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

Junho, 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

Junho de 2022

A Ficha Catalográfica é elaborada exclusivamente pela Biblioteca. Substitua esta página pelo documento gerado na versão final da sua monografia.

A Folha de aprovação deverá ser gerada pelo(a) professor(a) orientador(a) após a realização das correções sugeridas pela banca examinadora.

As instruções para elaboração desse documento eletrônico estão disponíveis em <https://www.monografias.ufop.br> (menu “Documentos”, opção “Tutorial Professor Orientador”).

Após gerar a folha de aprovação, o(a) professor orientador(a) enviará o documento ao(à) orientado(a), o qual deverá inseri-lo nesta página.

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 utilizar das plataformas de publicidades das redes sociais, que vem se mostrando cada vez mais tecnológicas e precisas, como forma de a 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 o desenvolvimento de uma espécie de censo demográfico com um monitoramento da variação temporal, a partir de dados coletados da plataforma de propagandas do Facebook. 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

This is the english abstract.

Key-words: latex. abntex. text editoration.

Lista de ilustrações

Lista de tabelas

Sumário

1 Introdução

1.1 Contextualização e definição do problema abordado

Um constante crescimento das redes sociais vem 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.

Analisando uma determinada rede social mobile, segundo (DONG et al., 2014), é possível descobrir padrões de comportamento que os usuários utilizam para manter suas conexões. Segundo o autor, pessoas de menor faixa etária são mais prováveis de serem ativas em aumentar seus círculos sociais, ao passo que pessoas de maior faixa etária procuram manter conexões apenas com pessoas próximas, e se manterem mais estáveis.

Utilizando-se de texto, publicamente disponível, extraídos também de redes sociais como *Facebook* e *Twitter*, foi possível determinar com um alto nível de precisão a faixa etária e o gênero dos autores(SAP et al., 2014). Adicionalmente, pessoas também utilizaram de informações disponíveis na rede para inferir atributos como idioma nativo(ARGAMON et al., 2009a), origem(RAO et al., 2010a) e localização(JONES et al., 2007).

No último ano, diversas plataformas de publicidade online sofreram criticas pelo fato de admitirem que anunciantes discriminam usuários, proibindo determinados grupos de raça, ou gênero de receber seus anúncios. Um anunciante pode criar propagandas altamente discriminatórias, sem a necessidade de utilizar atributos sensíveis (SPEICHER et al., 2018). 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.

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.

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).

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

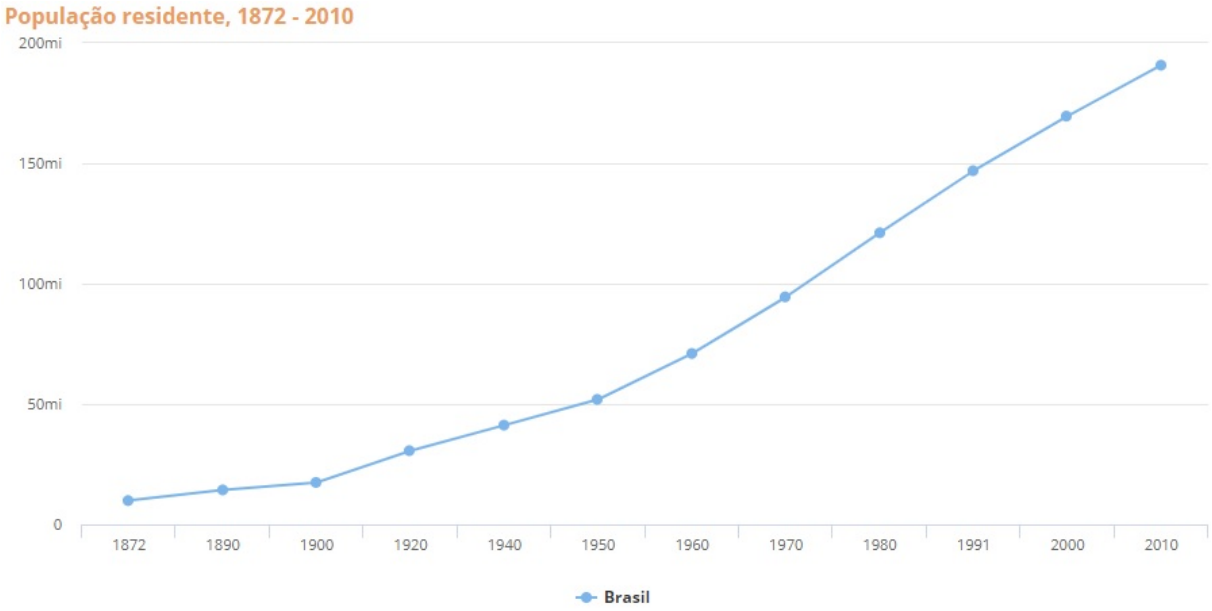


Figura 1 – Censo demográfico no Brasil ao longo dos anos.

Os censos, realizados pelo mundo inteiro, são uma fonte chave de dados que guiam investimentos governamentais e políticas públicas. Realizados durante séculos, são altamente necessários na sociedade moderna, além de serem cruciais para a definição da prioridade de investimentos para educação, infraestrutura e outras políticas públicas do país.

O Censo é uma grande representação em extensão e profundidade da população analisada e de suas características socioeconômicas ao mesmo passo que serve como referência para ser utilizada como base para o planejamento público e privado para os próximos anos.

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 e em 2020 o valor estimado

chegava a um total de R$3,4 bilhões. Nos Estados Unidos esse valor chega a $15 bilhões de dólares. A maior parcela desses valores, é destinada a contratação de cerca de 240 mil funcionários temporários, e essa é uma das características que impedem a redução significativa de orçamentos da pesquisa.

