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

Innan Plínio Rangel Amorim[1], **Filipe Nunes Ribeiro**[1]

[1]Departamento de Computação e Sistemas de Informação
Instituto de Ciências Exatas e Aplicadas
Universidade Federal de Ouro Preto (UFOP) – João Monlevade, MG – Brasil

innan.amorim@aluno.ufop.edu.br, filipe.ribeiro@ufop.edu.br

Resumo. *Este trabalho 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.*

1. Introdução

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].

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.

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, tendo início no ano de 1890, falhando apenas nos anos 1910 e 1930 onde 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.

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.

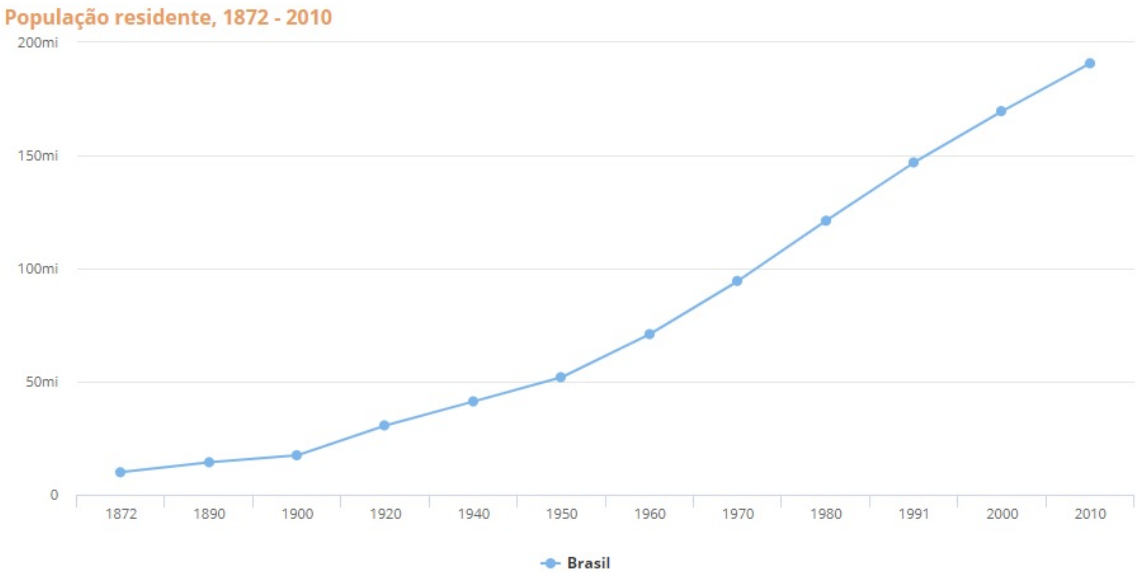


Figura 1. Censo demográfico no Brasil ao longo dos anos.

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 possui um total necessário 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, onde 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.

2. Objetivos propostos

2.1. Objetivos gerais

No presente trabalho, objetivamos a reprodução de uma espécie de censo demográfico, utilizando dados coletados a partir de uma rede social, para nosso estudo utilizamos o Facebook. 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*.

Em nossa base de dados de coletas direcionada para pessoas que moram nos Estados Unidos, incluímos 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), entretanto para o Brasil alguns desses atributos como *Inclinação política*, são tratados de forma diferente, portanto será necessário fazer uma avaliação prévia de quais atributos serão utilizado para a coleta de dados no caso do Brasil, onde inclusive temos a intenção de fazer a análise em um nível mais detalhado, considerando-se os estados.

Lista de interesses		
Inclinação Política	Raça	Gênero
Muito Liberal, Liberal, Moderado, Conservativo, Muito Conservativo	Hispânico Africano Asiático Outros	Masculino Feminino

Tabela 1. Exemplo de subgrupos para atributos primários.

