UNIVERSIDADE FEDERAL RURAL DE PERNAMBUCO

UNIDADE ACADÊMICA DE EDUCAÇÃO A DISTÂNCIA E TECNOLOGIA

BACHARELADO EM SISTEMAS DE INFORMAÇÃO

Análise do comportamento através dos dados coletados na internet

Por

Priscilla Amarante de Lima

Recife,
Abril/2021

UNIVERSIDADE FEDERAL RURAL DE PERNAMBUCO
UNIDADE ACADÊMICA DE EDUCAÇÃO A DISTÂNCIA E TECNOLOGIA
CURSO DE BACHARELADO EM SISTEMAS DE INFORMAÇÃO

PRISCILLA AMARANTE DE LIMA

Análise do comportamento através dos dados coletados na internet

Trabalho de Conclusão de Curso apresentada ao Curso de Bacharelado em Sistemas de Informação da Unidade Acadêmica de Educação a Distância e Tecnologia da Universidade Federal Rural de Pernambuco como requisito parcial à obtenção do grau de Bacharel.

Orientadora: Prof[a]. Dr[a]. Juliana Basto Diniz

Recife,
Abril/2021

Dados Internacionais de Catalogação na Publicação
Universidade Federal Rural de Pernambuco
Sistema Integrado de Bibliotecas
Gerada automaticamente, mediante os dados fornecidos pelo(a) autor(a)

L732a Lima, Priscilla Amarante
Análise do comportamento através dos dados coletados na internet / Priscilla Amarante Lima. - 2021.
46 f.

Orientadora: Juliana Basto Diniz.
Inclui referências.

Trabalho de Conclusão de Curso (Graduação) - Universidade Federal Rural de Pernambuco, Bacharelado em Sistemas da Informação, Recife, 2021.

1. Big Tech. 2. Big Data. 3. Machine Learning. I. Diniz, Juliana Basto, orient. II. Título

CDD 004

UNIVERSIDADE FEDERAL RURAL DE PERNAMBUCO
UNIDADE ACADÊMICA DE EDUCAÇÃO A DISTÂNCIA E TECNOLOGIA
BACHARELADO EM SISTEMAS DE INFORMAÇÃO

PRISCILLA AMARANTE DE LIMA

Análise do comportamento através dos dados coletados na internet

Trabalho de Conclusão de Curso julgado adequado para obtenção do título de Bacharel em Sistemas de Informação, defendida e aprovada por unanimidade em xx/xx/xxxx pela banca examinadora.

Banca Examinadora:

Juliana Diniz

Prof. M.e Nome Sobrenome
Orientador
Universidade Federal Rural de Pernambuco

Prof. M.e Nome Sobrenome
Universidade Federal Rural de Pernambuco

Prof. M.e Nome Sobrenome
Universidade Federal Rural de Pernambuco

Profª M.ª Nome Sobrenome
Universidade Federal Rural de Pernambuco

Dedico este trabalho a minha família e amigos.

AGRADECIMENTOS

Aos meus pais, por todo apoio e carinho de sempre em todos os momentos da minha vida.

Aos meus amigos, por ser fonte de inspiração e pelos momentos de alegria.

A minha orientadora, Juliana, pelos ensinamentos e ajuda no desenvolvimento deste trabalho.

RESUMO

Este trabalho apresenta uma análise sobre o comportamento humano através dos dados coletados na internet. Serão apresentadas as Big Techs e o estudo de caso da Cambridge Analytica. Os registros digitais de comportamento podem ser acessados, através das curtidas no Facebook e serem usadas para prever de forma automática e precisa um intervalo de atributos pessoais altamente confidenciais, incluindo: orientação sexual, etnia, pontos de vista religiosos e políticos, traços de personalidade, inteligência, felicidade, uso de substâncias viciantes, separação dos pais, idade e sexo. A análise apresentada é baseada em um conjunto de dados de mais de 58.000 voluntários que forneceram curtidas no Facebook, perfis demográficos detalhados e os resultados de vários testes psicométricos. O modelo proposto usa redução de dimensionalidade para processar os dados de curtidas, que são então inseridos em regressão linear para prever perfis psicodemográficos individuais de curtidas. O modelo classifica corretamente entre homens homossexuais e heterossexuais em 88% dos casos, afro-americanos e Americanos caucasianos em 95% dos casos, e entre democratas e Republicanos em 85% dos casos. Para o traço de personalidade "Abertura", a precisão da previsão está próxima da precisão teste-reteste de um padrão teste de personalidade. São apresentados exemplos de associações entre atributos e curtidas e discutidas as implicações para a personalização online e privacidade.

Palavras-chave: Aprendizagem de máquina, Big Tech, Big Dados, Rede Social.

ABSTRACT

This work presents an analysis of human behavior through data collected on the internet. They will be confirmed as Big Techs and the Cambridge Analytica case study. We show that digital records of behavior easily obtained, through likes, through Facebook can be used to automatically and accurately predict a range of highly confidential personal attributes, including: sexual orientation, ethnicity, religious and political views, personality traits, intelligence, happiness, use of addictive substances, separation from parents, age and sex. The based analysis is based on a data set of more than 58,000 volunteers who provided Facebook likes, detailed demographic profiles and the results of various psychometric tests. The proposed model uses dimensionality reduction to pre-process the tanned data, which is then inserted in linear regression to predict individual psych demographic profiles of tanned. The model correctly discriminates between homosexual and heterosexual men in 88% of cases, African-Americans and Caucasian Americans in 95% of cases, and between Democrats and Republicans in 85% of cases. For the personality trait "Aperture", prediction accuracy is close to the test-retest accuracy of a personality test pattern. We give examples of associations between attributes and likes and discuss it as a conclusion for online personalization and privacy.

Keywords: Big Tech, Big Data, Social Networks, Machine Learning

LISTA DE FIGURAS

Figura 1 – Desenho do estudo 8
Figura 2 - Previsão da classificação para dicotomia - Atributos expressos pela AUC 20
Figura 3 - Coeficiente de correlação de Pearson 21
Figura 4 - Análise dos números de curtidas 23
Figura 5 – Mudança de comportamento 32
Figura 6 – Teoria Big Five da Personalidade 33
Figura 7 – Permissões de aplicativos do Facebook e os itens de perfil correspondentes 35

SUMÁRIO

1 INTRODUÇÃO 13

1.1 OBJETIVOS 14

1.1.1. Geral 14

1.1.2. Específicos 14

1.2 METODOLOGIA 15

1.3 ESTRUTURA DA MONOGRAFIA 16

2 ANÁLISE DE DADOS COLETADOS NA INTERNET 17

2.1 Estudo do algoritmo para análise do comportamento. 18

2.2 Predição de variáveis dicotômicas. 20

2.3 Predição de variáveis numéricas. 22

3 ESTUDO DE CASO - CAMBRIDGE ANALYTICA 27

3.1 – Etapa 1 – Modelo de personalidade 29

3.2 – Etapa 2 – Mineração de dados 34

3.3 – Etapa 3 – Propagandas direcionadas 34

4 CONSIDERAÇÕES FINAIS 37

REFERÊNCIAS BIBLIOGRÁFICAS 39

1 INTRODUÇÃO

Uma nova *commodity*[1] vem criando uma indústria lucrativa e de crescimento rápido. Um século atrás, o recurso em questão era o petróleo. Agora, quem domina o mercado são as *Big Techs* que negociam dados, o petróleo da era digital afirma Maurício Ruiz, presidente da Intel no Brasil (LOUREIRO, 2018). Dentre as *Big Techs* mais valiosas do mundo estão a Google, Amazon, Facebook e Microsoft.

Elas possuem algoritmos que podem prever o comportamento do usuário na rede, assim o Google pode saber o que as pessoas estão pesquisando, o Facebook conhece o que elas compartilham e a Amazon o que elas compram. O sucesso dessas gigantes beneficiou seus consumidores, pois poucas pessoas querem viver sem a ferramenta de pesquisa do Google, a entrega em um dia da Amazon ou o *feed* de notícias do Facebook.

A crescente proporção de atividades humanas, como interações sociais, entretenimento, compras e coleta de informações são agora mediadas por serviços e dispositivos digitais. Tal comportamento mediado digitalmente pode ser facilmente registrado e analisado, alimentando o surgimento de ciências sociais computacionais (LAZER, 2009) e novos serviços como motores de busca personalizados, sistemas de recomendação (KOREN, 2009) e marketing online direcionado (CHEN, 2009). No entanto, a ampla disponibilidade de extensos registros de comportamento individual, juntos com o desejo de aprender mais sobre clientes e cidadãos, apresenta sérios desafios relacionados à privacidade e propriedade de dados (BUTLER, 2007, SHMATIKOV, 2008).

Há uma distinção entre os dados que são realmente registrados e as informações que podem ser previstas estatisticamente a partir de tais registros. As pessoas podem optar por não revelar certas informações sobre suas vidas, como sua orientação sexual ou

[1] Segundo o Ministério do Desenvolvimento, Indústria e Comércio Exterior (MDIC), **Commodity** é um termo de língua inglesa (plural commodities), que significa mercadoria. É utilizado nas transações comerciais de produtos de origem primária nas bolsas de mercadorias. O termo é usado como referência aos produtos de base em estado bruto (matérias-primas) ou com pequeno grau de industrialização, de qualidade quase uniforme, produzidos em grandes quantidades e por diferentes produtores. Estes produtos "in natura", cultivados ou de extração mineral, podem ser estocados por determinado período sem perda significativa de qualidade. Possuem cotação e negociabilidade globais, utilizando bolsas de mercadorias.

idade, e ainda estas informações podem ser previstas em um sentido estatístico a partir de outros aspectos de suas vidas que eles revelam. Por exemplo, uma importante rede de varejo usava registros de compras de clientes para prever a gravidez de suas clientes do sexo feminino e enviar-lhes ofertas oportunas e bem direcionadas (DUHIGG, 2012).

