---|---|---|
| IBGE | 32.136.136 | 34.236.060 | 32.848.925 | 26.897.948 | 21.973.753 | 14.785.338 | 14.073.625 |
| Facebook | 9.100.000 | 32.100.000 | 40.900.000 | 33.800.000 | 20.700.000 | 12.500.000 | 7.200.000 |

Figura 7 – População agrupada por idade no Brasil: Facebook vs IBGE

De forma similar, é possível realizar a mesma análise para a distribuição racial e étnica para a população do EUA (CENSUS, 2019). Na Figura 8, é possível visualizar lado a lado a porcentagem de pessoas em cada grupo étnico ou racial. A plataforma

de publicidade do Facebook, possui em seus atributos, dentro da categoria “afinidade”, subgrupos que se correlacionam com os grupos étnicos em censos reais.




(a) Facebook (b) US Census Bureal

Figura 8 – Distribuição racial e étnica no EUA, em comparação com o Facebook.

Mesmo os números não sendo exatamente iguais, é possível notar uma grande similaridade na distribuição desse atributo em específico. Para essa comparação, foi utilizado uma coleta de dados realizado em Maio de 2019 para os dados do Facebook, e o Censo de 2020, para os dados do EUA. Quando a análise é estendida para todas as coletas, visando observar a variação desses dados, encontramos o resultado mostrado na Figura 9

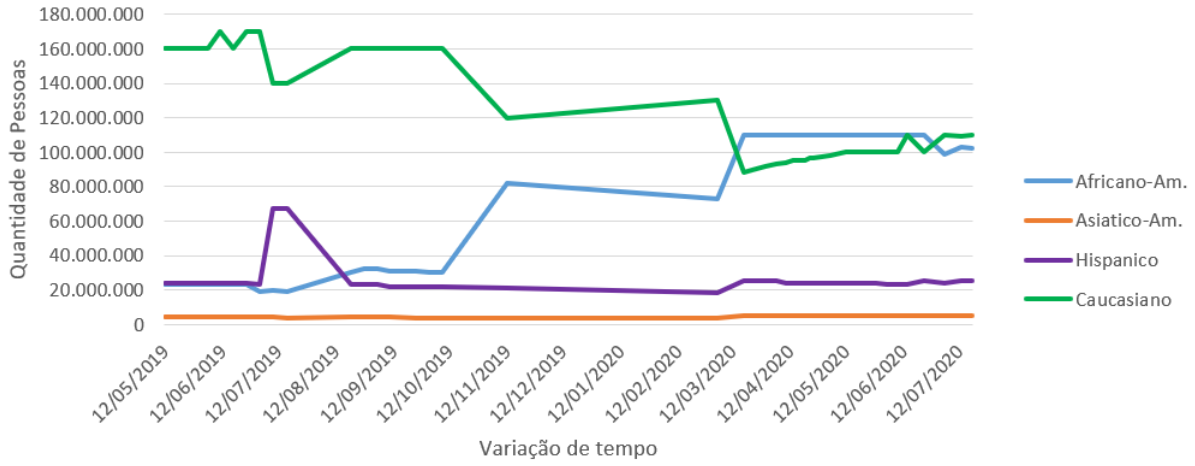



Figura 9 – Variação da distribuição racial no EUA, segundo o Facebook.

Como é possível observar, os subgrupos de Caucasianos e Africano-Americanos sofreram variações abruptas nesse período de aproximadamente um ano. Se fosse aplicado novamente o cálculo de percentual para estes mesmos dados, porém utilizando a coleta de julho de 2020, seriam encontradas as seguintes porcentagens: Caucasiano 46%; Africano-Americano 42%; Hispânico 10%; Asiático-Americano 2%; Das quais não trazem uma boa representação dos números reais. Essa estimativa é claramente um erro do Facebook, pois ela coloca o número de Africano-Americano como a maior população do EUA, e isso está longe de ser verdade.

Portanto, para esse atributo em específico, é possível verificar que os dados podem não ser confiáveis, devido a alta variação de seus subgrupos. Isso pode ser verificado em trabalhos futuros que deem continuidade com a coleta de dados destes atributos.

Assim como na distribuição racial e étnica para os EUA, a distribuição educacional no Brasil também apresenta grandes diferenças em seus valores. Quando comparamos os dados da plataforma de publicidade Facebook (coleta de novembro de 2021) com o censo realizado em 2010 pelo IBGE (IBGE, 2010a), é possível notar isso com clareza.

A figura 10 permite a visualização gráfica desses dados, quando observados lado a lado. Quando consideramos dados produzidos por usuários de redes sociais online, não é possível garantir que a informação está correta. Cada usuário tem a liberdade de preencher o seu perfil conforme desejar, e isso inclui inserir informações falsas, ou até escrevendo uma informação qualquer, apenas para não deixar o campo sem preenchimento. Outros usuários podem deixar de preencher informações por questões de privacidade, ou até mesmo por que não desejam gastar seu tempo com isso. É possível notar que, aproximadamente 60 milhões de usuários não preencheram o campo de escolaridade, em seus cadastros. Os dados também mostram uma superestimativa do Facebook com relação a usuários que possuem ensino superior completo, ultrapassando em quase 20 milhões de pessoas nessa subcategoria.