Atualmente, 65% da população nacional possui acesso à internet, porém, 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 2,5% da amostra responderam através dessa opção. Além do fato de não possuírem uma atualização tão constante quanto o desejado, tendo em vista que os censos são decenais.

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 propaganda, a qual representa a principal fonte de lucro destas empresas.

Com o proposto, pretende-se complementar as estatísticas oficiais, oferecendo um forte poder de monitoramento na variação temporal dos dados demográficos de forma fácil e eficiente além de estimativas oportunas entre os censos.

1.2 Estado da arte

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., 2009a), origem (RAO et al., 2010a) 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).

Os serviços e percepções úteis que podem ser aproveitadas usando dados demográficos não se limitam a recomendar páginas da Web adequadas ao perfil do usuário. Esforços nessa área incluem estudos que tentam inferir a tendência política de usuários de Redes Sociais Online (OSN) e detectar gênero para auxiliar investigações forenses. Em particular, alguns estudos recentes exploraram plataformas de publicidade OSN para inferir

dados demográficos de informações agregadas sobre os usuários. Esses tipos de plataformas contam com uma rica fonte de dados dos usuários, como local de trabalho, locais visitados, postagens publicadas e "curtidas", para inferir as características demográficas dos usuários em um nível refinado (RIBEIRO; BENEVENUTO; ZAGHENI, 2020).

1.3 Gap e solução proposta

Com os Censos necessitando de muito tempo, esforços e recursos em grande quantidade, países em desenvolvimento podem enfrentar dificuldade em sua realização. Além de que, os censos são realizados a cada 10 anos, o que faz com que a informação não esteja sempre atualizada.

O presente trabalho trata da elaboração de uma espécie de Censo demográfico, com dados coletados a partir da plataforma de propaganda do Facebook. A coleta de dados, feita de forma online, garante um alto poder de monitoramento dos dados coletados, possibilitando análises mais frequentes.

Acredita-se que, 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.

Com o proposto, é explorado uma maneira eficiente de obter conjuntos de dados fornecidos a partir das plataformas de publicidade do Facebook, e inferir demografia a partir deles. Com isso, espera-se criar a possibilidade de economia de tempo, e recursos financeiros em futuros períodos de recenseamento.

1.4 Justificativa

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.

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 senso 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.

1.5 Objetivos gerais

No presente trabalho, tem-se como objetivo a reprodução de uma espécie de censo demográfico, utilizando dados coletados a partir de uma rede social, o Facebook.

Para a base de dados coletados que são direcionados para pessoas que moram nos 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 a base de dados coletados que são direcionados para pessoas que moram no Brasil, foram coletados 6 tipos de atributos que são: *Gênero, idade, nível educacional, status de relacionamento, religião e imigrantes*, esses atributos possuem subgrupos também mostrados na tabela 1.

Lista de interesses			
Religião	Raça	Gênero	Idade
Católico, Protestante, Espirita, Ateu	Hispânico, Africano, Asiático, Caucasiano	Masculino Feminino	13 - 17; 18 - 24; 25 - 34; 34 - 44; 45 - 54; 55 - 64; 65+

Tabela 1 – Exemplo de subgrupos para atributos primários.

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.

1.5.1 Objetivos específicos

- Definir interesses a serem coletados.
- Inferir demografia a partir de dados online.
- Realizar a coleta dos dados periodicamente.
- Analisar a variação sob o tempo.
- Comparar resultados com fontes oficiais.
- Publicar os resultados e disponibilizar os dados.

1.6 Organização do trabalho

Este trabalho

2 Revisão bibliográfica

Para início do trabalho, foi realizada uma revisão bibliográfica para compreender as metodologias de inferências demográficas a partir de dados coletados online, ao mesmo tempo que foi necessário um estudo do pacote *facebook_business* desenvolvido para a linguagem *python*, 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 o surgimento dos portais de mídia social, há uma abundância de conteúdo gerado pelo usuário na web. Além disso, a popularidade de sites como Twitter, MySpace e Facebook está crescendo ininterruptamente. O Twitter sozinho tem mais de 10 milhões de usuários globais, em comparação com cerca de 6 milhões do Facebook (O'NEILL, 2009). Os usuários e comunidades nesses sites têm um amplo alcance para outros usuários dessas plataformas. O principal usuário do Twitter atinge cerca de 3,2 por cento da base total de usuários.

Segundo o Rao et al. (2010b) isso tem consequências importantes na publicidade direcionada e na personalização. No entanto, ao contrário do Facebook ou do MySpace, o Twitter tem metadados limitados disponíveis sobre seus usuários. Atributos importantes do usuário, como idade e sexo, que são diretamente úteis para fornecer serviços personalizados, geralmente não estão disponíveis. Além disso, pode-se estar interessado em conhecer outros atributos do usuário, como etnia, opiniões e outras propriedades e preferências que o usuário pode não divulgar.

Rao et al. (2010b) investigam e avaliam o desempenho em três atributos biográficos do autor - gênero, idade e origem regional - e um atributo de personalização - orientação política. Isso é limitado apenas por nosso acesso atual a dados de verdade para avaliação e pode-se aprender uma grande variedade de outros atributos, incluindo preferências alimentares (vegetariano vs não vegetariano), orientação sexual (homossexual vs heterossexual), status de estudante (estudante vs não estudante) e assim por diante, restringido apenas pelos dados de treinamento e avaliação disponíveis.

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. Rao et al. (2010b) ignoram ainda todos os candidatos com uma alta contagem de seguidores, visto que atendem a celebridades ou organizações. Cada um dos candidatos restantes é anotado manualmente por dois anotadores independentes. Os anotadores não foram expostos aos métodos experimentais utilizados ou às abordagens de modelagem adotadas. Para evitar um problema de viés de

etiqueta, restringiram o número de usuários em cada classe para ser semelhante.