Em alguns contextos, uma inundação inesperada de vouchers para vitaminas pré-natais e roupas de maternidade pode ser bem-vinda, mas também pode levar a um desfecho trágico, por exemplo, ao revelar (ou sugerir incorretamente) a gravidez de uma mulher solteira para sua família em uma cultura onde isso é inaceitável (INCE, 2009). Como este exemplo mostra prever as informações pessoais para melhorar produtos, serviços e segmentação também podem levar a invasões perigosas de privacidade.

Prever características e atributos individuais com base em várias pistas, como amostras de texto escrito (FAST, 2008), resposta a um teste psicométrico (COSTA, 1992), ou o aparecimento de espaços que as pessoas habitam (GOSILING, 2002), tem uma longa história. A migração humana para o ambiente digital torna possível basear tais previsões em registros digitais do comportamento humano.

Nesse sentido, o presente trabalho estuda como os dados dos usuários do Facebook podem ser coletados, através de curtidas (*Likes*), e utilizados para formular um algoritmo para compreender o comportamento humano usando os traços de personalidade dos usuários.

1.1 OBJETIVOS

1.1.1. Geral

- Verificar a estratégia utilizada pelas corporações para análise do comportamento humano através dos dados coletados na internet.

1.1.2. Específicos

- Verificar o algoritmo utilizado na análise do comportamento humano;
- Apresentar o estudo de caso sobre a Cambridge Analytica.

1.2 METODOLOGIA

Sob o aspecto metodológico o estudo consiste em uma pesquisa qualitativa com base em levantamento bibliográfico (livros e artigos científicos) sobre análise do comportamento através dos dados coletados na internet.

A metodologia utilizada no desenvolvimento deste trabalho se constitui das seguintes fases:

- Revisão bibliográfica dos conceitos e etapas da análise do comportamento através dos dados coletados na internet – buscou-se estudar a tecnologia utilizada na análise do comportamento e como foi utilizada a manipulação dos dados na internet através de pesquisa por artigos científicos em plataformas especializadas, por exemplo: Scielo e periódicos CAPES;
- A tecnologia utilizada na análise do comportamento - procurou-se pesquisar e estudar a ferramenta utilizada no caso da Cambridge Analítica para manipulação e mapeamento de perfis o modelo Ocean, a coleta de dados dos usuários através da Graph API e a plataforma de Big data e mineração de dados - Ripon;

Desse modo, o referido trabalho se coloca como um estudo exploratório que abre caminhos para outros estudos, que despertem o interesse em discutir sobre o cenário da análise do comportamento através dos dados coletados na internet.

1.3 ESTRUTURA DA MONOGRAFIA

A monografia será composta por:

- Introdução: neste capítulo, foi apresentada a motivação para o trabalho, tomando como base a relação das *Big Techs* com a análise do comportamento humano através dos dados coletados na internet. Também foi apresentada a metodologia utilizada na monografia;
- Análise de dados coletados na internet: neste capítulo, será analisado o estudo do algoritmo para análise do comportamento desenvolvido por Kosinski;
- Estudo de caso da Cambridge Analytica: neste capítulo, será apresentado como é possível coletar os dados na internet utilizando o caso de repercussão internacional da Cambridge Analytica;
- Considerações finais: serão apresentados os efeitos provocados pela tecnologia através da manipulação dos dados coletados na internet causam na sociedade;

2 ANÁLISE DE DADOS COLETADOS NA INTERNET

As redes sociais se tornaram meios amplamente usados e populares para a disseminação de informações, bem como facilitadores de interações sociais. As contribuições e atividades dos usuários fornecem uma visão valiosa sobre o comportamento, experiências, opiniões e interesses individuais. Considerando que a personalidade, que identifica de maneira única cada um de nós, afeta muitos aspectos do comportamento humano, processos mentais e reações afetivas, há uma enorme oportunidade para adicionar novas qualidades baseadas na personalidade às interfaces de usuário. Sistemas personalizados usados em domínios como e-learning, filtragem de informações, colaboração e e-commerce podem se beneficiar muito de uma interface de usuário que adapta a interação (por exemplo, estratégias motivacionais, estilos de apresentação, modalidades de interação e recomendações) de acordo com a personalidade do usuário. Ter capturado as interações anteriores do usuário é apenas um ponto de partida para explicar o comportamento do usuário do ponto de vista da personalidade.

Em um estudo desenvolvido por Kosinski (2012), foi demonstrado que idade, sexo, ocupação, nível de escolaridade, e até mesmo a personalidade pode ser prevista no site através do histórico de navegação do usuário. No estudo de Schutz (2006), foi mostrado que a personalidade pode ser prevista com base no conteúdo de sites pessoais, coleções de música (GOSILING, 2002), propriedades de perfis do Facebook ou Twitter como o número de amigos ou a densidade de redes de amizade, ou linguagem usada por seus usuários (GOLBECK, 2011). Além disso, a localização dentro de uma rede de amizade no Facebook mostrou ser preditivo da orientação sexual (MISTREE, 2009).

O estudo de Kosinski (2012) se baseou nos *Likes* do Facebook, através do mecanismo usado por seus usuários para expressar sua associação positiva com conteúdo online através das curtidas, como fotos, atualizações de status de amigos, páginas de produtos no Facebook, esportes, músicas, livros, restaurantes ou sites populares. As curtidas representam uma classe muito genérica de registros digitais, semelhante a consultas de pesquisa na web, históricos de navegação na web e das compras com cartão de crédito.

Por exemplo, observar as curtidas dos usuários relacionadas à música, fornece informações semelhantes à observação de registros de músicas ouvidas online e artistas pesquisados usando um mecanismo de pesquisa da Web ou assinaturas de canais do Twitter relacionados. Contudo, esses outros registros digitais ainda estão disponíveis para várias partes (por exemplo, governos, desenvolvedores de navegadores da Web, mecanismos de pesquisa, ou aplicativos do Facebook) e, portanto, previsões semelhantes são improváveis que sejam limitadas ao ambiente do Facebook.

2.1 Estudo do algoritmo para análise do comportamento.

As técnicas de mineração de dados desempenham um papel fundamental na extração de padrões de correlação entre personalidade e variedade de dados do usuário capturados de fontes múltiplas. No estudo desenvolvido por Kosinski (2013), foram adotadas duas abordagens para estudar traços de personalidade dos usuários nas redes sociais. A primeira abordagem usa uma variedade de algoritmos de aprendizado de máquina para construir modelos baseados em redes sociais. O segundo estende as características relacionadas à personalidade com dicas linguísticas (Mairesse et al. 2007; Oberlander e Nowson 2006). Várias técnicas de classificação e regressão foram usadas para construir modelos de personalidade preditivos ao longo das cinco dimensões de personalidade usando as características linguísticas de um conjunto de dados composto por alguns milhares de ensaios solicitados de alunos de introdução à psicologia (Mairesse et al. 2007).

Foram usados no estudo duas técnicas de regressão m5sup/ Processos de Rules and Gaussian aplicados para construir modelos preditivos de personalidade, além dos atributos linguísticos extraídos com LIWC (http://www.liwc.net) usado como uma ferramenta para análise linguística. O parâmetro H4Lvd uma ferramenta para análise de conteúdo de dados, que incluem 182 tags de palavras. O algoritmo para predição utilizado o Support Vector Machines (SVM) e seus versões mais eficientes e otimizadas, Simple Minimal Otimização (SMO) e o Boost (MultiBoostAB e AdaBoostM1) (KOSINSKI, 2013).

Foram selecionadas características e atributos que revelam quão precisos e potencialmente intrusivos essa análise preditiva, pode ser incluindo "orientação sexual", "Origem étnica", "visões políticas", "religião", "personalidade", "inteligência",

"satisfação com a vida", uso de substâncias ("álcool", "drogas", "cigarros"), "se os pais do indivíduo permaneceram juntos até que o indivíduo fizesse 21 anos" e atributos demográficos básicos, como "idade", "gênero"," status de relacionamento" e "tamanho e densidade da rede de amizades".

Os cinco fatores sobre os escores de personalidade do modelo (COSTA, 1992) foram estabelecidos usando o questionário International Personality Item Pool (IPIP) com 20 itens (GOLDBERG, 2006). A inteligência foi medida usando Matrizes Progressivas Padrão de Raven (SPM[2]) (RAVEN, 2000) e a satisfação com a vida foi medida usando a escala Satisfaction With Life - SWL (DIENER, 1985).