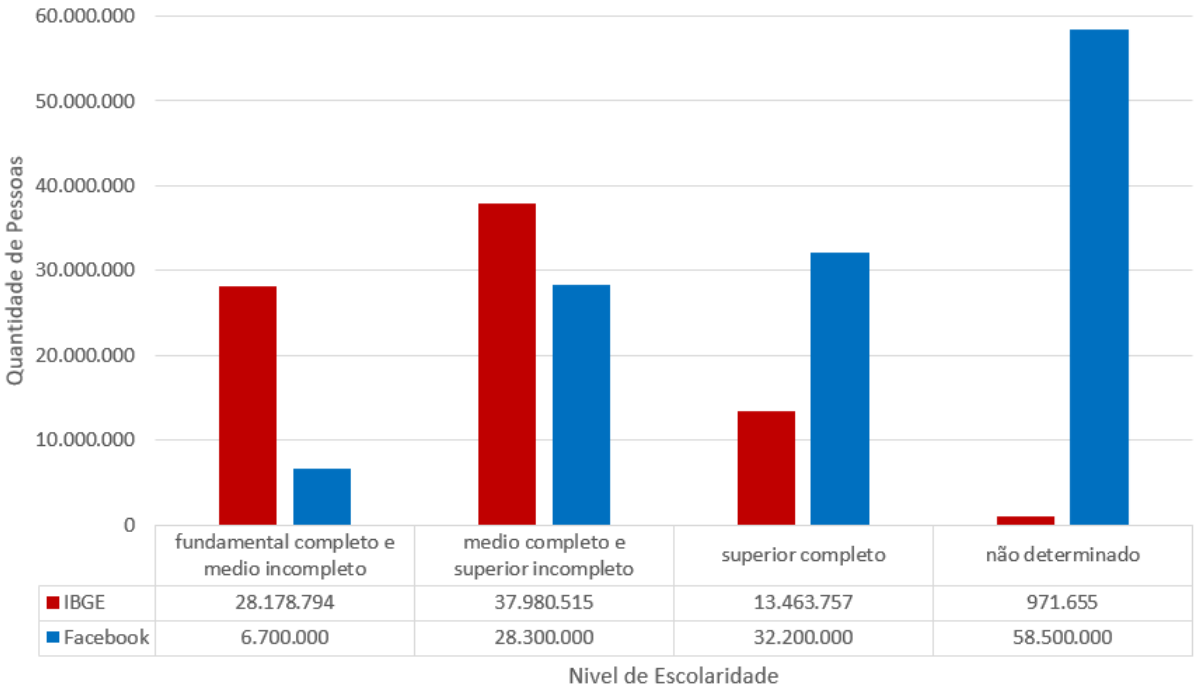



Figura 10 – Distribuição educacional no Brasil: IBGE versus Facebook

Para a próxima análise, será comparado o atributo Expatriados, que está relacionado a Imigrantes. Porém, antes é necessário entender como os fluxos de movimentação internacional foram afetados nos últimos anos.

A pandemia de COVID-19 implicou na maior redução dos movimentos de entrada e saída do Brasil na última década (OLIVEIRA T; CAVALCANTI, OBMigra, 2020). Tendo seu início a partir de março de 2020, a pandemia de COVID-19 fez com que o governo federal brasileiro impusesse uma série de portarias regulando os fluxos migratórios no país.

Com uma movimentação mensal de quase 2,5 milhões de entradas e saídas em 2019, em junho de 2020 esse número era de apenas 40 mil, que representa uma queda sem precedentes na série histórica, segundo dados utilizados pelo relatório anual da OBMigra, de agosto de 2020 (OLIVEIRA T; CAVALCANTI, OBMigra, 2020).

As movimentações de entrada e saída, são divididas em categorias, sendo elas: Brasileiro, Turista, Trânsito, Residente, Temporário, Fronteiriço e Não nacionais, deportados ou extraditados. Praticamente todos os tipos de movimentações tiveram quedas expressivas em 2020, quando comparadas com o ano anterior. A única categoria em que teve aumento no período, foi a movimentação de saída da categoria "Não nacionais, deportados ou extraditados" (+969,1%). Esse aumento pode estar relacionado as portarias decretadas em razão da pandemia. Entretanto, mesmo com essas variações, o numero de entradas foi maior do que saídas (Aproximadamente 50 mil).

Com todas as restrições impostas e os fluxos migratórios diminuindo drasticamente, como resultado a quantidade de registro de entrada de imigrantes também teve queda considerável. Com queda registrada a partir da segunda quinzena de março, o maior impacto foi registrado em abril, segundo relatório conjuntural da OBMigra referente ao primeiro quadrimestre de 2020 (SIMÕES A; HALLAK NETO, 2020).

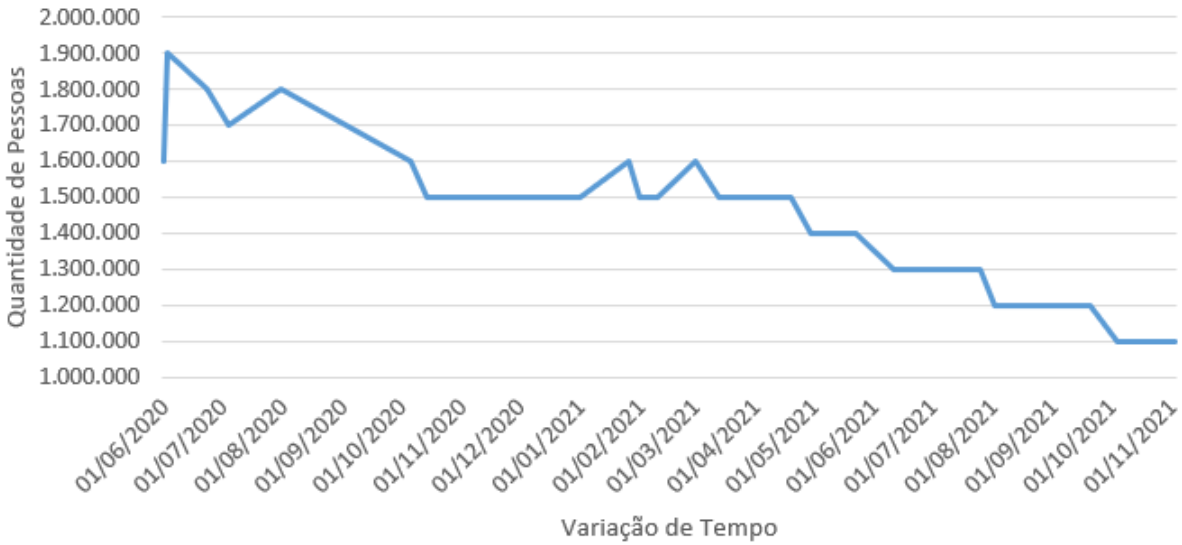



Figura 11 – Variação total de Imigrantes para o Brasil, segundo o Facebook

Solicitações de refúgio, autorização de residência fixa e temporária, autorização para trabalhadores qualificados, dentre outros tipos de registros de imigração apresentaram quedas expressivas em seus números durante o período analisado.