A pesquisa em sociolinguística apresentada por Rao et al. (2010b) explorou os efeitos do gênero, idade, classe social, religião, educação e outros atributos do orador no discurso coloquial e no monólogo. Os primeiros trabalhos nesta área envolveram o estudo de características morfológicas e fonológicas, respectivamente. De todos os atributos, o gênero foi amplamente estudado, ao que tudo indica por causa de suas implicações antropológicas. Os primeiros modelos computacionais para detectar gênero a partir do texto estavam principalmente interessados em texto formal.

Ao contrário do problema de atribuição de autoria (determinar o autor de um texto a partir de um determinado conjunto de candidatos), o perfil de autoria, mostrado por Argamon et al. (2009b), não tem início com um conjunto de amostras de escrita de autores candidatos. Em vez disso, os autores trazem a observação sociolinguística de que diferentes grupos de pessoas falando ou escrevendo em um determinado gênero e em um determinado idioma usam essa linguagem de maneira diferente. Ou seja, eles variam na frequência que eles usam certas palavras ou construções sintáticas. As dimensões particulares do perfil que os autores consideram são gênero do autor , idade, idioma nativo e personalidade.

Argamon et al. (2009b) mostram uma abordagem para o perfil de autoria que é aplicar o aprendizado de máquina ao texto de categorização. O processo é o seguinte:

- Primeiro, os autores pegam um determinado corpus de documentos de treinamento, cada um rotulado de acordo com sua categoria para uma dimensão de perfil particular.
- Cada documento é então processado para produzir um vetor numérico, cada um dos elementos representa algum recurso do texto que pode ajudar a discriminar as categorias relevantes.
- Um método de aprendizado de máquina, então, calcula um classificador que, na medida do possível, classifica os exemplos de treinamento corretamente.
- Por fim, o poder preditivo do classificador é testado em dados fora do treinamento.

Existem dois tipos básicos de recursos que podem ser usados para o perfil de autoria: recursos baseados em conteúdo e recursos baseados em estilo. Este reflete o fato de que diferentes populações podem tender a escrever sobre diferentes tópicos bem como se expressarem de maneira diferente sobre o mesmo assunto. Argamon et al. (2009b) trazem que devemos considerar esses tipos separadamente, começando com recursos baseados em estilo.

Deve-se observar que o uso de recursos baseados em conteúdo para estudos de autoria pode ser problemático. Embora seja plausível que os marcadores baseados em estilo possam realmente distinguir uma classe de autores de outra, deve-se ter cuidado, pois os

marcadores de conteúdo podem ser artefatos de uma situação de escrita particular ou configuração experimental e podem, assim, produzir resultados excessivamente otimistas que não serão confirmados em aplicativos da vida real (ARGAMON et al., 2009b). Devemos portanto,ter cuidado para distinguir os resultados que exploram recursos baseados em conteúdo daqueles que não são.

Os censos têm sido usados por muitos séculos para avaliar as quantidades demográficas. São necessários e de extrema importância para o funcionamento ordenado das sociedades modernas. Os censos são fundamentais para definir investimentos prioritários em educação, infraestrutura e outras políticas públicas. Em países como os EUA, a coleta de dados por meio de censos é determinada pela Constituição. Os censos são necessários, mas o custo e o tempo necessários para fazer um censo da população são bastante elevados. Um relatório recente publicado pelo U.S. Census Bureau estima que o custo esperado para o censo decenal de 2020 é de 15 bilhões de dólares (CENSUS, 2017).

Formas complementares de coleta de dados para censos foram testadas por diversos países. Na Noruega, por exemplo, as autoridades realizaram o Censo com uma abordagem baseada em registro, que usa informações de uma fonte administrativa existente e coleta informações sobre famílias, moradias e indivíduos para complementar os dados sobre as características demográficas da população (RIBEIRO; BENEVENUTO; ZAGHENI, 2020).

Os autores recolheram estimativas das características demográficas dos utilizadores do Facebook através da plataforma de publicidade do Facebook, nomeadamente Anúncios no Facebook. Em particular, analisaram sete categorias demográficas coletadas por meio da plataforma de publicidade: gênero, raça, idade, renda, educação, tendência política e país de residência anterior - e as compararam com estatísticas oficiais. Os resultados apresentados pelos autores mostram que parte dos dados demográficos extraídos dos Anúncios do Facebook são bastante semelhantes aos dados oficiais, notadamente no que diz respeito à raça, inclinação política e nível de educação de pós-graduação. Para as categorias em que os dados online se desviam das estatísticas oficiais, avaliaram o quanto os grupos demográficos online estão mais ou menos representados no Facebook e calcularam os fatores de correção.

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 propaganda conseguiram direcionar as propagandas para nichos específicos de usuários considerando certos tipos de atributos como nome, aspectos demográficos, comportamentos, entre outros.

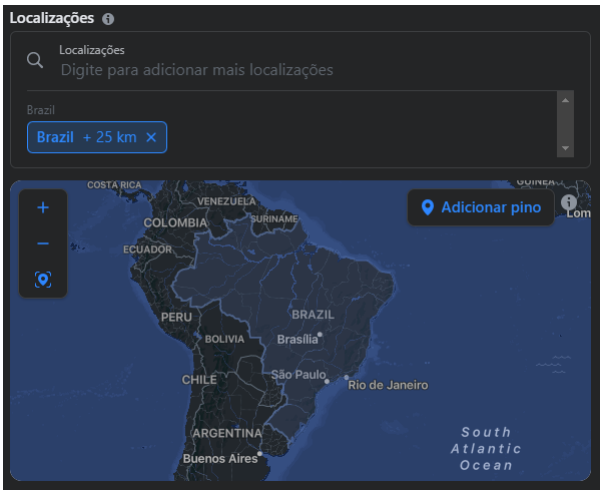
As plataformas de propaganda providenciam três maneiras para definir a audiência de determinado anúncio, *Personally Identifiable Information (PII) targeting*, *Look-alike audience targeting* e *Attribute-based targeting*. Para a automatização da coleta de dados utilizaremos a API de Marketing do Facebook, disponibilizada na linguagem *Python*.

