SOCIAL BIG DATA, ELEIÇÕES E FACEBOOK:

Indícios digitais de previsão eleitoral aplicados ao pleito presidencial de 2018 no Brasil [1]

SOCIAL BIG DATA, ELECTIONS AND FACEBOOK:

Digital traces about electoral prevision applied on the Brazilian presidential elections in 2018

Felipe Murta [2]
Leonardo Magalhães [3]
Raul Pimentel [4]

***Resumo**: Esta pesquisa dedica-se a entender se o grau de engajamento aos candidatos no Facebook é capaz de prever a intenção de votos no 1º turno das eleições presidenciais no Brasil em 2018. Para tal, foram coletadas cerca de 10 mil publicações de 9 diferentes campanhas na plataforma social durante os dias 1 de junho e 7 de outubro de 2018. Foram testados 90 modelos preditivos. Os resultados vão apontar correlações entre os dados observados e o resultado atingido nas urnas, reforçando as teorias que defendem a relevância desta fonte de dados como bons preditores eleitorais.*

***Palavras-Chave:** Mídias Socais; Eleições; Opinião Pública; Previsão Eleitoral.*

***Abstract**: This research is dedicated to observe, from digital data produced by Brazilian campaigns competing for the office of president of the republic in 2018 during the electoral period, if the candidate engagement degree in Facebook is a good predictor for the 1st round elections. To that end, 10.000 publications of 9 different campaigns on the social media were collected during June 1st and October 7th. The analyses tested 90 predictive models. The results will point out correlations between the data observed and the result achieved at the polls, reinforcing the theories that defend the relevance of these data source as being good predictors.*

***Keywords:** Social Media; Elections; Public Opinion; Elections Forecast*

[1] Trabalho apresentado ao Grupo de Trabalho Mídia e Eleições do VIII Congresso da Associação Brasileira de Pesquisadores em Comunicação e Política (VIII COMPOLÍTICA), realizado na Universidade de Brasília (UnB), de 15 a 17 de maio de 2019. O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES).

[2] Doutorando em comunicação Social pela PUC-Rio, femurpi@gmail.com.
[3] Doutorando em comunicação Social pela PUC-Rio e Coordenador de Projetos para a América Latina e Caribe em Atlas Político, leonardo_firmino@msn.com.
[4] Graduado em Comunicação Social - Jornalismo pela PUC-Rio, raulmpimentel@hotmail.com.

1 Introdução

Passadas as eleições de 2018 ficou notório para boa parte dos eleitores brasileiros o potencial das mídias digitais, em particular das redes sociais como eficientes plataformas de campanha eleitoral no que se refere à capacidade de disseminar mensagens por parte do candidato, dialogar com seus eleitores e impactar públicos segmentados com conteúdo específico. Com a massificação do uso do Facebook pelos diversos públicos que compõem a esfera pública online brasileira, abrem-se diversas possibilidades para entender como a plataforma em questão pode influenciar disputas eleitorais e, em particular modo, se e como o uso dessa ferramenta poderia ajudar a prever os resultados das eleições no Brasil.

Afirmar que estudos sobre a manifestação da opinião pública em redes sociais como o Facebook são capazes auxiliar na previsão de resultados eleitorais reforça a ideia de que a plataforma já vem sendo usada como meio de expressão da própria subjetividade, e de forma tão ampla que poderia considerar-se como análoga à uma amostra representativa da população brasileira. Com as devidas estratificações amostrais seria possível realizar estudos tão ou mais incisivos que os meios tradicionalmente usados para este fim – *surveys* e modelos estruturais, porém com a vantagem de poder realizar, em tempo real, análises mais profundas e variadas. De fato, o estudo da opinião pública nas redes sociais só é possível graças à exposição pública e voluntária da própria opinião em diversos formatos e meios. Tais dados ficam armazenados e possuem um rico repertório de informações, onde acredita-se ser possível, inclusive, entender suas conexões com outros elementos online.