Segundo o relatório anual de 2020 da OBMigra, o Brasil possui 1.3 milhões de imigrantes, porém, segundo o Facebook, em junho de 2020 o Brasil possuía 1.9 milhões de imigrantes. Apenas ao final de 2021 que esse número cai para 1.1 milhão.

Conforme é possível verificar na figura 11, essa queda ocorreu gradativamente ao longo de 2020 e 2021, sem variações destoantes.

Analisando os dados de imigração de forma estratificada, olhando para os subgrupos de imigrantes, e gerando um gráfico para se observar a variação desse atributo em relação ao tempo, obtemos a figura 12. É possível verificar uma queda constante nos números para o ano de 2020 (Aproximadamente 100 mil para o subgrupo América Latina), seguidos de um pequeno aumento no começo de 2021, onde a partir de abril teve queda constante. O mesmo comportamento de queda nos dados verificado na análise do atributo como um todo se manteve quando é realizado a análise para o atributo dividido em seus subgrupos.

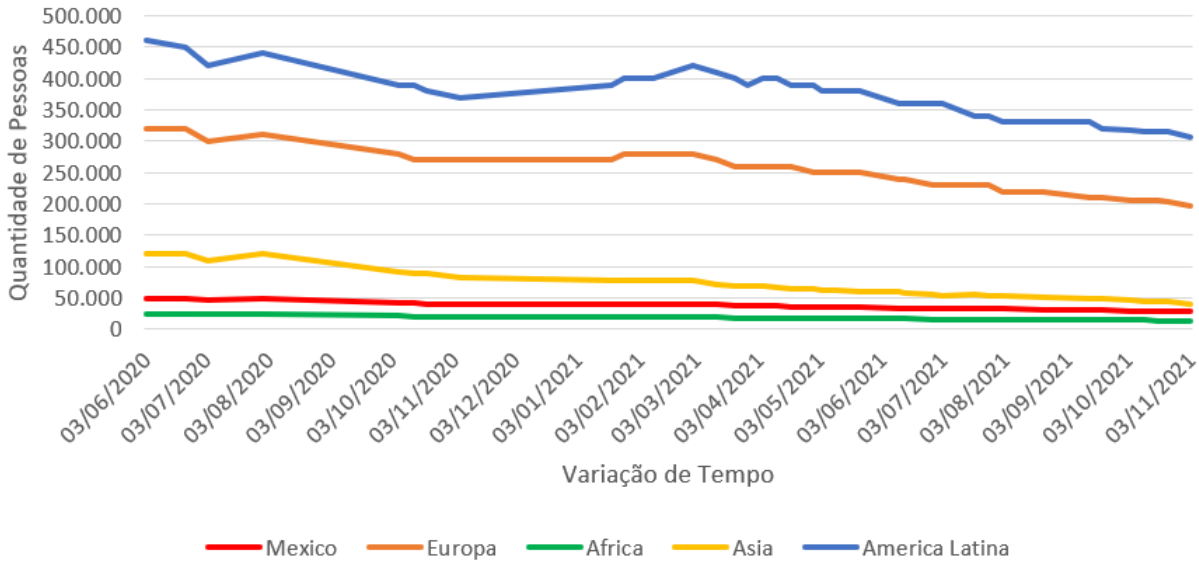
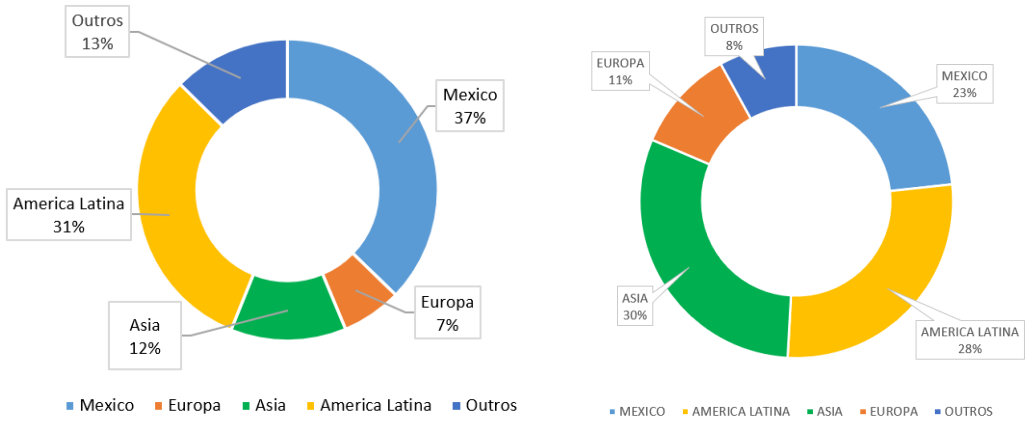



Figura 12 – Variação de imigrantes para o Brasil, segundo o Facebook

Para os Estados Unidos, é possível fazer uma análise da distribuição dos imigrantes no país, em comparação com dados oficiais do US Census Bureau. As ultimas pesquisas oficiais realizadas, estimam que o EUA possui aproximadamente 45 milhões de imigrantes, aproximadamente 15% da população americana.

O Facebook estimou em 2020 uma população de 26 milhões de imigrantes, o que representava aproximadamente 11% da população naquele momento. Apesar do número de usuários continuar crescendo, chegando em até 270 milhões em meados de 2021, o número de expatriados diminuiu, chegando em 19 milhões em novembro de 2021, apenas 3,5%. Esse fato, pode ter associação provável relacionada a pandemia do Covid-19, onde diversas pessoas no mundo inteiro retornaram a seus países de orígem.