Nos últimos anos, as plataformas de propaganda em diversas redes sociais mostraram uma grande evolução, com isso surgiram novas formas de se definir a audiência a qual uma determinada propaganda será direcionada, como citado anteriormente. No presente trabalho, aproveitaremos 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 interesses para definir a audiência alvo.

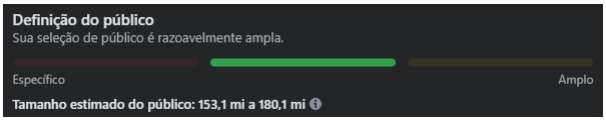
As plataformas de publicidade do facebook fornecem algumas maneiras de definir o público que um anúncio deve atingir, a maneira escolhida para este trabalho é a *Attribute-based targeting* (Segmentação baseada em atributos). Na segmentação baseada em atributos: o anunciante pode definir o público-alvo com base em uma série de atributos que variam desde os interesses do usuário até o dispositivo que ele usa ao acessar o facebook. Atributos estes como:

- Dados demográficos básicos que vão incluir o que o indivíduo possui de mais básico, 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 são os dados básicos, bem como universidade em que o indivíduo frequentou até qual o seu nível de renda.

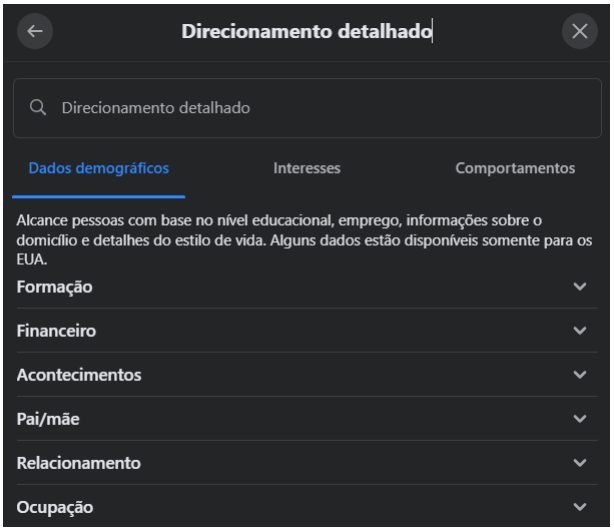
- E por fim as conexões é neste momento em que o anunciante poderá filtrar qual é o seu público-alvo, ele poderá adicionar ou excluir indivíduos de sua página.



(a) Localização



(b) Alcance



(c) Direcionamento detalhado

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

Foi utilizado uma fórmula simples 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, foi possível incluir uma nova camada de seleção de atributos com intenção de nos permitir estratificar os dados de uma forma abrangente.

Figura 2 representa um exemplo de como escolher uma audiência alvo através da plataforma de propaganda do Facebook. Adicionalmente, podemos obter as quantidades de pessoas 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.

Uma simples fórmula foi utilizada para representar diferentes subpopulações das quais utilizaremos para inferir os dados demográfico, como por exemplo, religião. Por fim, a audiência de pessoas que vivem no Brasil foi dividida em quatro subgrupos diferente de religião: Católico (Xc), Protestante (Xp), Espírita (Xe) e Ateu (Xa).

A partir disso aplica-se a formula para calcular a porcentagem de qualquer atributo que desejarmos, como por exemplo, a porcentagem de usuários que se reconhecem como católicos.

$$pc = \frac{Xc}{Xc + Xp + Xe + Xa}$$

Uma formula geral para calcularmos a porcentagem (pe) de uma população específica pertencente ao subgrupo (Xs), é descrita como:

$$porcentagem(pe) = \frac{Xs}{\sum_{i=1}^{n} Xi}$$

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 propaganda que um negócio pode rodar. Apesar de esforços para que as estimativas sejam precisas, as estimativas não são desenhadas para corresponder com as estimativas 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.

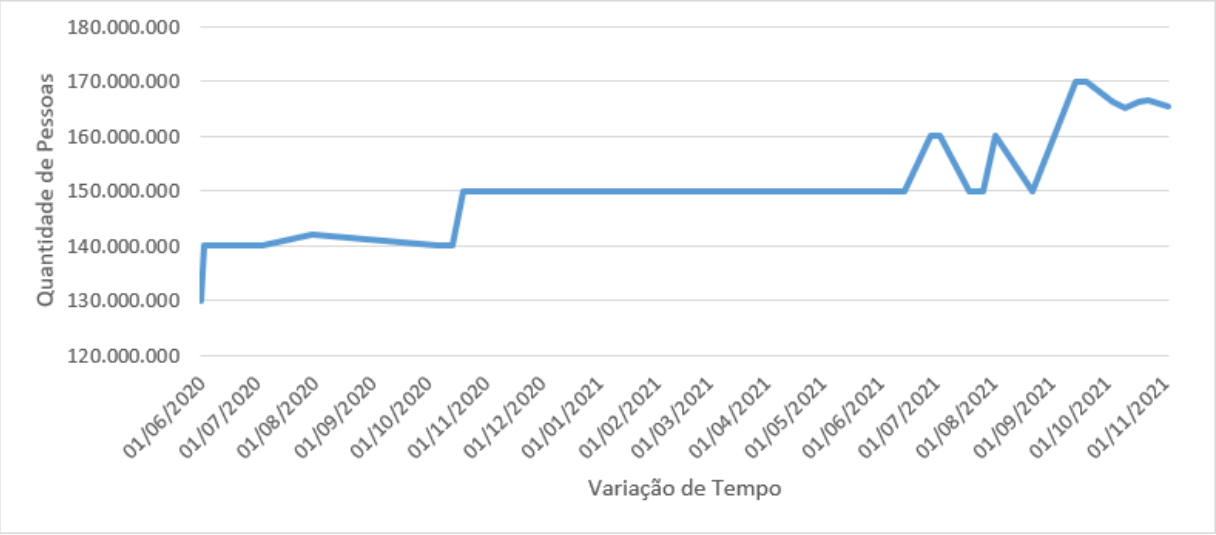
Outro ponto importante, é o fato do Facebook possuir um mecanismo que proteja a identidade dos seus usuários. Em outras palavras, caso os atributos selecionados no momento da pesquisa, correspondam a uma subpopulação menor do que mil usuários, o valor retornado será sempre 1000. Apesar desse mecanismo de proteção a 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 estimular a quantidade de informação falsa sobre as características informadas pelos próprios usuários em seu perfil. O próprio Facebook faz esse tipo de tratamento, na busca de uma maior confiabilidade em sua plataforma de propaganda.