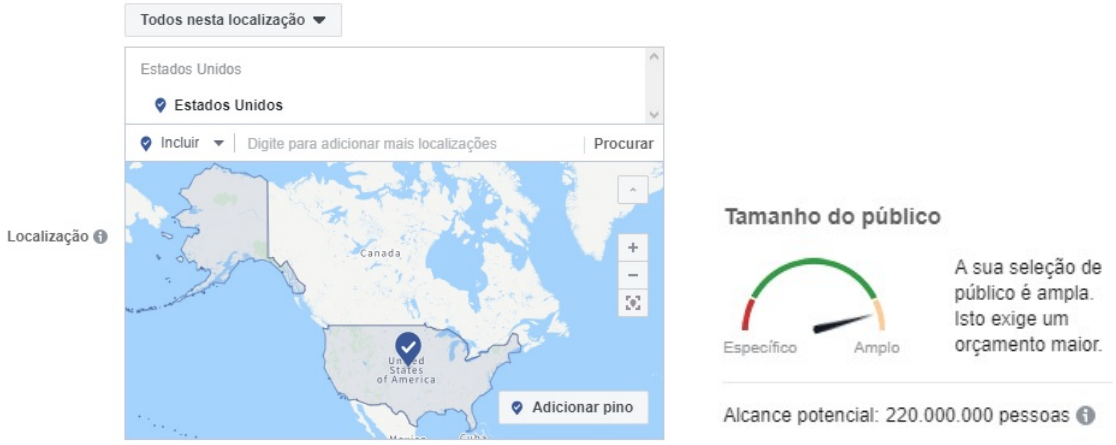
Como proposto, serão feitas coletas de dados periódicas, onde a partir desses dados, será feito uma análise da variação sob o tempo. Os resultados obtidos, bem como os dados coletados, serão disponibilizados através de alguma plataforma para que possa ser facilmente utilizada em um trabalho futuro ou a pesquisadores da área.

2.2. 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.

3. Metodologia

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



(a) Localização (b) Alcance

Figura 2. Exemplo de publicidade direcionada para usuários que moram nos Estados Unidos para a plataforma de propaganda do Facebook

[Ribeiro et al. 2018], onde foi utilizada a *Attribute-based targeting*, em que o anunciante escolhe entre uma variedade de interesses para definir a audiência alvo.

Utilizamos uma fórmula simples que seleciona todos usuários do Facebook que moram nos Estados Unidos. Dessa forma obtemos um potencial de aproximadamente 220 milhões de usuários de ambos gêneros e com idade acima de 13 anos (idade mínima permitida pelo Facebook). A partir disso, conseguimos 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 EUA também são conservativos ou liberais por exemplo, obtendo então as seguintes quantias respectivas: 20 milhões e 26 milhões aproximadamente.

Tomamos como base essa simples fórmula para representarmos diferentes subpopulações das quais utilizaremos para inferir os dados demográfico, como por exemplo, inclinação política. Por fim, conseguimos dividir a audiência de pessoas que vivem nos Estados Unidos em cinco categorias diferente de alinhamento político: muito conservativo (Xmc), conservativo (Xc), moderado (Xm), liberal (Xl) e muito liberal (Xml).

A partir disso podemos aplicar a formula para calcular a porcentagem de qualquer atributo que desejarmos, como por exemplo, a porcentagem de usuários que possuem inclinação política definida como conservativa.

$$pc = \frac{Xc}{Xmc + Xc + Xm + Xl + Xml}$$

Podemos então definir uma formula geral para calcularmos a porcentagem (pe) de uma população específica pertencente ao subgrupo (Xs).

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

4. Etapas Concluídas

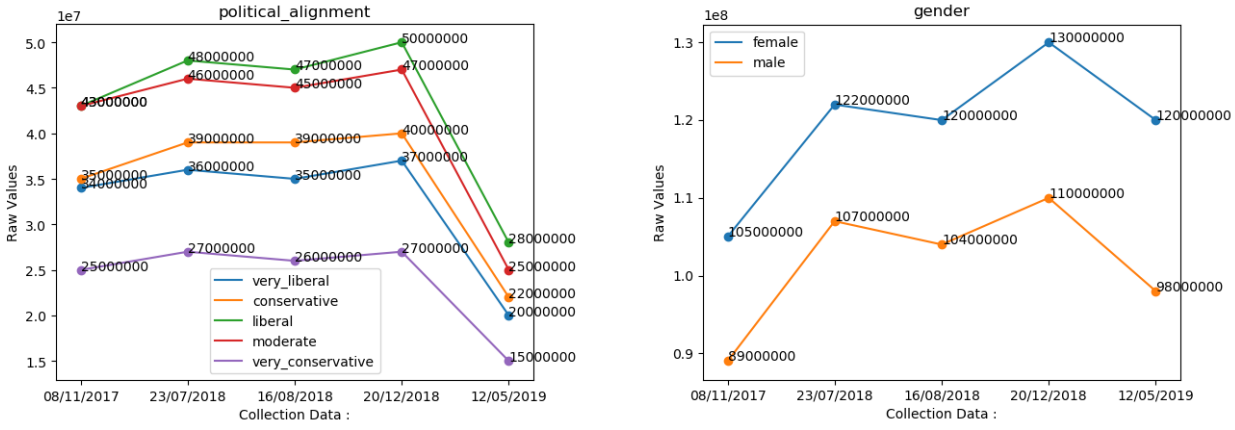
4.1. Revisão Bibliográfica

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.

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 onde os anúncios discriminavam usuários de grupos sensíveis de receberem seus anúncios [Speicher et al. 2018].