A figura 13 mostra a distribuição em porcentagem dos principais países que possuem




(a) Facebook (b) US Census Bureal

Figura 13 – Distribuição de expatriados no EUA, em comparação com o Facebook.

imigrantes morando no EUA, contra um corte de junho de 2020 do Facebook.

Comparando as duas figuras, nota-se uma grande diferença nas estimativas das populações asiáticas e mexicanas. Os demais atributos obtiveram uma estimativa similar, com uma diferença percentual de 5% ou menos. Mesmo considerando diferentes períodos de coleta, essa distribuição permanece praticamente a mesma, o que pode ser um indício de um certo grau confiabilidade dos dados para este atributo.

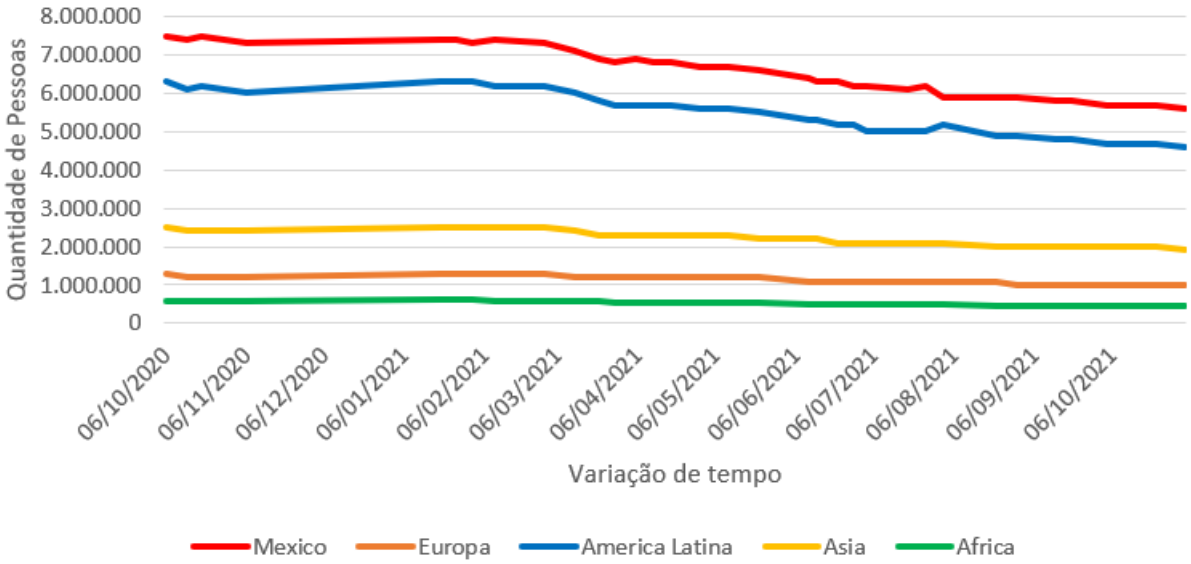
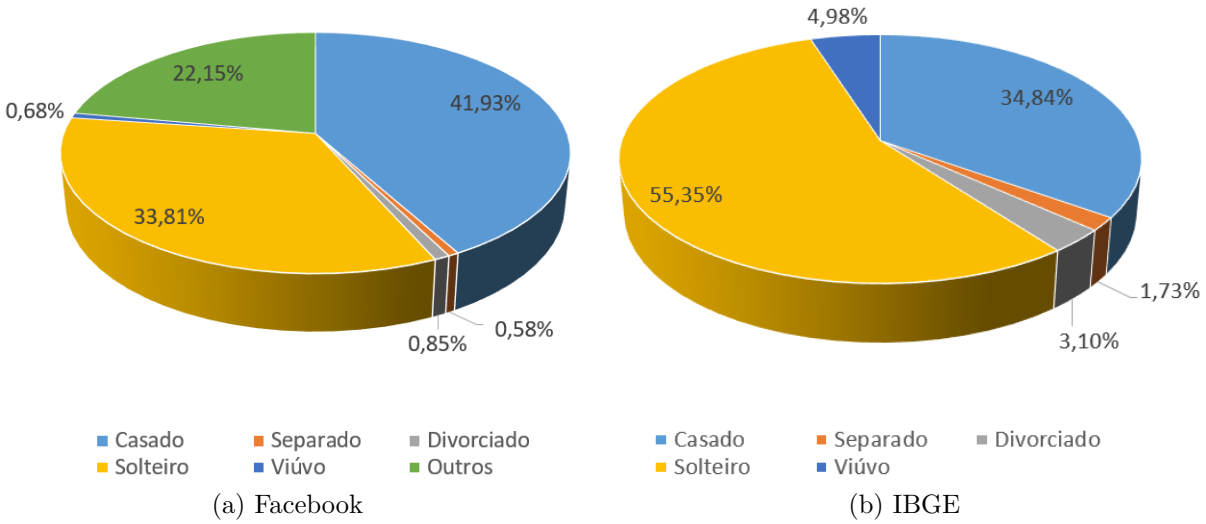



Figura 14 – Variação de imigrantes para o EUA, segundo o Facebook

Para visualização da variação dos dados expostos nas figuras anteriores, na figura 14, verifica-se a variação de imigrantes no EUA. Nota-se que as maiores variações ocorreram para os subgrupos “México” e “América Latina”, em que em ambos os casos a população caiu em mais de 1 Milhão no período de 1 ano. Pode-se notar aqui, uma similaridade no comportamento dos atributos coletados para o EUA com os do Brasil, no qual ambos

apresentaram quedas significativas em seus subgrupos. Esse comportamento também pode ser justificado pela pandemia, onde diversas pessoas retornaram ao seu país de origem.

A seguir, foi analisado o atributo "Status de Relacionamento" referente aos usuários que moram no Brasil. Buscando encontrar similaridades com fontes oficiais, foi utilizado uma coleta de junho de 2020 do Brasil, e a amostra de nupcialidade do censo de 2010 do IBGE (IBGE, 2010d).