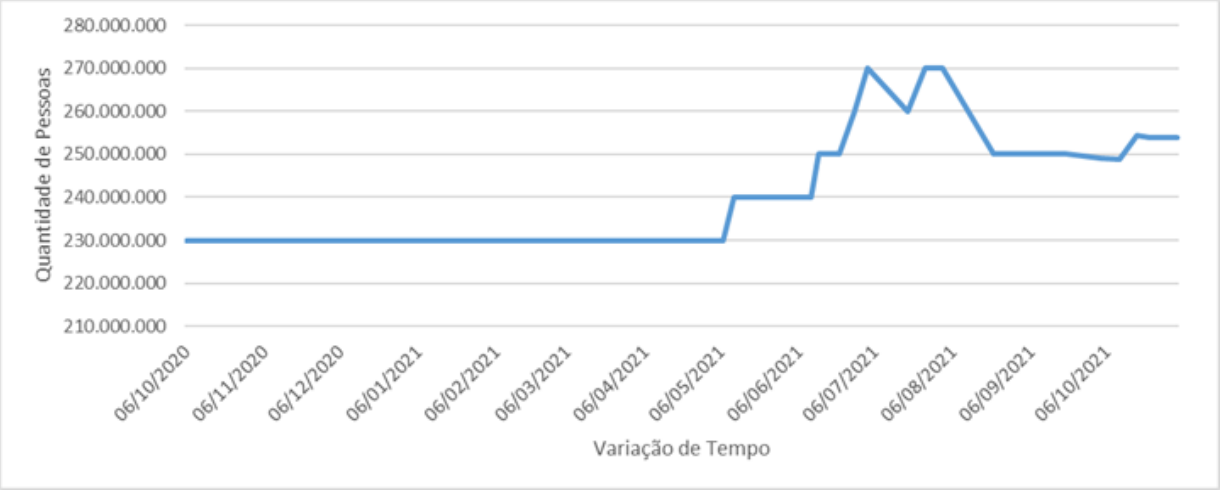
4 Resultados

Nesta seção, tem-se como objetivo a visualização dos dados coletados através da plataforma de propaganda 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 de usuários do Facebook como por exemplo é mostrado na figura 3.



(a) Brasil



(b) EUA

Figura 3 – Variação da população, segundo coleta de dados no Facebook.

Alguns dados oficiais, como por exemplo o tamanho da população estimada, é

divulgado no do IBGE, juntamente com os dados do ultimo 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 4.

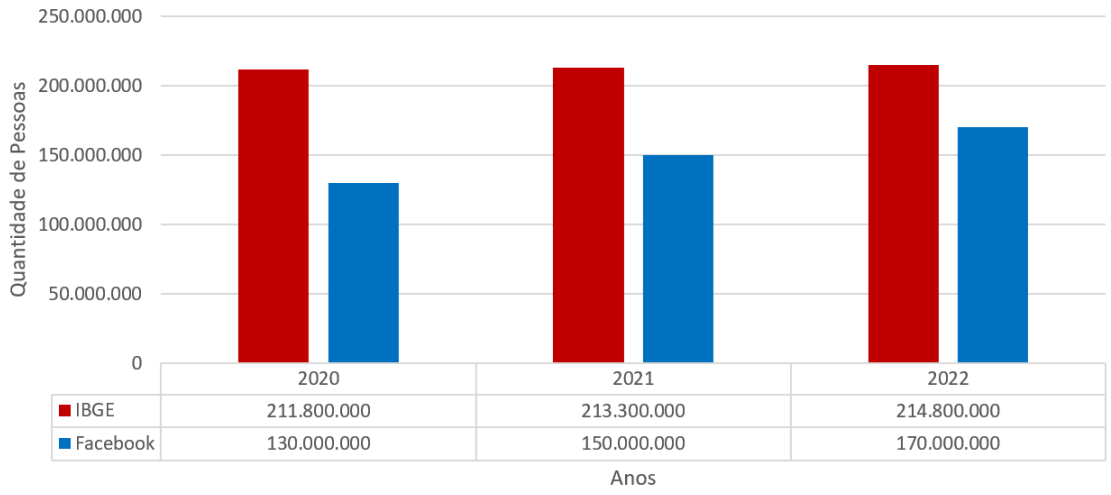
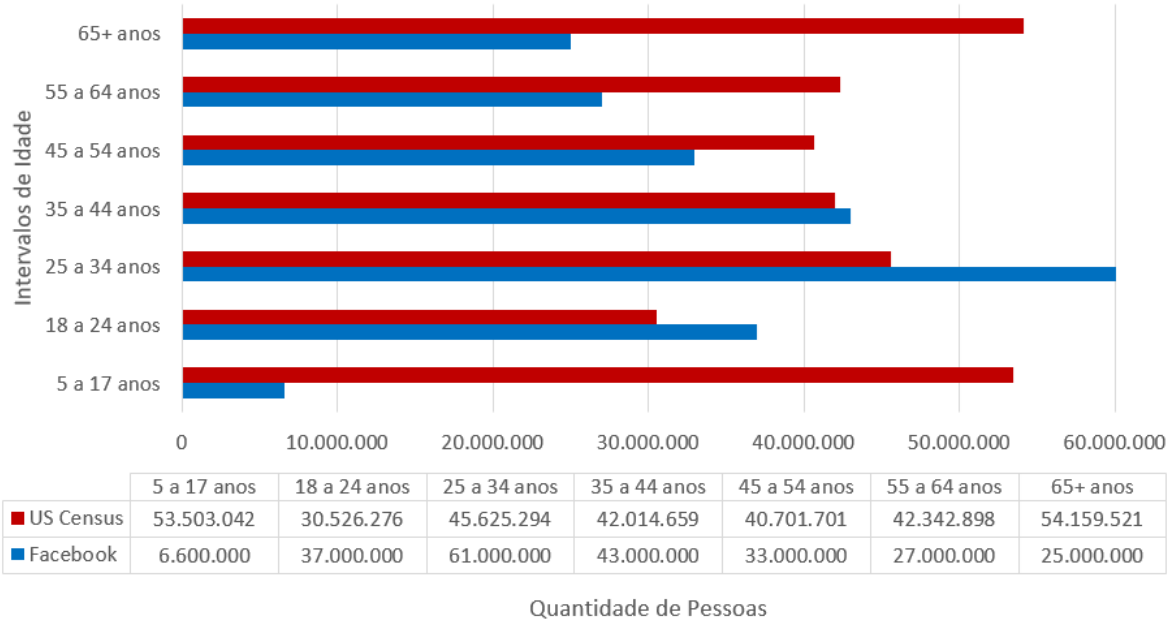