4.2. Coleta e Análise dos Dados

As coletas começaram a ser efetuadas em meados de maio do presente ano sendo feitas uma vez por semana. Nós aproveitamos coletas similares feitas durante os anos de 2017 e 2018 para utilizarmos de comparação com a nova base de dados adquirida, foi possível verificar uma variação notável em vários dos atributos, como exemplificado na figura 3.



(a) Inclinação política

(b) Gênero

Figura 3. Exemplo de comparação entre coletas.

Ao realizarmos o mesmo procedimento, porém aplicado as coletas efetuadas para o desenvolvimento deste trabalho, podemos notar uma variação bem menor do que quando comparamos dados de períodos distantes, como mostrado na figura 4.

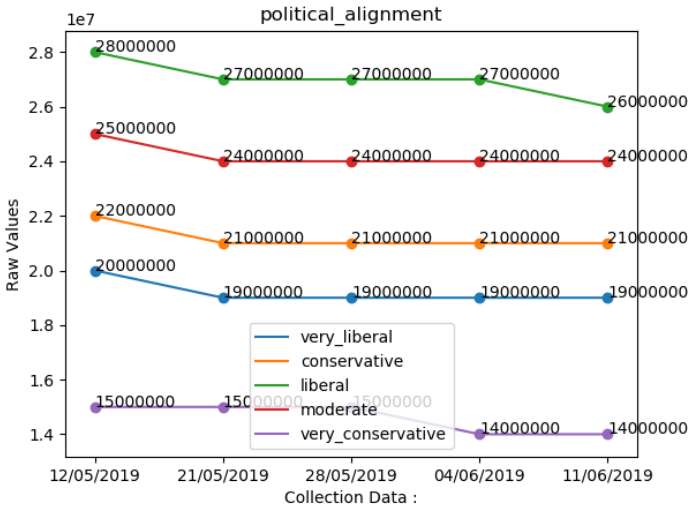


Figura 4. Variação da inclinação política em um período de 5 semanas.

5. Avaliação do andamento do trabalho e considerações finais

O planejamento do trabalho foi concluído com sucesso, assim como foram iniciadas as coletas de dados com uma pequena análise inicial. Caso necessário, alterações podem ser efetuadas, com intenção de buscar sempre resultados mais proveitosos da coleção de dados a ser analisada. Os próximos passos do presente trabalho incluem:

1. Definir interesses a serem direcionados na pesquisa para o Brasil.
2. Dar continuidade as coletas em ambos países.
3. Realizar a analise temporal e estatística de toda a base de dados.

Referências

[Araujo et al. 2017] Araujo, M., Mejova, Y., Weber, I., and Benevenuto, F. (2017). Using facebook ads audiences for global lifestyle disease surveillance: Promises and limitations. In *Proceedings of the 2017 ACM on Web Science Conference*, WebSci '17, pages 253–257, New York, NY, USA. ACM.

[Argamon et al. 2009] Argamon, S., Koppel, M., Pennebaker, J. W., and Schler, J. (2009). Automatically profiling the author of an anonymous text. *Commun. ACM*, 52(2):119–123.

[Dong et al. 2014] Dong, Y., Yang, Y., Tang, J., Yang, Y., and Chawla, N. V. (2014). Inferring user demographics and social strategies in mobile social networks. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 15–24. ACM.

[Jones et al. 2007] Jones, R., Kumar, R., Pang, B., Tomkins, A., Tomkins, A., and Tomkins, A. (2007). I know what you did last summer: query logs and user privacy. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 909–914. ACM.

[Nations 2017] Nations, U. (2017). *Principles and Recommendations for Population and Housing Censuses, Revision 3*.

[Rao et al. 2010] Rao, D., Yarowsky, D., Shreevats, A., and Gupta, M. (2010). Classifying latent user attributes in twitter. In *Proceedings of the 2nd international workshop on Search and mining user-generated contents*, pages 37–44. ACM.

[Ribeiro et al. 2018] Ribeiro, F., Henrique, L., Benevenuto, F., Chakraborty, A., Kulshrestha, J., Babaei, M., and Gummadi, K. P. (2018). 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*, ICWSM'18, Stanford, USA.

[Sap et al. 2014] Sap, M., Park, G., Eichstaedt, J., Kern, M., Stillwell, D., Kosinski, M., Ungar, L., and Schwartz, H. A. (2014). Developing age and gender predictive lexica over social media. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1146–1151.

[Speicher et al. 2018] Speicher, T., Ali, M., Venkatadri, G., Ribeiro, F. N., Arvanitakis, G., Benevenuto, F., Gummadi, K. P., Loiseau, P., and Mislove, A. (2018). On the Potential for Discrimination in Online Targeted Advertising. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT*'18)*.