(a) Facebook (b) IBGE

Figura 15 – Distribuição nupcial no Brasil: IBGE versus Facebook

Para essa comparação, são exibidos os subgrupos em que são considerados pelas estimativas oficiais, e que também estão presentes na plataforma do Facebook. Ainda que a rede social considere todos tipos de relacionamentos oficiais considerados pelo IBGE, o Facebook disponibiliza para seus usuários outras opções de status de relacionamento que não estão inclusas no Censo oficial. Essas outras opções são representadas na figura 15 como o subgrupo "Outros", e sendo elas as seguintes: Em Relacionamento, Noivo, União Civil, Parceira Doméstica, Relacionamento Aberto e Complicado.

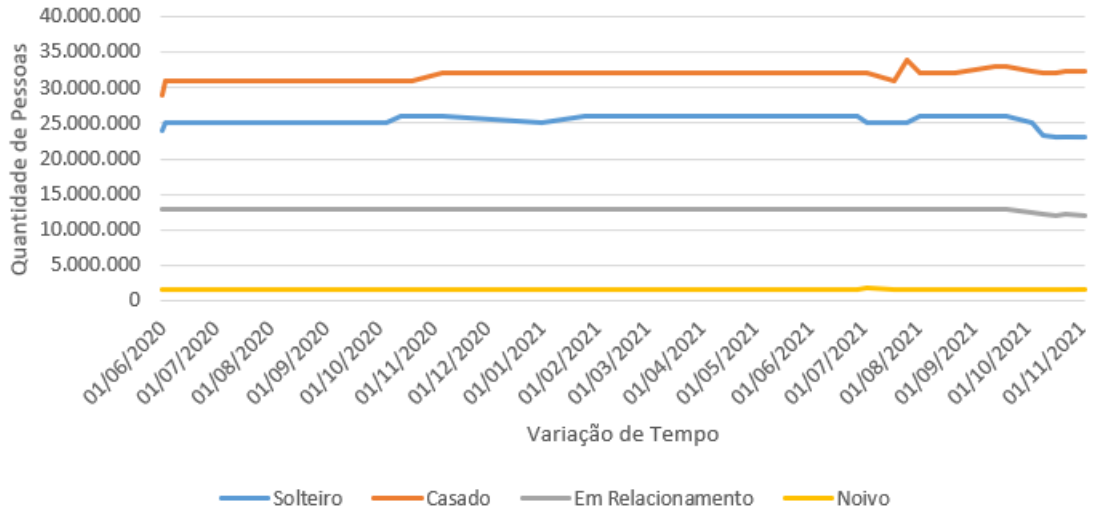
Logo de imediato é possível verificar uma grande diferença na estimativa de solteiros entre a coleta online e o censo oficial. Um grande fator influenciador para que isso ocorra, é o fato do IBGE não considerar outras formas de relacionamento, da mesma forma que o Facebook considera. Para o censo oficial, mesmo que uma pessoa esteja namorando ou noiva ela é considerada na pesquisa como solteira.

Na tentativa de buscar uma estimativa mais próxima da estimativa oficial, se agrupássemos o subgrupo "Outros" com o subgrupo "Solteiro", este subgrupo passaria a representar 55,96% da amostra. Essa estimativa se aproxima bem mais do censo do IBGE, que é de 55,35%. Entretanto, não é possível afirmar que essa é a forma correta de manipular os dados para que se aproximem das estimativas oficiais.

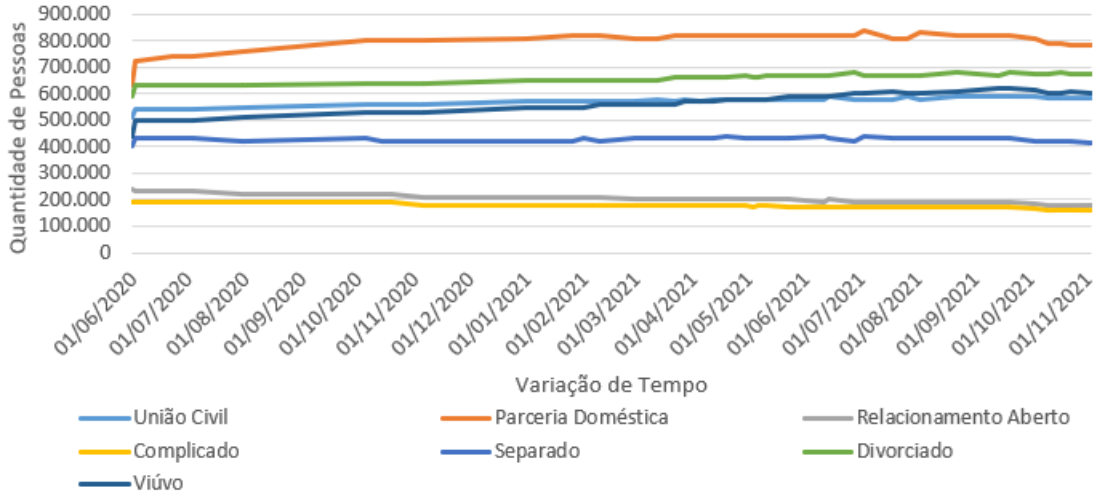
O campo de status de relacionamento não é um campo obrigatório no cadastro de

usuários na rede social, além do fato de existirem várias opções para escolha. Desse modo, os usuários são livres para definir da forma que quiserem qual é o status de relacionamento atual, e mudarem quantas vezes desejarem. Com isso, similarmente como ocorre para análise do Nível de escolaridade, não é possível estimar a quantidade de informação falsa que é inserida pelos usuários da plataforma.