Figura 4 – Censo oficial Versus Facebook

A partir da figura 4, 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. Esse crescimento pode ter como um dos fatores influenciadores, a pandemia do Covid-19.

Além de uma analise 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 5, por exemplo. O Facebook possui aproximadamente 260 milhões de usuários que residem no EUA. 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. O mesmo comportamento pode ser observado 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 18 e 34 anos, o Facebook possui

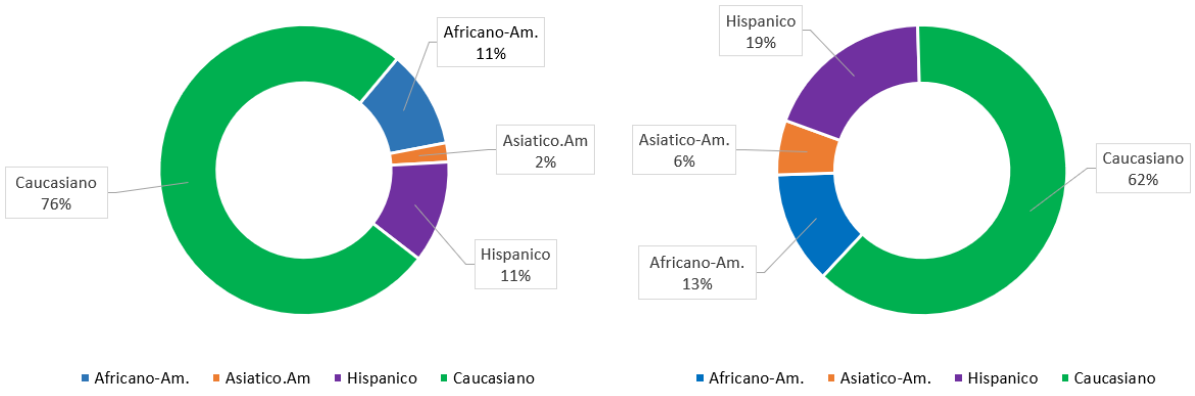
uma população maior do que a indicada pelo censo oficial. Novamente, esse resultado pode ser um reflexo da criação de múltiplas contas por alguns usuários, incluindo contas comerciais, porém não é possível afirmar com exatidão.



	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
US Census	53.503.042	30.526.276	45.625.294	42.014.659	40.701.701	42.342.898	54.159.521
Facebook	6.600.000	37.000.000	61.000.000	43.000.000	33.000.000	27.000.000	25.000.000

Figura 5 – População agrupada por idade no EUA

De forma similar, é possível realizar a mesma análise para a distribuição racial e étnica para a população do EUA (CENSUS, 2019). Na figura 6, é possível visualizar lado a lado a porcentagem de pessoas em cada grupo étnico ou racial. A plataforma de propaganda 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 6 – 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 estendemos para todas as coletas, visando observar a variação desses dados, encontramos o resultado mostrado na figura 7

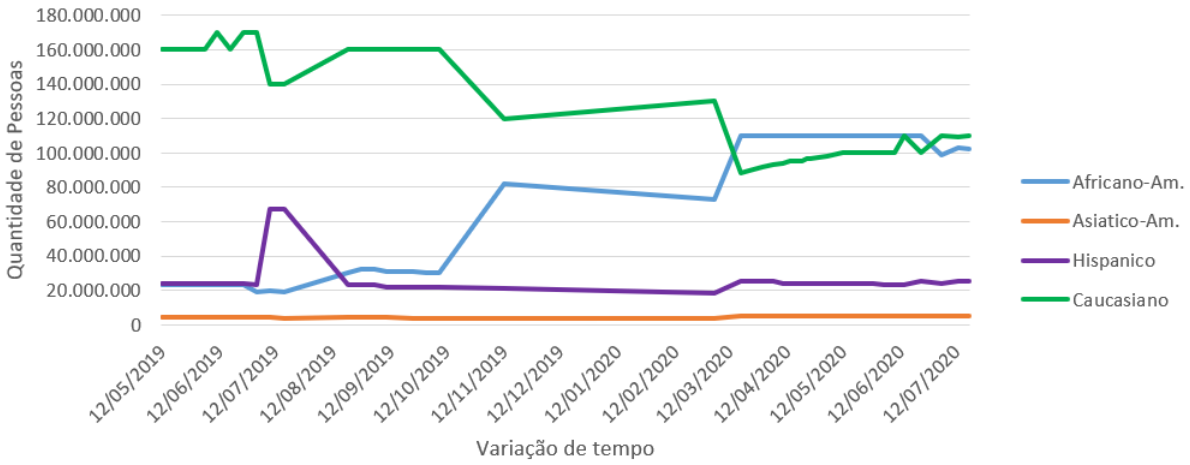


Figura 7 – Variação da distribuição racial sobre o tempo, com base em dados do Facebook.