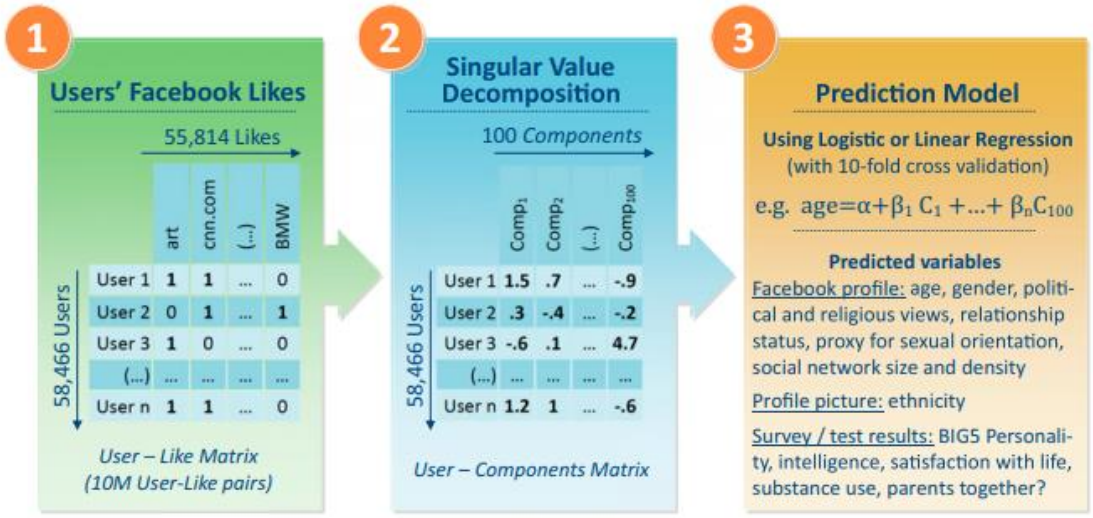
A média da idade foi aproximadamente de 25,6, gênero (62% feminino), status de relacionamento ("solteiro" / "em relacionamento"; 49% solteiro), opiniões políticas ("Liberal" / "Conservador"; 65% liberal), religião ("muçulmano" / "cristão"; 90% Cristão) e as informações da rede social foram obtidas dos perfis dos usuários do Facebook.

Consumo de álcool pelos usuários (50% bebem), drogas (21% usam drogas) e cigarros (30% fumam) e se os pais do usuário ficaram juntos até o usuário completar 21 anos (56% permaneceram juntos) foram registrados por meio de pesquisas online. A inspeção visual das fotos do perfil foi usada para atribuir origem étnica a uma pessoa selecionada aleatoriamente usando uma sub amostra de usuários (73% caucasianos; 14% africanos Americano; 13% outros). A orientação sexual foi atribuída usando o Campo "Interessado em" do perfil do Facebook; usuários interessados apenas em outros do mesmo sexo foram rotulados como homossexuais (4,3% homens; 2,4% mulheres), enquanto os interessados em usuários do gênero oposto foram rotulados como heterossexuais.

---

[2] O SPM-Standard Progressive Matrices é um teste de inteligência não verbal de múltipla escolha baseado na teoria geral da habilidade de Spearman. O SPM é um teste de inteligência padrão comprovado usado em pesquisas e configurações clínicas, bem como em contextos de alto risco, como na seleção de militares e processos judiciais (RAVEN, 2000).

Figura 1 – Desenho do estudo



Fonte: Kosinski (2013)

O estudo ilustrado na figura 1 foi baseado em uma amostra de 58.466 voluntários dos Estados Unidos, obtida por meio do aplicativo myPersonality no Facebook (www.mypersonality.org /wiki), que incluía as informações do perfil do Facebook, uma lista de seus likes (n = 170 likes por pessoa em média), pontuações de testes psicométricos e informações de pesquisas. Os usuários e suas curtidas foram representados como uma matriz esparsa do tipo usuário, cujas entradas foram definidas como 1 se existisse uma associação entre um usuário e a Curtir e 0 caso contrário.

Variáveis numéricas como idade ou inteligência foram previstas usando um modelo de regressão linear, enquanto variáveis dicotômicas, como sexo ou orientação sexual, foram previstas usando regressão logística (KOSINSKI, 2013).

2.2 Predição de variáveis dicotômicas.

A predição de variáveis dicotômicas é equivalente à probabilidade de classificar corretamente dois usuários selecionados aleatoriamente, um de cada gênero (por exemplo, masculino e feminino). A figura 2 mostra a previsão das variáveis dicotômicas obtidas através do estudo de Kosinski (2013) expressas em termos da área sob a curva característica de operação do receptor Area Under Curve (AUC). A maior precisão foi

alcançada para origem étnica e gênero. Africano Americano e caucasiano americano foram corretamente classificados em 95% dos casos, homens e mulheres foram classificados corretamente em 93% dos casos, sugerindo que os padrões de comportamento online como expressos por curtidas diferem significativamente entre esses grupos permitindo uma classificação quase perfeita.

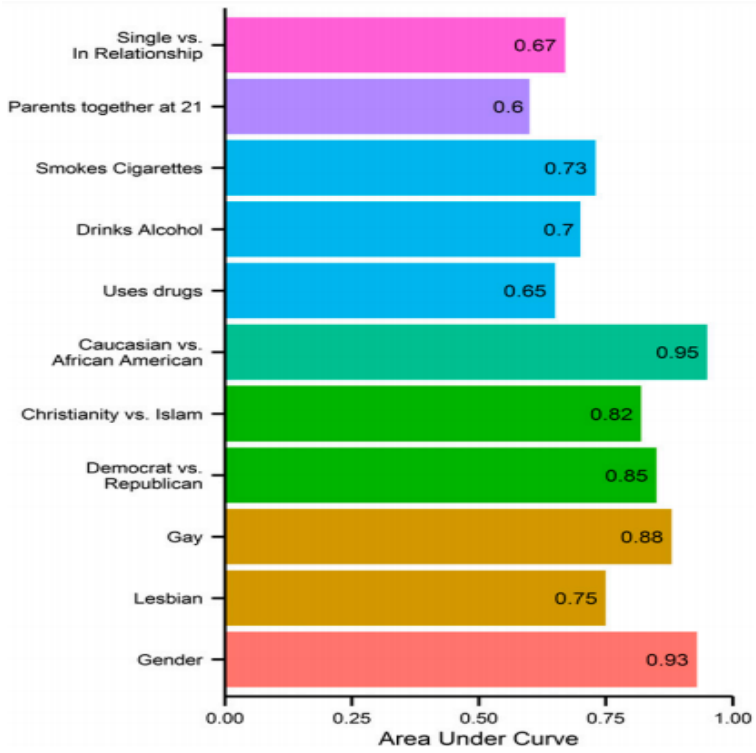
Cristãos e muçulmanos foram classificados corretamente em 82% dos casos, e resultados semelhantes foram alcançados para democratas e republicanos (85%). A orientação sexual era mais fácil de distinguir entre os homens (88%) do que as mulheres (75%), o que pode sugerir um comportamento mais amplo a divisão (conforme observado a partir do comportamento online) entre homens heterossexuais e homossexuais.

Uma boa precisão de previsão foi alcançada para o status de relacionamento e uso de substâncias (entre 65% e 73%). A precisão relativamente mais baixa para o status de relacionamento pode ser explicada por sua variabilidade em comparação com outras variáveis dicotômicas (por exemplo, sexo ou orientação sexual).

A precisão do modelo foi mais baixa (60%) ao inferir se os pais dos usuários ficavam juntos ou separados antes dos usuários completarem 21 anos. Embora seja sabido que o divórcio dos pais tem efeitos de longo prazo no bem-estar dos jovens adultos (MUSICK, 2010), é notável que isso possa ser detectável por meio de seus likes no Facebook. Indivíduos com pais separados têm maior probabilidade de gostar das declarações preocupadas com relacionamentos, como "Se eu estiver com você então eu estou com você eu não quero mais ninguém" (KOSINSKI, 2013).

Figura 2 – Previsão da classificação para dicotomia

Atributos expressos pela AUC



Fonte: Kosinski (2013)

2.3 Predição de variáveis numéricas.

A predição de variáveis numéricas trás a previsão de regressão para atributos numéricos e características expressas pelo coeficiente de correlação de Pearson entre os valores de atributos previstos e reais, todas as correlações são significativas no nível P <0,001. O coeficiente de correlação de Pearson é um teste que mede a relação estatística entre duas variáveis contínuas podendo ter um intervalo de valores de +1 a -1.

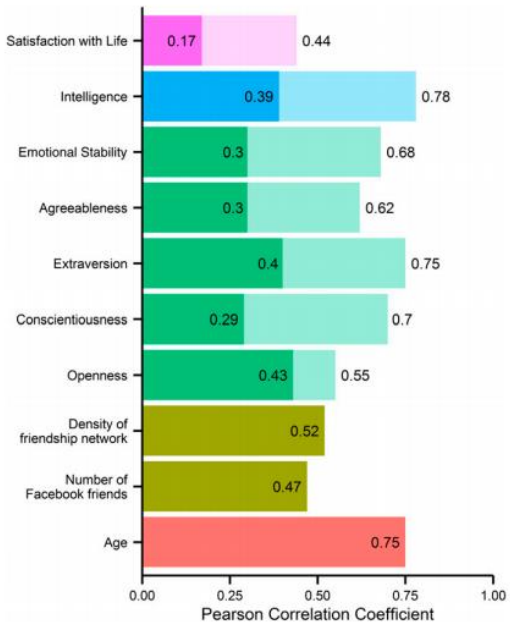
A Figura 3 apresenta a precisão de variáveis numéricas expressas pelo produto Pearson - coeficiente de correlação entre o momento real e os valores previstos. No estudo desenvolvido por Kosinski (2013) a maior correlação foi obtida para a idade (r = 0,75), seguido pela densidade (r = 0,52) e tamanho (r = 0,47) do Facebook rede de amizade. Seguindo de perto estavam os traços de personalidade de “Abertura” (r = 0,43), “Extroversão” (r = 0,40) e “Inteligência” (r = 0,39). Os traços de personalidade restantes e a satisfação com a vida foram previstos com uma precisão um pouco menor (r = 0,17 a 0,30). Traços psicológicos são exemplos de traços latentes (ou seja, traços que não pode

ser medido diretamente). Como consequência, seus valores só podem ser medidos aproximadamente, por exemplo, avaliando as respostas dos questionários.