A figura 16 mostra a variação histórica dos subgrupos do atributo relacionamento. Nota-se que os valores não sofreram variações abruptas, e que apesar de variarem, essa variação ocorreu de forma suave e gradativa, o que se aproxima de uma mudança mais natural sem indicar possíveis erros nas estimativas. O fato dos dados variarem pouco, ou de forma gradativa na série histórica não significa que as estimativas são precisas ou corretas. Esse comportamento apenas indica que o dado não sofre variações abruptas, e que caso seja possível encontrar uma correlação de algum desses dados com dados de fontes oficiais, esse monitoramento pode se tornar bastante útil.



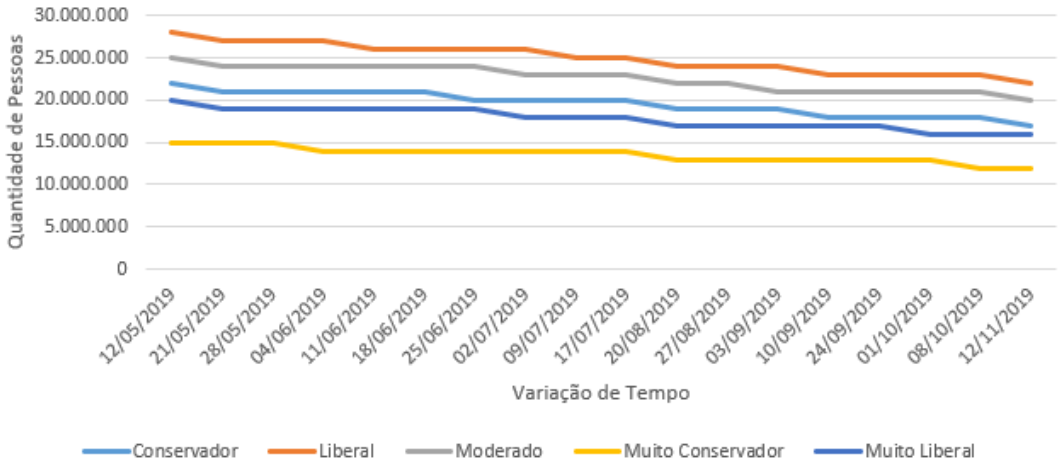

(a) População maior que 1 Milhão



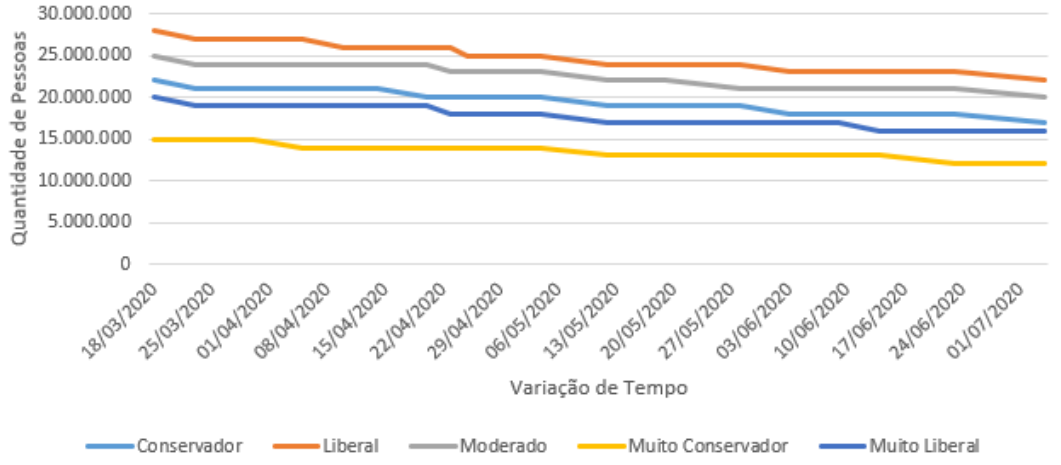

(b) População menor que 1 Milhão

Figura 16 – Variação de relacionamentos para o Brasil, segundo o Facebook

Por fim, na figura 17 é exibida a variação dos subgrupos do atributo Inclinação Política: Muito Conservador, Conservador, Moderado, Liberal e Muito Liberal. É importante destacar que o atributo "Inclinação Política" está disponível apenas para o EUA, portanto não foi possível realizar esta análise para o Brasil. A figura é dividida nos dois períodos em que foram feitas coletas para esse atributo. Em ambas figuras nota-se que não houve variações abruptas repentinas, em que o dado apresenta queda constante para todos os subgrupos. Esse é mais um caso em que não é possível explicar essa variação, por causa da metodologia de caixa preta da plataforma de publicidade do Facebook, no qual apenas a estimativa numérica é apresentada, e não há uma explicação para aqueles números.




(a) 2019




(b) 2020

Figura 17 – Variação da inclinação política do EUA, segundo o Facebook

# 5 Conclusão

Este trabalho pretendeu realizar a coleta de dados demográficos, o monitoramento da variação desses dados ao longo do tempo, e a comparação com dados reais, a partir da plataforma de publicidade do Facebook. A coleta de dados de forma online, além de ser rápida, ser facilmente automatizada e ter um alto poder de monitoramento, também traz a possibilidade de redução de custos na realização do censo. Com um framework que permite explorar a plataforma de publicidade do Facebook a partir API de Marketing (RIBEIRO et al., 2019), foi possível realizar coletas de dados semanais de forma automatizada, em que os atributos para as coletas podem ser escolhidos e definidos conforme desejado.

Primeiramente, foram definidos os atributos a serem coletados para cada uma das regiões analisadas. Com os atributos definidos, foi possível a partir deles, inferir demografia, ou seja, analisar populações humanas e suas características gerais. Com o estudo, foi possível verificar tamanho, distribuição e estrutura das populações selecionadas. A partir da coleta de dados periódica, e com a consolidação dos dados coletados, foi possível visualizar de forma gráfica a variação temporal de qualquer atributo desejado.