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 calculo 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.

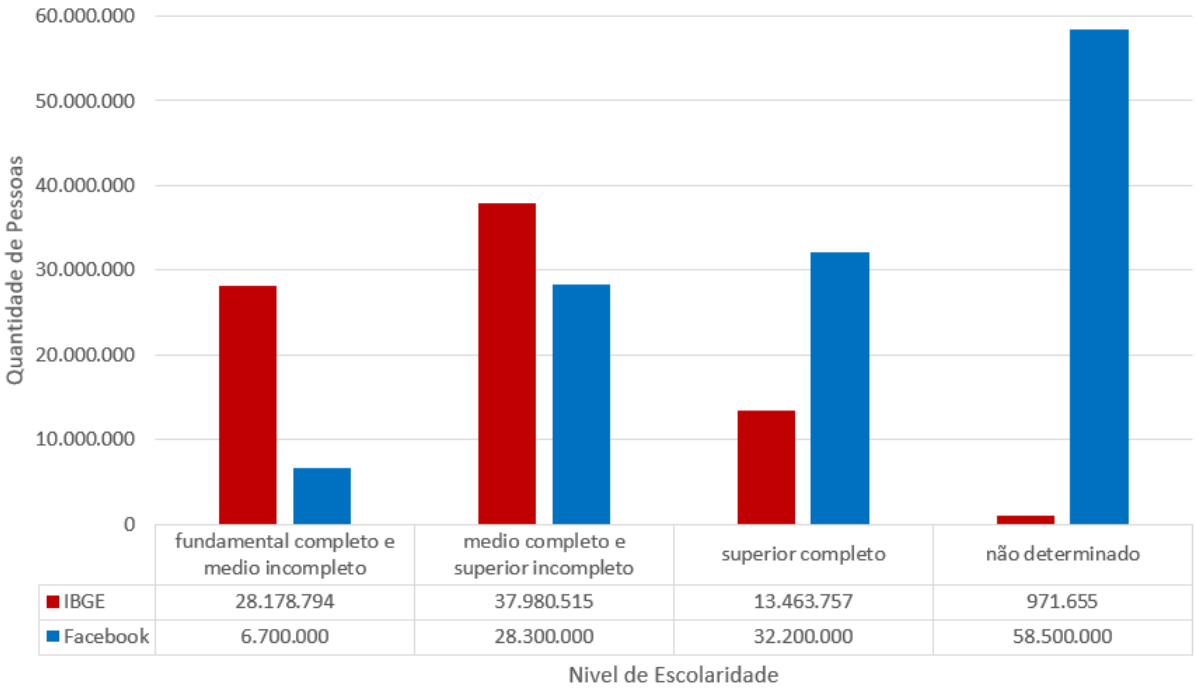


Figura 8 – Distribuição educacional no Brasil

Assim com 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 propaganda Facebook (coleta de novembro de 2021) com o censo realizado em 2010 pelo IBGE (IBGE, 2010a). A figura 8 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.

...EM FINALIZAÇÃO... IMIGRANTES:

A pandemia de COVID-19 implicou na maior redução dos movimentos de entrada e saída do país na década (CITE).

2019: 181.584 imigrantes registrados. 2020: 92.544 imigrantes registrados

Figura 9, mostra a variação de imigrantes no Brasil.

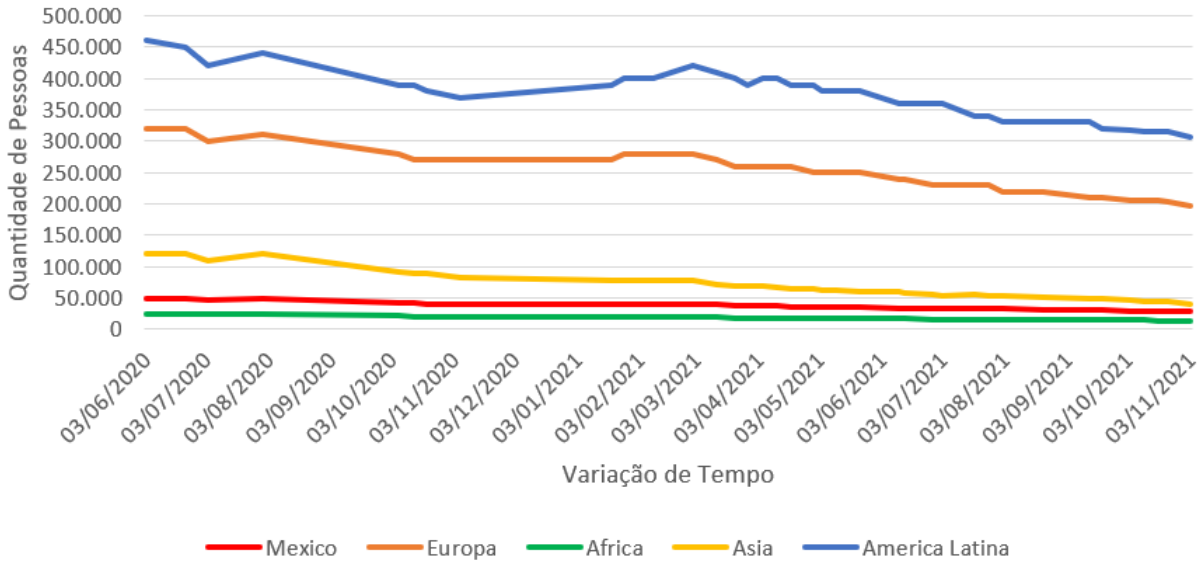


Figura 9 – Variação de imigrantes para o Brasil, segundo o Facebook

Figura 10, mostra a variação de imigrantes no EUA.