As barras transparentes apresentadas na figura 3 indicam a precisão dos questionários usados conforme expresso por suas confiabilidades teste-reteste (correlação produto-momento de Pearson entre as pontuações do questionário obtidas pelo mesmo respondente em dois momentos). A correlação entre o previsto e o escore de abertura real (r = 0,43) foi muito próximo ao teste-reteste confiabilidade para abertura (r = 0,50). Isso indica que para o traço de abertura, a observação das curtidas do usuário é quase tão informativa quanto usar a própria pontuação do teste de personalidade. Para as características restantes, as precisões de previsão correspondem a cerca de metade confiabilidade teste-reteste do questionário.

A precisão de predição relativamente mais baixa para satisfação com a vida (r = 0,17) pode ser atribuída à dificuldade de separar ao longo do tempo a felicidade (SCHIMMACK, 2002) de mudanças de humor, que variam com o tempo. Por isso, embora a pontuação de satisfação com a vida inclui a variabilidade atribuível ao humor, gostos dos usuários são acumulados ao longo de um período mais longo e, portanto, podem ser adequados apenas para prever a felicidade ao longo do tempo.

Figura 3 – Coeficiente de correlação de Pearson



Fonte: Kosinski (2013)

Os resultados apresentados por Kosinski (2013) contam com indivíduos para os quais entre um e 700 curtidas estavam disponíveis. O número médio de curtidas foi 68 por indivíduo. Portanto, qual é a precisão esperada dada a um indivíduo aleatório e como a precisão da previsão muda com o número de curtidas observadas? Usando uma sub amostra (n = 500) de usuários para os quais pelo menos 300 curtidas foram executados modelos preditivos com base em subconjuntos de n = 1, 2, ..., 300 curtidas. Os resultados apresentados na Figura 4 mostram que mesmo sabendo um único like aleatório para um determinado usuário pode resultar em precisão de predição não desprezível. Quanto maior for as Curtidas aumentará a precisão, mas com retornos decrescentes de cada informação adicional.

Traços e atributos individuais podem ser previsto com um alto grau de precisão com base em registros de gostos dos usuários. A Tabela 1 apresenta uma amostra de likes altamente preditivos relacionados a cada um dos atributos. Por exemplo, os melhores preditores de alta inteligência incluem os atributos "Thunderstorms", "The Colbert Report, "" Science "e "Curly Fries", enquanto a baixa inteligência foi indicada por “Sephora”, “I Love Being A Mom”, “Harley Davidson ”e“ Lady Antebellum ”.

Os preditores do sexo masculino com tendência a homossexualidade incluíram os atributos “No H8 Campaign,” “Mac Cosmetics,” e "Wicked The Musical", enquanto fortes preditores do sexo masculino com tendência a heterossexualidade incluiu os atributos "Wu-Tang Clan", "Shaq" e "Being”. Embora alguns dos gostos se relacionassem claramente com seu atributo previsto, como no caso de “No H8 Campaign” e homossexualidade, outros pares são mais evasivos; não há conexão óbvia entre Curly Fries e alta inteligência.

Tabela 1 – Traços e atributos (amostras de likes preditivos)

Trait		Selected most predictive Likes		
IQ	*High*	The Godfather	Jason Aldean	*Low*
		Mozart	Tyler Perry	
		Thunderstorms	Sephora	
		The Colbert Report	Chiq	
		Morgan Freemans Voice	Bret Michaels	
		The Daily Show	Clark Griswold	
		Lord Of The Rings	Bebe	
		To Kill A Mockingbird	I Love Being A Mom	
		Science	Harley Davidson	
		Curly Fries	Lady Antebellum	

Sexual Orientation				
	Homosexual Males	No H8 Campaign Kathy Griffin Kurt Hummel Glee Human Rights Campaign Mac Cosmetics Adam Lambert Ellen DeGeneres Juicy Couture Sue Sylvester Glee Wicked The Musical	X Games Nike Basketball Bungie WWE Sportsnation Wu-Tang Clan Foot Locker Shaq Bruce Lee Being Confused After Waking Up From Naps	*Heterosexual Males*
	Homosexual Females	Girls Who Like Boys Who Like Boys Rupauls Drag Race No H8 Campaign Gay Marriage Human Rights Campaign The L Word Sometimes I Just Lay In Bed And Think About Life Not Being Pregnant Gay Marriage Tegan And Sara	Lipton Brisk Yahoo Adidas Originals Foot Locker WWE Inbox 1 Makes Me Nervous Thinking Of Something And Laughing Alone I Just Realized Immature Spells I'm Mature Did You Get A Haircut No It Grew Shorter Nike Women	*Heterosexual Females*

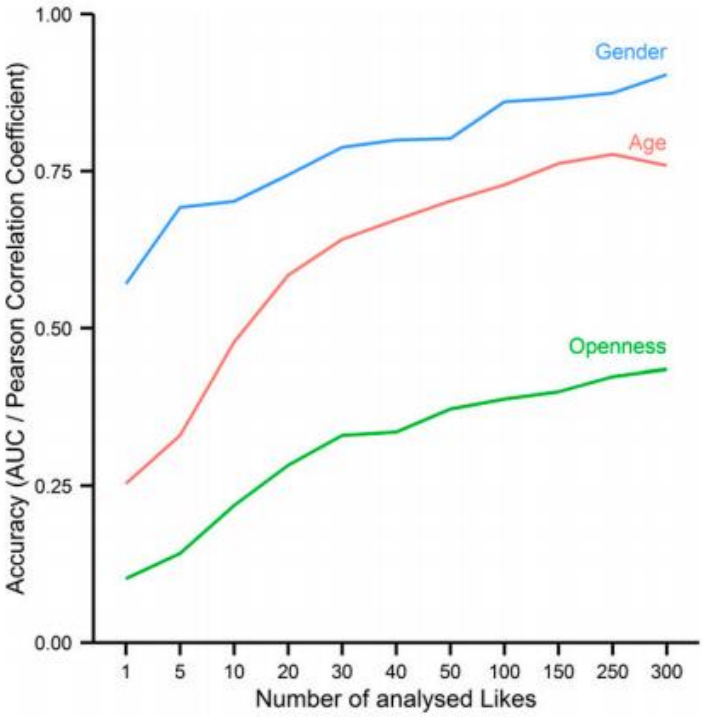
Fonte: Kosinski (2013)

Além disso, observa-se que poucos usuários foram associados a curtidas revelando explicitamente seus atributos. Por exemplo, menos de 5% de usuários rotulados como gays eram conectados a grupos explicitamente gays, como os atributos No H8 Campaign, "Being Gay," "Gay Marriage," "I love Being Gay"," Nós não escolhemos ser gays, fomos escolhidos "(KOSINSKI, 2013).

O gráfico ilustrado na figura 4 apresenta previsões selecionadas em função do número de Curtidas disponíveis. A precisão é expressa na relação entre a Area Under Curve - AUC (gênero) e coeficiente de correlação de Pearson (idade e abertura). Cerca de 50% dos usuários nesta amostra tiveram pelo menos 100 curtidas e cerca de 20% tiveram pelo menos 250 curtidas. Observe que para gênero (variável dicotômica) a linha de base corresponde a um AUC = 0,50.

Os resultados apresentados na figura 4 mostram que mesmo sabendo um único like aleatório para um determinado usuário pode resultar em precisão de predição não desprezível. Quanto mais curtidas tiver aumenta a precisão, mas com retornos decrescentes de cada informação adicional.

Figura 4 – Análise dos números de curtidas



Fonte: Kosinski (2013)

Os traços e atributos individuais podem ser previsto com um alto grau de precisão com base em registros a partir de likes dos usuários. Embora alguma das curtidas se relacionasse claramente com seu atributo previsto, como no caso de No H8 Campaign e homossexualidade, outros pares são mais evasivos. Além disso, poucos usuários foram associados a curtidas explicitamente revelando seus atributos. Por exemplo, menos de 5% de usuários rotulados como gays eram conectados a grupos explicitamente gays, como no No H8 Campaign, “Being Gay,” “Gay Marriage,” “I love Being visto anteriormente na tabela 1.

3 ESTUDO DE CASO - CAMBRIDGE ANALYTICA

Em 2013, pesquisadores da Universidade de Cambridge analisaram resultados de voluntários com relação a um teste de personalidade no Facebook para avaliar seu perfil psicológico com base no OCEAN, que traduzido para o português refletem respectivamente: abertura, conscienciosidade, extroversão, afabilidade e neuroticismo (CABRAL, 2018).

Esta pesquisa utilizou os resultados dos 350.000 participantes dos EUA e estabeleceu um relacionamento claro entre a atividade do Facebook (e outros indicadores online) e este perfil de personalidade de cinco fatores. De acordo com Hanna (2018) este resultado demonstrou que o perfil OCEAN para qualquer indivíduo pode ser deduzido, razoavelmente preciso, olhando as métricas e sem a necessidade de um instrumento psicográfico formal.

Não há indícios, de que esta pesquisa expôs usuários do Facebook ou seus amigos para qualquer abuso de privacidade específico. Há indícios que a Universidade recusou compartilhar dados com o que se tornaria Cambridge Analytica.