A proposta descrita no presente trabalho, de fato permite realizar os passos necessários para o estabelecimento de uma base de dados demográficos e sua utilização em um possível censo, de forma complementar. Entretanto, como foi verificado a partir da análise dos resultados, nem sempre os dados coletados representam (mesmo que apenas em forma percentual, e não bruta), um retrato preciso da realidade. Foi possível observar categorias de dados em que a discrepância entre os dados coletados do Facebook e dados de fontes oficiais eram muito grandes, indicando uma estimativa errônea. De forma similar, a análise da variação temporal dos atributos revelou que, alguns deles, sofrem variações abruptas, em períodos curtos de tempo (de uma semana para a outra), sem nenhum motivo aparente.

Sendo assim, foi observado que, a realização de coletas, monitoramento e comparação com fontes reais, de dados demográficos coletados online, além de ser possível também é uma ferramenta poderosa. Embora ainda seja necessário a validação das informações, foi possível notar diversas oportunidades onde os dados analisados eram bastante similares, dadas as proporções. Em um cenário otimista, é esperado que a informação disponibilizada possa auxiliar em pesquisas do Censo no futuro. Com isso, cria-se a possibilidade de economizar tempo e dinheiro, além de uma nova oportunidade de desenvolvimento acelerado para países em desenvolvimento, onde há um grande deficit de recursos econômicos.

Para trabalhos futuros, acredita-se que expandir a coleta de dados para outras redes sociais, como Twitter e LinkedIn pode ajudar a minimizar a dependência do Facebook. Com essas novas fontes, novas formas de análises podem ser efetuadas, bem como o

cruzamento de informação entre as plataformas para validação. Outras redes sociais como o LinkedIn, por exemplo, trazem uma grande oportunidade de melhora na qualidade dos dados. Isso se verifica, pois por se tratar de uma plataforma para contatos profissionais, as pessoas tendem a inserir informações reais, e terem seu perfil completamente preenchido.

Como contribuição final, todos os dados coletados e utilizados para a realização deste trabalho estarão disponíveis em um repositório online (GitHub). Espera-se que o conjunto de dados possa abrir novos caminhos de pesquisa para aqueles interessados.

# Referências

ACKER, A. H. *Censo Experimental abaixo da expectativa prova fracasso da intervenção no IBGE*. 2019. [Acesso em: 15 maio. 2022.]. Disponível em: <https://assibge.org.br/censo-experimental-abaixo-da-expectativa-prova-fracasso-da-intervencao-no-ibge/>. Citado na página 10.

ACKER, H. *Falta de dinheiro e de pessoal ameaça Censo 2020*. 2019. Acesso em: 15 maio. 2022. Disponível em: <https://assibge.org.br/falta-de-dinheiro-e-de-pessoal-ameaca-censo-2020/>. Citado na página 9.

ARAUJO, M. et al. Using facebook ads audiences for global lifestyle disease surveillance: Promises and limitations. In: *Proceedings of the 2017 ACM on Web Science Conference*. New York, NY, USA: ACM, 2017. (WebSci '17), p. 253–257. ISBN 978-1-4503-4896-6. Disponível em: <http://doi.acm.org/10.1145/3091478.3091513>. Citado na página 14.

ARGAMON, S. et al. Automatically profiling the author of an anonymous text. *Commun. ACM*, Citeseer, v. 52, n. 2, p. 119–123, 2009. Citado na página 14.

CENSUS. *SELECTED POPULATION PROFILE IN THE UNITED STATES*. 2019. [Acesso em: 20 maio. 2022.]. Disponível em: <https://data.census.gov/cedsci/table?q=foreign&g=0100000US&tid=ACSSPP1Y2019.S0201>. Citado 2 vezes nas páginas 22 e 25.

CENSUS, U. S. *2020 Census Life-cycle Cost Estimate Executive Summary*. 2017. [Acesso em: 6 março. 2022.]. Disponível em: <https://www2.census.gov/programs-surveys/decennial/2020/program-management/planning-docs/2020-cost-estimate1.pdf>. Citado na página 9.

CESARE, N. et al. Promises and pitfalls of using digital traces for demographic research. *Demography*, Springer, v. 55, n. 5, p. 1979–1999, 2018. Citado na página 11.

DONG, Y. et al. Inferring user demographics and social strategies in mobile social networks. In: ACM. *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. [S.l.], 2014. p. 15–24. Citado 2 vezes nas páginas 12 e 14.

GAO, G. A. O. *2020 Census: Innovations Helped with Implementation, but Bureau Can Do More to Realize Future Benefits*. 2021. Acesso em: 15 maio. 2022. Disponível em: <https://www.gao.gov/products/gao-21-478#:~:text=With%20the%20decennial%20census%20coming,recent%20estimate%20of%20%2415.6%20billion.> Citado na página 9.

GIL-CLAVEL, S.; ZAGHENI, E. Demographic differentials in facebook usage around the world. In: *Proceedings of the International AAAI Conference on Web and Social Media*. [S.l.: s.n.], 2019. v. 13, p. 647–650. Citado na página 25.

IBGE. *AMOSTRA - EDUCAÇÃO*. 2010. [Acesso em: 15 maio. 2022.]. Disponível em: <https://cidades.ibge.gov.br/brasil/pesquisa/23/22469?detalhes=true>. Citado na página 27.

IBGE. *CENSO - PANORAMA*. 2010. [Acesso em: 15 maio. 2022.]. Disponível em: <https://cidades.ibge.gov.br/brasil/panorama>. Citado 2 vezes nas páginas 22 e 24.

IBGE. *Censo Demográfico - Séries Históricas*. 2010. [Acesso em: 15 maio. 2022.]. Disponível em: <https://www.ibge.gov.br/estatisticas/sociais/populacao/9662-censo-demografico-2010.html?=&t=series-historicas>. Citado na página 10.

IBGE. *Nupcialidade - Amostra Censo 2010*. 2010. Acesso em: 15 maio. 2022. Disponível em: <https://cidades.ibge.gov.br/brasil/pesquisa/23/%2022714?detalhes=true>. Citado 2 vezes nas páginas 22 e 31.