Expatriados porcentagem 11

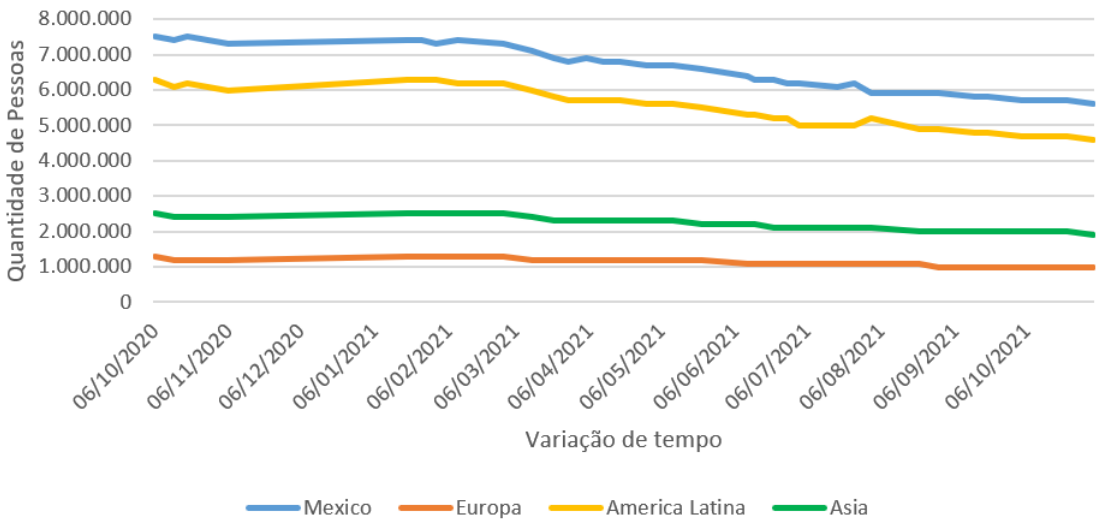


Figura 10 – Variação de imigrantes para o Brasil, segundo o Facebook

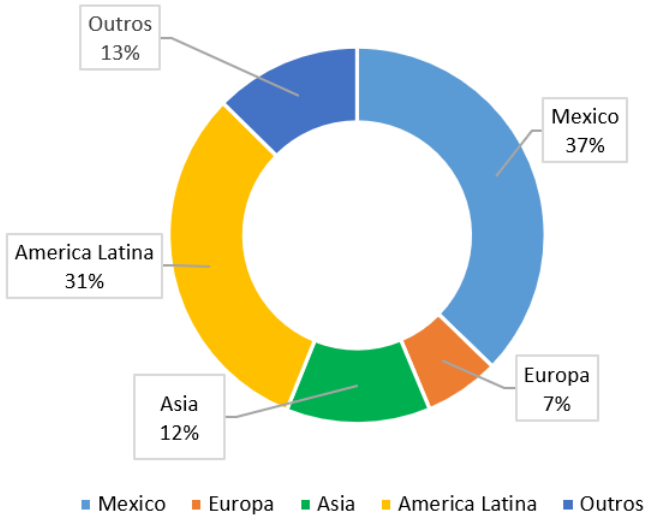


Figura 11 – Variação de Expatriados para o EUA

5 Conclusão

Este trabalho pretendeu desenvolver uma espécie de censo demográfico, com dados coletados a partir da plataforma de propaganda 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 propaganda 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. 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 um censo demográfico a partir de dados coletados online, além de ser possível também é uma ferramenta poderosa, em termos de monitoramento. Embora ainda seja necessário a validação das informações, foi possível verificar 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

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 12.

ARGAMON, S. et al. Automatically profiling the author of an anonymous text. *Commun. ACM*, Citeseer, v. 52, n. 2, p. 119–123, 2009. Citado 2 vezes nas páginas 10 e 12.

ARGAMON, S. et al. Automatically profiling the author of an anonymous text. *Communications of the ACM*, ACM New York, NY, USA, v. 52, n. 2, p. 119–123, 2009. Citado 2 vezes nas páginas 17 e 18.

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 na página 24.

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 18.

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 10.

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 10 e 12.

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 23.

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 26.

IBGE. *CENSO - PANORAMA*. 2010. [Acesso em: 15 maio. 2022.]. Disponível em: <https://cidades.ibge.gov.br/brasil/panorama>. Citado na página 23.

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 na página 23.

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 2 vezes nas páginas 10 e 12.

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 11.

O'NEILL, N. *Twitter Versus Facebook, A Comparison Of The Top Users*. 2009. [Acesso em: 16 fevereiro. 2022.]. Disponível em: <http://www.allfacebook.com/2009/03/twitter-facebook-comparison/>. Citado na página 16.

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 2 vezes nas páginas 10 e 12.

RAO, D. et al. Classifying latent user attributes in twitter. In: *Proceedings of the 2nd international workshop on Search and mining user-generated contents*. [S.l.: s.n.], 2010. p. 37–44. Citado 2 vezes nas páginas 16 e 17.

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 na página 19.

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 2 vezes nas páginas 13 e 18.

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 28.

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 10 e 12.

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 10 e 12.