Entretanto, houve um segundo projeto de pesquisa iniciado pela Global Science Research (GSR) - em cooperação com Cambridge Analytica— para identificar os parâmetros necessários para desenvolver os perfis OCEAN usando um questionário de personalidade na Amazon - Plataforma Mechanical Turk e Qualtrics, uma plataforma de pesquisa. O quiz aplicado aos usuários concede acesso obrigatório ao GSR a seu perfil no Facebook, que concedeu acesso aos dados dos amigos dos usuários por meio de a API aberta do Facebook até maio de 2015. O objetivo era estabelecer uma metodologia para perfis psicográficos de indivíduos com base nas redes sociais e outros indicadores (HANNA, 2018).

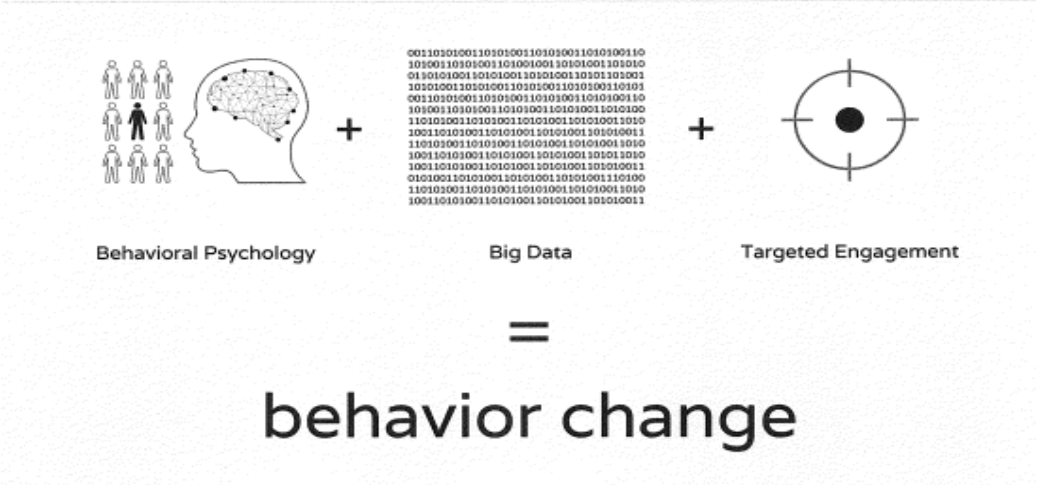
De acordo com Galloway (2020), as *Big Techs* (Google, Facebook, Amazon, Apple) já utilizam de tecnologia para rastrear e analisar cada movimento de um indivíduo on-line bem como quando eles estão lidando com seus negócios diários. As câmeras estão em quase toda parte e os algoritmos de reconhecimento facial são capazes de reconhecer qualquer pessoa.

A política de “privacidade” da empresa afirma que grupos externos de empresas patrocinadoras não tem acesso a informações ou perfis pessoais. No entanto, diz que as informações podem ser compartilhadas com outras empresas, advogados, agentes ou agências governamentais a fim de cumprir a lei como aconteceu no caso da Cambridge Analytica (HARRIS, 2018).

A Cambridge Analytica fazia uma análise da psicologia comportamental, a partir dos dados fornecidos pelos indivíduos no big data em suas redes sociais (Google, Snapchat, Twitter, Facebook e YouTube) e traçava um alvo de engajamento, ilustrado na figura 5, usando os dados de cada usuário do Facebook para prever sua personalidade.

Eles objetivaram quantificar a personalidade, pontuando usuários individuais em cinco traços de personalidade principais: abertura, consciência, extroversão, afabilidade, neuroticismo, que se refere ao modelo de personalidade Big five ou OCEAN a partir disso conseguia traçar o perfil da personalidade e armazenar as informações no big data. Para atingir o alvo, era preciso utilizar os recursos da inteligência artificial e propagandas direcionadas nas redes sociais a partir das curtidas e compartilhamento com outros usuários. Dessa forma, era possível atingir o objetivo da mudança de comportamento.

Figura 5 – Mudança de comportamento



Fonte: Cambridge Analytica (2013)

Mesmo para aqueles que não possuiam uma conta no Facebook puderam ter seus dados obtidos através de outros sites que possuem o logotipo do Facebook, o que permitiu o rastreamento de não membros. De acordo com Hanna (2018), existem muitas semelhantes fontes de rastreamento online - por exemplo, web beacons - a maioria dos

quais estão ligados a “cookies” que podem ser usados em sites, e o acesso pode ser vendido para compradores interessados.

De acordo com o artigo de Michal Kosinski e colegas (KOSINSKI, 2017), foi confirmado que o impacto significativo foi obtido com uma base de amostra de 3,5 milhões usuários. Com uma ampla base de informações pessoais prontamente disponíveis, a Cambridge Analytica conseguiu transformar dados coletados em simples questionários de personalidade, em ferramentas fundamentais em campanhas políticas, conseguindo concretizar resultados que eram tidos como altamente improváveis (FLORES, 2017).

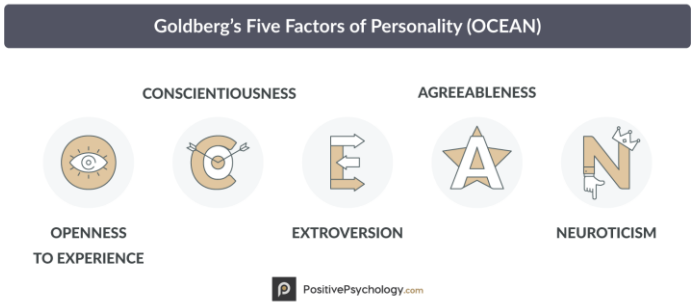
O processo de análise do comportamento dos usuários pode ser dividido em três etapas que serão apresentadas nas subseções a seguir.

3.1 – Etapa 1 – Modelo de personalidade

O trabalho inicia com a definição de qual modelo psicográfico que pode ser utilizado para modelar as personalidades e perfis das pessoas. De acordo com Digman (1990) o modelo OCEAN, também conhecido como a teoria Big Five da personalidade, é usado para classificar os indivíduos de acordo com o tipo de personalidade, e oferece aos pesquisadores várias ferramentas e ideias úteis para ajudar a explicar as tendências de líderes e seguidores de maneira consistente ao longo do tempo. Refere-se aos 5 fatores de personalidade pelo método lexical, ou seja, baseado em uma análise linguística. Conforme afirma Eysenck (1993), a maioria dos termos semelhantes às características que as pessoas usam para descrever outras pessoas pode ser categorizada de maneira confiável em cinco amplas dimensões de personalidade.

As cinco principais dimensões apresentadas na figura 6 incluem: abertura à experiência, consciência, extroversão, agradabilidade e neuroticismo. A partir das cinco categorias de personalidade, será possível entender o enquadramento dos perfis utilizado pela inteligência artificial para aprender sobre o comportamento humano (COURTNEY, 2021).

Figura 6 – Teoria Big Five da Personalidade



Fonte: https://positivepsychology.com/big-five-personality-theory/

- Openness to experience (abertura à experiência) preocupa-se com o pensamento inovador, a curiosidade, a assimilação de novas informações e a abertura a novas experiências.
- Conscientiousness (conscienciosidade) gira em torno da ideia de organização e perseverança. Líderes mais conscientes tendem a ser organizados, sinceros, levam a sério os compromissos e raramente se metem em problemas. Os que têm menos consciência tendem a serem mais espontâneos, criativos, inclina-se a regras e menos preocupados em cumprir os compromissos;
- Extroversion (extroversão) envolve comportamentos com maior probabilidade de serem exibidos em ambientes de grupo e geralmente preocupados em progredir na vida. Quando alguém está tentando influenciar ou controlar os outros, esses padrões comportamentais geralmente se tornam óbvios. Os líderes com alto nível de extroversão são autoconfiantes, sinceros, opinativos e competitivos;
- Agreeableness (agradabilidade) é uma dimensão da personalidade que se preocupa com o modo como a pessoa se dá bem, em vez de se antecipar aos outros. Indivíduos com muita simpatia se deparam com um ambiente charmoso, diplomático, acolhedor, acessível, enfático e otimista. Os que têm menos afabilidade são mais propensos a parecer socialmente ignorantes, frios, mal-humorados, insensíveis e um tanto pessimistas;
- Neuroticism (neuroticismo) está preocupado com a forma como as pessoas reagem ao estresse, mudança, crítica pessoal ou fracasso. Os líderes com menos neuroticismo

> tendem a serem calmos otimistas escondem suas emoções e não cometem erros ou falhas pessoalmente. Por outro lado, os que têm mais neuroticismo são apaixonados, mal-humorados, ansiosos e perdem a calma quando estressados ou criticados.

O método Big Five não é novo, muito menos o uso de modelos psicológicos pela ciência da persuasão. A grande inovação da CA está na metodologia. Eles pegaram um modelo computacional criado por Michal Kosinski para o Centro de Psicometria da Universidade de Cambridge e aplicaram ao marketing político. Em 2012, Kosinski gerou seu modelo: com 68 "curtidas" de um usuário do Facebook, ele poderia prever, com uma baixa margem de erro, sua cor de pele (95% de acerto), sua orientação sexual (88% de acerto) e sua afiliação ao Partido Democrata ou Republicano (85% de acerto). Inteligência, religião, consumo de álcool e tabaco também poderiam ser previstos. Analisando apenas 10 "likes" o modelo de Kosinski foi capaz de avaliar uma pessoa melhor do que um colega de trabalho, com 70 "likes", poderia fazer melhor do que um amigo e com 300, melhor do que seu parceiro (DUSSEL, 2018).