IBGE. *Projeções da População*. 2018. [Acesso em: 15 maio. 2022.]. Disponível em: <https://www.ibge.gov.br/estatisticas/sociais/populacao/9109-projecao-da-populacao.html?=&t=resultados>. Citado 2 vezes nas páginas 21 e 24.

IBGE. *Histórico dos Censos*. 2022. [Acesso em: 15 maio. 2022.]. Disponível em: <https://memoria.ibge.gov.br/historia-do-ibge/historico-dos-censos/censos-demograficos.html>. Citado na página 9.

IBGE. *Materiais/Guia do censo/Contratações*. 2022. Acesso em: 15 maio. 2022. Disponível em: <https://censo2010.ibge.gov.br/materiais/guia-do-censo/contratacoes.html>. Citado na página 10.

IBGE. *Materiais/Guia do censo/Operação censitária*. 2022. Acesso em: 15 maio. 2022. Disponível em: <https://censo2010.ibge.gov.br/materiais/guia-do-censo/operacao-censitaria.html#:~:text=Quanto%20custa%20realizar%20o%20Censo,R%24%201%2C%20677%20bilh%C3%A3o.> Citado na página 9.

JONES, R. et al. I know what you did last summer: query logs and user privacy. In: ACM. *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*. [S.l.], 2007. p. 909–914. Citado na página 14.

JÚNIOR, R. L. I.; RIBEIRO, F. N. Inferring cultural similarity among brazilian states based on data from social media advertising platforms. In: SBC. *Anais do XXVI Simpósio Brasileiro de Sistemas Multimídia e Web*. [S.l.], 2020. p. 204–211. Citado na página 16.

LEITE, V.; SALLES, R. Design e implementação de uma arquitetura conceitual para a criação de social botnet em redes sociais. In: SBC. *Anais do XIX Simpósio Brasileiro de Segurança da Informação e de Sistemas Computacionais*. [S.l.], 2019. p. 337–350. Citado na página 25.

NATIONS, U. *Principles and Recommendations for Population and Housing Censuses, Revision 3*. [s.n.], 2017. 315 p. Disponível em: <https://www.un-ilibrary.org/content/publication/bb3ea73e-en>. Citado na página 9.

OLIVEIRA T; CAVALCANTI, L. M. M. Dados consolidados da imigração no brasil 2022. *Observatório das Migrações Internacionais*, 2022. Ministério da Justiça e Segurança Pública/ Departamento de Migrações [Acesso em: 15 maio. 2022.]. Disponível em: <https://portaldeimigracao.mj.gov.br/images/Obmigra_2020/OBMigra_2022/DADOS_CONSOLIDADOS/Dados_Consolidados_2022.pdf>. Citado na página 22.

OLIVEIRA T; CAVALCANTI, L. M. M. Imigração e refúgio no brasil. relatório anual 2020. *Série Migrações. Observatório das Migrações Internacionais; Ministério da Justiça e Segurança Pública/ Conselho Nacional de Imigração e Coordenação Geral de Imigração Laboral. Brasília, DF*, OBMigra, 2020. Ministério da Justiça e Segurança Pública/ Departamento de Migrações. [Acesso em: 15 maio. 2022.]. Disponível em: <https://portaldeimigracao.mj.gov.br/images/dados/relatorio-anual/2020/OBMigra_RELAT%C3%93RIO_ANUAL_2020.pdf>. Citado 2 vezes nas páginas 22 e 28.

QIU, J. et al. Deepinf: Social influence prediction with deep learning. In: *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*. [S.l.: s.n.], 2018. p. 2110–2119. Citado na página 14.

RAO, D. et al. Classifying latent user attributes in twitter. In: ACM. *Proceedings of the 2nd international workshop on Search and mining user-generated contents*. [S.l.], 2010. p. 37–44. Citado na página 14.

RIBEIRO, F. et al. Media bias monitor: Quantifying biases of social media news outlets at large-scale. In: *Proceedings of the International AAAI Conference on Web and Social Media*. Stanford, USA: [s.n.], 2018. (ICWSM'18). Citado 2 vezes nas páginas 15 e 17.

RIBEIRO, F. N.; BENEVENUTO, F.; ZAGHENI, E. How biased is the population of facebook users? comparing the demographics of facebook users with census data to generate correction factors. In: *12th ACM Conference on Web Science*. [S.l.: s.n.], 2020. p. 325–334. Citado na página 15.

RIBEIRO, F. N.; KANSAON, D.; BENEVENUTO, F. Leveraging the facebook ads platform for election polling. 2019. Citado na página 16.

RIBEIRO, F. N. et al. Inference of demographic data from digital advertising platforms based on social media. Universidade Federal de Minas Gerais, 2019. Citado na página 34.

SAP, M. et al. Developing age and gender predictive lexica over social media. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. [S.l.: s.n.], 2014. p. 1146–1151. Citado 2 vezes nas páginas 12 e 14.

SIMÕES A; HALLAK NETO, J. C. L. O. T. M. M. Relatório conjuntural: tendências da imigração e refúgio no brasil, 1º quadrimestre/2020. *Observatório das Migrações Internacionais; Ministério da Justiça e Segurança Pública/ Coordenação Geral de Imigração Laboral. Brasília, DF - OBMigra*, 2020. [Acesso em: 15 maio. 2022.]. Disponível em: <https://portaldeimigracao.mj.gov.br/images/dados/relatorios_conjunturais/2020/Novo_Conjuntural_1QUAD_1.pdf>. Citado 2 vezes nas páginas 22 e 28.

SPEICHER, T. et al. On the Potential for Discrimination in Online Targeted Advertising. In: *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT*'18)*. [S.l.: s.n.], 2018. Citado 2 vezes nas páginas 14 e 16.

ZAGHENI, E.; WEBER, I.; GUMMADI, K. Leveraging facebook's advertising platform to monitor stocks of migrants. *Population and Development Review*, JSTOR, p. 721–734, 2017. Citado na página 15.