Além da psicologia, a CA usa big data, que é a agregação de dados. Cada vez que fazemos algo, deixamos "impressões digitais" que são registradas, coletadas e analisadas. Diferentes tipos de informação podem ser coletados sobre uma pessoa para entender seus pontos de vista: dados demográficos (idade, sexo, religião, etc.), dados atitudinais (que carros dirigem, quais revistas eles leem, que mídia eles consomem, o que hobbies que praticam a aparência dos filmes) e dados comportamentais (quantos contatos ou fotos você tem no Facebook, quantas ligações você faz, a que horas está acordado, etc). Por meio de modelos computacionais, esses dados são cruzados e um perfil individual é criado de acordo com um modelo de personalidade que permite a comunicação individualizada, sabendo antecipadamente que mensagem cada um dos destinatários quer ouvir (DUSSEL, 2018).

Em seguida, houve a escolha da forma de coleta de dados dos usuários através da Graph API, uma interface criada pelo próprio Facebook para que desenvolvedores consigam construir aplicações que possam utilizar as informações dos usuários (e de suas conexões) (SCHONFELD, 2010).

A introdução da API Graph foi anunciada pelo Facebook como uma forma revolucionária de compreender e acessar a vida social das pessoas via compartilhamento nas redes sociais. O que tornou a primeira versão v1.0 da API Graph do Facebook altamente problemática foram suas permissões estendidas. Os aplicativos podem solicitar uma grande variedade de informações de amigos dos usuários sem muito atrito ou

comunicar o (s) motivo (s) para fornecer consentimento. Uma vez autorizado, o aplicativo v1.0 poderia permanecer em segundo plano coletando e processando os dados das pessoas - e de toda a rede de amigos - por anos. Além disso, os aplicativos v1.0 também podem solicitar mensagens privadas dos usuários (ou seja, através da caixa de entrada do Facebook) por meio da solicitação de API " read_mailbox " (ALBRIGHT, 2018) .

As informações disponíveis dos amigos de seus usuários (apresentadas na figura 7) estão listadas na coluna itens do perfil "profile items" na linha propriedades estendida do perfil "Extended profile properties": sobre mim, ações, atividades, aniversário, check-ins, educação, eventos, jogos, grupos, cidade natal, interesses, curtidas, localização, notas, status online, tags, fotos, perguntas, relacionamentos, religião/política, status, assinaturas, site, histórico de trabalho.

Figura 7 - Permissões de aplicativos do Facebook e os itens de perfil correspondentes

Permission Group	Permissions	Profile Items
Public profile (default)	public profile	id, name, first name, last name, link, gender, locale, timezone, updated time, verified
App friends	user friends	bio, birthday, education, first name, last name, gender, interested in, languages, location, political, relationship status, religion, quotes, website, work,

Extended Profile Properties	friends about me, friends actions, friends activities, friends birthday friends checkins, friends education history, friends events, friends games activity, friends groups, friends hometown, friends interests, friends likes, friends location, friends notes, friends online presence, friends photo video tags, friends photos, friends questions, friends relationship details, friends relationships, friends religion politics, friends status, friends subscriptions, friends website, friends work history	about me, actions, activities, birthday checkins, history, events, games activity, groups, hometown, interests, likes, location, notes, online presence, photo video tags, photos, questions, relationship details, relationships, religion politics, status, subscriptions, website, work history
Extended Permissions	read mailbox	inbox

Fonte: Symeonidis, I. etall (2015)

Para dar uma ideia da enormidade do gráfico social subjacente ao Facebook, em 2012 foi anunciado que o Facebook tinha 901 milhões de usuários, e o gráfico social consiste em muitos tipos além de apenas usuários (WEAVER AND TARJAN, 2012). Partindo das questões de privacidade que surgem na instalação de terceiros aplicativos (Apps) por meio dos amigos do usuário no Facebook, que ilustra como a divulgação de dados do usuário ocorre por meio dos aplicativos instalados de seus amigos. Vários aplicativos coletam permissões para informações confidenciais consideradas, como e-mail (68,99%), informações de amigos (10,23%) e, mais invasivos, privilégios de caixa de correio privada (1%) (SYMEONIDIS, 2015).

Considerando que as permissões aos dados de amigos parecem limitados, a total falta de transparência e opção (falta de consentimento válido) para o usuário é bastante preocupante. O impacto da privacidade que surge com a aquisição de dados pessoais dos usuários por meio de aplicativos instalados por seus amigos no Facebook e o

agrupamento de dados pessoais dos usuários por meio de Apps, expondo esses dados fora do app e sem o conhecimento prévio dos usuários (DUSSEL, 2018).

Ser capaz de aumentar a conscientização sobre a coleta de itens de perfil está em linha com o legal princípio da proteção de dados por padrão, pois pode potencialmente apoiar decisões e promover o controle do usuário sobre a divulgação de dados pessoais.

3.2 – Etapa 2 – Mineração de dados

A plataforma de big data e mineração de dados utilizados pela Cambridge Analytica foi o Ripon, software desenvolvido pela empresa canadense AggregateIQ (AIQ). De acordo com o Washington Post (2018), o Ripon foi o software que utilizou os algoritmos dos dados do Facebook, através de mecanismos de inteligência artificial permitia a uma campanha gerenciar seu banco de dados de eleitores, atingir eleitores específicos, conduzir campanhas de angariação de fundos e realizar pesquisas. A ferramenta utilizava o modelo OCEAN com algumas modificações para fazer o perfil dos usuários da rede social.

De acordo com Lombardo (2019), combinando mineração de dados, inteligência da mídia social, traços psicológicos, aprendizagem de máquina e algoritmo, o grupo de participação Cambridge Analytica, criou perfis detalhados de milhões de eleitores.

Segundo Henriksen (2019) o Projeto Ripon, como também foi classificado, era um programa de software que usava algoritmos sofisticados para permitir que as campanhas segmentem os eleitores em grupos com base em características psicológicas, como neurótico ou introvertido. Uma vez que os indivíduos foram identificados e agrupados, a plataforma então forneceu imagens pré-selecionadas e testadas em grupo e palavras-chave com maior probabilidade de alterar o comportamento desses indivíduos.

3.3 – Etapa 3 – Propagandas direcionadas

Em seguida, foram propostas ações específicas de acordo com as preferências dos usuários. Basicamente utilizando técnicas de publicidade para enviar anúncios e

mensagens políticas para as pessoas certas, no momento e lugares mais adequados. As estratégias online incluíam principalmente o Facebook, já que a empresa permite que o conteúdo patrocinado seja apresentado para perfis específicos de usuários.

Conforme Kanakia (2019), uma ampla base de informações pessoais prontamente disponíveis a partir da microssegmentação de indivíduos pode ser facilmente implantada. Mensagens direcionadas podem ser aplicadas para afetar seu comportamento, contornando os regulamentos existentes sobre divulgação e informação consentida. Esses fatores sugerem que os consumidores e os dados pessoais dos usuários sejam protegidos e que possam ser notificados da afiliação daqueles que procuram influenciá-los, e que eles tenham uma melhor oportunidade para participar como cidadãos informados.

É possível notar como que um mero aplicativo de perguntas e respostas no Facebook consegue traçar um perfil tão completo de um potencial eleitor para o candidato específico que no caso seria o alvo desejado. Analisando que a tecnologia e as ferramentas de ciência dos dados consigam resultados incríveis, ela ainda continua apenas um meio, e não a mensagem em si.

De acordo com Campos (2018) a estrutura de microssegmentação da Cambridge Analytica se dedicou a influenciar indivíduos coletando dados, construindo perfis psicológicos com base nas suas informações e os alimentando de anúncios conforme o seu perfil político específico. O método de mineração foi um questionário de personalidade em um aplicativo do Facebook desenvolvido por Acadêmico da Cambridge University Aleksandr Kogan. Quando o usuário (eleitor dos EUA para ser elegível) utilizou o quiz voluntariamente, as suas informações e de seus amigos seria extraído. Com base nos dados, um perfil de personalidade seria desenvolvido para medir a "abertura para experiência, consciência, extroversão, agradabilidade e neuroticismo" (HERN, 2018). Para completar os perfis, mais de 253 algoritmos foram produzidos por perfil para usar as curtidas do Facebook e os resultados de questionários para adivinhar traços de personalidade (HERN, 2018).

Usando esses perfis, a Cambridge Analytica direcionou grupos específicos com mensagens para fazer com que votem ou permaneçam em casa (CADWALLADR & GRAHAMHARRISON, 2018). A propaganda era não apenas individualizado ao longo das linhas políticas, mas também por personalidade. Essas operações eram semelhantes às usadas por marketing político (MURRAY, 2010; ZAFARANI, 2014), mas a formação de um perfil psicológico era principalmente exclusivo para o método de Kogen. O próprio sistema era projetado especificamente para influenciar o comportamento do

eleitor e exercer poder por meio de publicidade de personalidade específica. Isto é a intenção refletida em suas táticas em associação com campanhas políticas.

Pesquisas foram realizadas para avaliar se os canadenses mudariam seus hábitos de internet em resposta ao escândalo. Quase 73% disseram que diminuiriam seu uso, alterar hábitos ou mudar suas configurações de privacidade e 38% disseram que sua opinião sobre o Facebook piorou (CBC NEWS, 31 2018). No entanto, trimestralmente a taxa de crescimento de usuários ativos mensais do Facebook permaneceu praticamente inalterada (STATISTA, 2018). Apesar da relativa falta de impacto no uso da Internet, as implicações para este estudo de caso permanecem significativas. Em primeiro lugar, verifica-se que empresas de marketing personalizado têm esquemas desenvolvidos para exercer poder sobre usuários por meio não consensual dos sistemas de microssegmentação. No caso da Cambridge Analityca, eles procuraram ajudar a criar perfis psicológicos almejando determinados grupos com mensagens para influenciar o usuário. Em segundo lugar, o sucesso real de microssegmentação baseada em psicologia não é fortemente apoiada pela pesquisa empírica. No entanto, mais investigação é garantida por causa de lacunas de conhecimento na personalização do marketing político. Independentemente do impacto real da empresa sobre os usuários, é aparente que eles colaboraram com campanhas para beneficiar certas partes por tendências geopolíticas e tecnológicas.

4 CONSIDERAÇÕES FINAIS

O modelo de negócio desenvolvido pelas Big Techs está em vender a ideia de conforto e segurança para um serviço “gratuito” em troca de dados. Muitas vezes agimos de forma impulsiva sendo seduzida por propagandas direcionada e através dos cliques e curtidas as Big Techs são monetizadas. Entender o processo de controle e manipulação é um avanço que a sociedade vem observando como foi o caso da Cambridge Analytica que teve repercussão mundial.

Foi apresentado que através de uma grande variedade de atributos pessoais dos usuários, variando de orientação sexual a inteligência, pode ser inferida de forma automática e precisa usando seus likes no Facebook. Semelhança entre curtidas no Facebook e outros tipos difundidos nos registros digitais, como históricos de navegação, consultas de pesquisa ou o histórico de compra sugerem que o potencial de revelar os atributos dos usuários é improvável que seja limitado a likes. Além disso, a grande variedade de atributos previstos neste estudo indica que, com o dado adequado, pode ser possível revelar também outros atributos.

Prever os atributos e preferências individuais dos usuários pode ser usado para melhorar vários produtos e serviços. Por exemplo, sistemas e dispositivos digitais (como lojas online ou carros) podem ser projetados para ajustar seu comportamento para melhor se adequar ao perfil inferido de cada usuário (NASS, 2000). Além disso, a relevância do marketing e das recomendações de produtos pode ser melhorada com a adição de dimensões para modelos de usuário atuais. Por exemplo, seguro online e os anúncios podem enfatizar a segurança ao enfrentar as emoções dos usuários instáveis, mas enfatizam ameaças potenciais ao lidar com os emocionalmente estáveis. Além disso, os registros digitais de comportamento podem fornecer uma maneira conveniente e confiável de medir traços psicológicos. A avaliação automatizada com base em grandes amostras de comportamento pode não apenas ser mais precisa e menos propensa a trapaça e deturpação, mas também pode permitir avaliação ao longo do tempo para detectar tendências. Além disso, inferência baseada em observações de comportamento registrado digitalmente pode abrir novas portas para pesquisas em psicologia humana.

Por outro lado, a previsibilidade dos atributos individuais de registros digitais de comportamento pode ter implicações negativas, porque pode ser facilmente aplicado a um grande número de pessoas sem obter seu consentimento individual e sem eles

perceberem. Empresas comerciais, instituições governamentais, ou até mesmo seus amigos no Facebook podem usar software para inferir atributos como inteligência, orientação sexual ou pontos de vista políticos que um indivíduo pode não ter a intenção de compartilhar. Podem-se imaginar situações em que tais previsões, mesmo que incorretas, podem representar uma ameaça ao bem-estar, à liberdade ou mesmo à vida de um indivíduo. É importante notar que, dada à quantidade cada vez maior de rastros digitais, que as pessoas expõem, torna-se difícil para os indivíduos controlar quais seus atributos estão sendo revelados.

Os algoritmos desenvolvidos pelas corporações auxiliam a traçar ações para determinado objetivo como foi apresentado no estudo de caso da Cambridge Analityca, onde os usuários do Facebook e seus amigos foram monitorados através das curtidas, compartilhamentos e das propagandas direcionadas para que tivessem determinado comportamento e assim atingir o alvo. A tecnologia apresentada mostra na prática como as empresas costumam agir na internet e cabe aos usuários ter o devido conhecimento de que está sendo observado enquanto navega na rede. Este trabalho contribui para auxiliar a sociedade sobre como se dá a análise do comportamento através dos dados coletados na internet e abre uma oportunidade de pesquisa para futuros trabalhos nas áreas de big data, redes sociais e machine learning.

REFERÊNCIAS BIBLIOGRÁFICAS

ALBRIGHT, Jonathan. The Graph API: Key Points in the Facebook and Cambridge Analytica Debacle. 2018. Disponível em: https://medium.com/tow-center/the-graph-api-key-points-in-the-facebook-and-cambridge-analytica-debacle-b69fe692d747;

BRASIL. Lei complementar nº 166, de 8 de ABRIL de 2019. Diário Oficial da União, Brasília, 9 de abril de 2019, Seção I, p.1. Disponível em http://www.in.gov.br/materia/-/asset_publisher/Kujrw0TZC2Mb/content/id/70693213/do1-2019-04-09-lei-complementar-n-166-de-8-de-abril-de-2019-70693117;

BLACK MIRROR: Temporada 3, Episódio 1. Direção: Joe Wright. Produção: Michael Schur e Rashida Jones. Local: Netflix, 2016;

BUTLER D (2007) Data sharing threatens privacy. Nature 449(7163):644–645;

CABRAL, C. Quatro perspectivas para entender o caso Cambridge Analytica&Facebook. Crypto ID. 2018. Disponível em https://cryptoid.com.br/banco-de-noticias/quatro-perspectivas-para-entender-o-escandalo-da-cambridge-analytica/;

CADWALLADR, C. & Graham-Harrison, E. (2018, March 17). How Cambridge Analytica turned Facebook "likes" into a lucrative political tool. The Guardian. Retrieved from https://www.theguardian.com/news/2018/mar/17/cambridge-analyticafacebook-influence-us-election;

CAMPOS, Matt. Cambridge Analytica, Microtargeting, and Power: "A Full-Service Propaganda Machine" in the Information Age. 2018. Disponível em: https://trailsix.sites.olt.ubc.ca/files/2019/03/TRAIL_SIX_Volume_13_online.pdf#page=24. Acesso em: 31/03/21;

CAMBRIDGE ANALYTICA, 2103. Disponível em https://cambridgeanalytica.org;

CARDOSO, Paula. O CADASTRO POSITIVO E O RANKING DO HOMEM ENDIVIDADO. 2019. Disponível em: http://medialabufrj.net/blog/2019/05/dobras-32-o-cadastro-positivo-e-o-ranking-do-homem-endividado/. Acesso em 13/11/2020;

CBC NEWS. (2018, March 26). 73% of Canadians to change Facebook habits after data mining furor, survey suggests. CBC News. Retrieved from https://www.cbc.ca/news/technology/facebook-use-data-mining-angusreid-survey-1.4592371;

CHEN Y, Pavlov D, Canny JF (2009) Large-scale behavioral targeting. International Conference on Knowledge Discovery and Data Mining, pp 209–218;

CITRON, D. K., & PASQUALE, F. A. (2014). The scored society: due process for automated predictions. *Washington Law Review*, *89*.;

COALIZÃO DIREITOS NA REDE. Carta aberta sobre a reforma do cadastro positivo e proteção de dados pessoais. Brasília, 28 de março de 2018. Disponível em: <https://direitosnarede.org.br/p/reforma-do-cadastro-positivo-plp441/>

COSTA PT, McCrae RR (1992) Revised NEO Personality Inventory (NEO-PI-R) and NEO Five-Factor Inventory (NEO-FFI) Manual (Psychological Assessment Resources, Odessa, FL).;

COURTNEY Ackerman, Big Five Personality Traits: The OCEAN Model Explained 2021. Disponível em: https://positivepsychology.com/big-five-personality-theory/;

CREEMERS, Rogier. Planning Outline for the Construction of a Social Credit System (2014-2020). Disponível em: https://chinacopyrightandmedia.wordpress.com/2014/06/14/planning-outline-for-the-construction-of-a-social-credit-system-2014-2020/. Acesso em 11/11/2020;

DIGMAN, J. Personality structure: Emergence of the five-factor model. Annual Review of Psychology. 41: 417–440. 1990. Disponível em: https://www.annualreviews.org/doi/10.1146/annurev.ps.41.020190.002221;

DIENER E, Emmons RA, Larsen RJ, Griffin S (1985) The satisfaction with life scale. J Pers Assess 49(1):71–75.;

DONEDA, D., & MENDES, L. S. (2014). Data protection in Brazil: new developments and current challenges. In: Reloading Data Protection (pp. 3-20). Springer Netherlands.;

DUHIGG C (2012) The Power of Habit: Why We Do What We Do in Life and Business (Random House, New York);

DUSSEL, Julieta. Cómo ganar elecciones contando "me gusta". 2018. Disponível em: https://www.pagina12.com.ar/104359-como-ganar-elecciones-contando-me-gusta. Acesso em: 20/12/2020;

EYSENCK, H. J. (1993). The structure of phenoytypic personality traits: Comment. American Psychologist, 48 (12), 1299-1300;

FAST LA, Funder DC (2008) Personality as manifest in word use: Correlations with selfreport, acquaintance report, and behavior. J Pers Soc Psychol 94(2):334–346;

FINLAY, Steven (2014). Predictive Analytics, Data Mining and Big Data. Myths, Misconceptions and Methods (1st ed.). Basingstoke: Palgrave Macmillan. p. 237;

FLORES, P. O que a Cambridge Analytica, que ajudou a eleger Trump, quer fazer no Brasil. Nexo Jornal. 2017. Disponível em https://www.nexojornal.com.br/expresso/2017/12/08/O-que-a-Cambridge-Analytica-que-ajudou-a-eleger-Trump-quer-fazer-no-Brasil;

GALLOWAY Scott. Os Quatro Apple, Amazon, Facebook e Google. O Segredo dos Gigantes da Tecnologia. Editora: Alta Books. 2020;

GIL, Antônio Carlos, Como elaborar projetos de pesquisa, 4 ed. São Paulo: Atlas, 2002;

GOLDBERG LR, et al. (2006) The international personality item pool and the future of public-domain personality measures. J Res Pers 40(1):84–96.;

GOSILING SD, Ko SJ, Mannarelli T, Morris ME (2002) A room with a cue: Personality judgments based on offices and bedrooms. J Pers Soc Psychol 82(3):379–398;

GOLBECK J, Robles C, Turner K (2011) Predicting personality with social media. Conference on Human Factors in Computing Systems, pp 253–262;

HANNA, Mina J. Jim Isaak. User Data Privacy: Facebook, Cambridge Analytica, and Privacy Protection. 2018. Disponível em: https://ieeexplore.ieee.org/document/8436400;

HARRIS, E.; Warner, A. Dark Arts: How Cambridge Analytica Used Facebook to Find Out Who You Are. 2018. Disponível em https://medium.com/@thenib/dark-arts-how-cambridge-analytica-used-Facebook-to-find-out-who-you-are-d10b150b9653;

HENRIKSEN Ellen Emilie. Big data, microtargeting, and governmentality in cybertimes. The case of the Facebook-Cambridge Analytica data scandal. 2019. https://www.duo.uio.no/bitstream/handle/10852/69743/Master.pdf?sequence=1&isAllowed=y

HERN, A. (2018, May 6). Cambridge Analytica: how did it turn clicks into votes?: Whistleblower Christopher Wylie explains the science behind Cambridge Analytica's mission to transform surveys and Facebook data into a political messaging weapon. The Guardian. Retrieved from https://www.theguardian.com/news/2018/may/06/cambridgeanalytica-how-turn-clicks-into-votes-christopher-wylie;

IDEC, Instituto Brasileiro de Defesa do Consumidor. Entenda como funciona o novo cadastro positivo. 22 de fevereiro de 2019. Disponível em: <https://idec.org.br/dicas-e-direitos/entenda-como-funciona-o-novo-cadastro-positivo>.

INCE HO, Yarali A, Özsel D (2009) Customary killings in Turkey and Turkish modernization. Middle East Stud 45(4):537–551;

KANAKIA Harshil, Giridhar Shenoy and Jimit Shah. Cambridge Analytica – A Case Study. Indian Journal of Science and Technology, Vol 12(29), DOI: 10.17485/ijst/2019/v12i29/146977, August 2019;

KOSINSKI M, Kohli P, Stillwell DJ, Bachrach Y, Graepel T (2012) Personality and website choice. ACM Web Science Conference, pp 251–254;

KOSINSKI, Michal, David Stillwell, and Thore Graepel. Private traits and attributes are predictable from digital records of human behavior. 2013. Disponível em: https://www.pnas.org/content/pnas/110/15/5802.full.pdf;

KOSINSKI, Michal, S. C. Matz,, G. Navec, and D. J. Stillwelld, Psychological targeting as an effective approach to digital mass persuasion. 2017. Disponível em: https://www.researchgate.net/publication/321043573_Psychological_targeting_as_an_effective_approach_to_digital_mass_persuasion;

KOREN Y, Bell R, Volinsky C (2009) Matrix factorization techniques for recommender systems. Computer 42(8):30–37;

LAZER D, et al. (2009) Computational social science. Science 323(5915):721–723.

LOUREIRO, Rodrigo. Os dados são o novo petróleo. Disponível em: https://www.istoedinheiro.com.br/os-dados-sao-o-novo-petroleo/. Acesso em: 20/08/2020. 2018;

LORRAN, Tácio, Lucas Marchesini. Cadastro positivo: consumidores reclamam de atrasos e falhas. Disponível em: https://www.metropoles.com/brasil/economia-br/cadastro-positivo-consumidores-reclamam-de-atrasos-e-falhas. 2020;

MAIRESSE, F.; Walker, M. A.; Mehl, M. R.; and Moore, R. K.2007. Using Linguistic Cues for the Automatic Recognition of Personality in Conversation and Text. Journal of Artificial Intelligence Research 30(1): 457–500.

MENÁRGUEZ, Ana Torres. Os privilegiados são analisados por pessoas; as massas, por máquinas. El País, 21 de novembro de 2018. Disponível em: <https://brasil.elpais.com/brasil/2018/11/12/tecnologia/1542018368_035000.html?%3Fid_externo_rsoc=FB_BR_CM&fbclid=IwAR03NTMI0fdZU_7AYDxuMlFNgx7JidrDw9VSncTVihgFmtNqvvwpDAh13fU>;

MEYER, Maximiliano. Como funcionará a "pontuação de cidadão" que está sendo implementada na China. 2018. Disponível em: https://www.oficinadanet.com.br/tecnologia/22197-como-funciona-o-score-social-da-china. Acesso em: 11/11/2020;

MDIC – Ministério do Desenvolvimento, Indústria e Comércio Exterior (2014). Base de dados histórica. Fonte: http://www.desenvolvimento.gov.br;

MISTREE Jernigan C, BF (2009) Gaydar: Facebook friendships expose sexual orientation. First Monday 14(10);

MUSICK K, Meier A (2010) Are both parents always better than one? Parental conflict and young adult well-being. Soc Sci Res 39(5):814–830;

MURRAY, G. R. (2010). Microtargeting and electorate segmentation: Data mining the American national election studies. Journal of Political Marketing, 9(3), 143-166. Doi: 10.1080/15377857.2010.497732;

NASS, C. , & Moon, Y. (2000). Machines and mindlessness: Social responses to computers. Journal of Social Issues, 56(1), 81-103;

LOMBARDO Silvia. The Bad, the Good, and the Rebellious Bots: World's First in Artificial Intelligence. 2019. Disponivel em: https://www.igi-global.com/viewtitlesample.aspx?id=262835&ptid=242997&t=The%20Bad,%20the%20Good,%20and%20the%20Rebellious%20Bots:%20World%27s%20First%20in%20Artificial%20Intelligence&isxn=9781799834991;

OBERLANDER, J., and Nowson, S. 2006. Whose Thumb Is It Anyway? Classifying Author Personality from Weblog Text. In Proceedings of the 44th Annual Meeting of the Association for Computational Linguistics, 627-634. Stroudsburg, PA, USA: ACM Press.

ORWELL, G. (1952).1984. Círculo de Lectores S.A. Ediciones Destino, S.L;

O'NEIL, Cathy. Weapons of math destruction: How big data increases inequality and threatens democracy. Broadway Books, 2017.

PASQUALE, F. A. (2011). Restoring Transparency to Automated Authority. *Seton Hall Research Paper*, (2010-28). Available at: http://digitalcommons.law.umaryland.edu/cgi/viewcontent.cgi?article=2357&context=fac_pubs;

PENNEBAKER, J. W.; and King, L. A. 1999. Linguistic Styles: Language Use As an Individual Difference. Journal of Personality and Social Psychology 77: 1296–1312.

RAVEN JC (2000) The Raven's progressive matrices: Change and stability over culture and time. Cognit Psychol 41(1):1–48;

STATISTA. (2018). Number of monthly active Facebook users in the United States as of 3rd quarter 2018 (in millions). Retrieved from https://www.statista.com/statistics/264810/number-of-monthly-activefacebook-users-worldwide/;

SCHONFELD, E. Zuckerberg: We are building a web where the default is social. TechCrunch. https://techcrunch.com/2010/04/21/zuckerbergs-buildin-web-default-social/;

SHMATIKOV Narayanan A, V (2008) Robust de-anonymization of large sparse datasets. IEEE Symposium on Security and Privacy, pp 111–125;

SIMÃO, Barbara, BESSA, Leonardo, R.,PEREIRA, Laudelina. Pontuação de crédito, proteção de dados, transparência: uma difícil conciliação? X FÓRUM DA INTERNET NO BRASIL COMITÊ GESTOR DA INTERNET. 2020. Disponível em: https://forumdainternet.cgi.br/programacao/detalhe/2/1989/. Acesso em: 11/11/2020;

SYMEONIDIS, I.; Tsormpatzoudi, P.; Preneel, B. Colla damage of Facebook Apps: an enhanced privacy scoring model. IACR. 2015. Disponível em https://eprint.iacr.org/2015/456.pdf Washington Post. Cambridge Analytica's 'Ripon' brochure;

SCHUTZ A, Marcus B, Machilek F, (2006) Personality in cyberspace: Personal Web sites as media for personality expressions and impressions. J Pers Soc Psychol 90(6):1014-1031;

SCHIMMACK U, Diener E, Oishi S (2002) Life-satisfaction is a momentary judgment and a stable personality characteristic: The use of chronically accessible and stable sources. J Pers 70(3):345–384;

ZAFARANI, R., Abbasi, A. M., & Liu, H. (2014). Social Media Mining: An Introduction. Cambridge: Cambridge UP. Doi:10.1017/CBO9781139088510;

ZUBOFF, Shoshana. The Age of Surveillance Capitalism: The Fight for a Human Future ate the New Frontier of Power. PublicAffairs: New York, 2019. 705 p;

WASHINGTON POST. 2018. Disponível em: https://www.washingtonpost.com/apps/g/page/politics/cambridge-analyticas-ripon-brochure/2293/?noredirect=on. Acesso em: 02/03/21;

WEAVER Jesse, Paul Tarjan. Facebook Linked Data via the Graph API. 2012. Disponível em: http://www.cs.rpi.edu/~weavej3/papers/swj2012-fbld.pdf;