Nesta linha, o presente estudo é dividido em quatro partes incluindo esta introdutória. Em seguida, uma breve discussão teórica acerca das pesquisas que sustentam e as que consideram inconclusivo o uso de dados digitais coletados em redes socais durante períodos eleitorais como índices validos e influentes no processo de previsão de votos. Em terceiro vem a metodologia da pesquisa, a qual se baseou em analisar métricas quantitativas produzidas durante o período de campanha presidencial no Facebook para identificar se houve, de fato, indícios que possam

apontar para um modelo de previsão eleitoral voltado para a interações do público com os candidatos. Sobre isso, é necessário lembrar que esta pesquisa reconhece sua importância parcial por considerar fundamental e de igual relevância uma análise qualitativa de sentimento sobre as interações contabilizadas. A pesquisa se desenvolve na busca por comprovar a hipótese (H_1) de que o engajamento no Facebook aos candidatos no 1º turno das eleições presidenciais do Brasil em 2018 prevê a intenção de voto aos mesmos. Para tanto, mediante a plataforma Atlas Político, foram coletadas todas as postagens dos candidatos à presidência da república para calcular o índice diário de engajamento. No que se refere aos dados de intenção de voto, os mesmos foram obtidos mediante *surveys* diários. Na conclusão, parte final, e em respeito à análise dos dados e resultados, foi encontrada uma forte correlação entre a intenção de voto nas urnas no primeiro turno e o engajamento aos candidatos no Facebook. Neste cenário, isoladamente observado, o Facebook se demonstrou um razoável indicador de intenção de voto no primeiro turno das eleições presidenciais de 2018 no Brasil. Mediante regressão lineares múltiplas foram testados 90 modelos combinatórios ao todo, confirmando a hipótese de relação causal entre as variáveis.

2 Índices Digitais e a Previsão Eleitoral

Não é de hoje que a busca por previsões de votos em períodos eleitorais mobiliza o mundo acadêmico. O uso cada vez mais disseminado e frequente das mídias digitais por parte da sociedade levou estudiosos a perceber a necessidade de incorporar os dados por elas produzidos nas recentes pesquisas de previsão eleitoral, indo muito além dos métodos costumeiramente utilizados nas pesquisas tradicionais de intenção de voto (FIGUEIREDO, 2007).

Mesmo com o recente protagonismo assumido pelas redes sociais, especificamente no contexto eleitoral, são usuais os trabalhos inseridos tanto no campo da Comunicação Política Digital quanto nas Ciências Sociais e da Computação que defendem a influência de determinados tipos de dados digitais em processos eleitorais. Porém, bilhões de usuários ativos compartilhando incessantemente mensagens e arquivos fizeram com que a relação entre mídia social e campanha

eleitoral fosse considerada parte fundamental não apenas do ambiente de mídia atual como do processo eleitoral. Dado o grande volume de informações pessoais que produzem, essas novas mídias vêm sendo cada vez mais exploradas para fins de monitoramento e previsão de eventos no mundo. Não à toa muitos consideram que a evolução tecnológica dos meios de comunicação, assim como a amplificação do acesso ao ambiente online em escala global, fizeram com que estudos na área de política passassem a coletar dados que os eleitores produziam nas redes sociais e outras informações extraídas do ambiente online para buscar validações em suas análises e conclusões científicas (DI GRAZIA, *et al.*, 2013; KRISTENSEN, *et al.*, 2017; SAEZ-TRUMPER, MEIRA e ALMEIDA, 2011; TUMASJAN, *et al.*, 2013). Twitter, YouTube, Facebook, Instagram e WhatsApp são exemplos de redes que desde 2008 estão sendo, cada uma no seu tempo, massivamente utilizados por campanhas eleitorais na disseminação do seu conteúdo diretamente para o seu público sem a necessidade de intermediários, papel até então desempenhado pela mídia tradicional.

Ao investigar se as redes sociais possuem indicadores válidos e expressivos sobre o comportamento político do eleitor, autores como Tumasjan, Sprenger e Sandner (2010) dedicaram-se a observar o comportamento de pessoas em plataformas digitais durante campanhas eleitorais e escolheram o Twitter como objeto de suas pesquisas por apresentar características propícias ao debate político, discussões, trocas de ideias e acesso à informações globais em tempo real. Argumentavam que, como previsor de resultado de eleições, a simples contagem de menções a partidos e/ou candidatos poderia ser capaz de acompanhar pesquisas eleitorais tradicionais e apresentar resultados próximos aos resultados finais da eleição (MIRANDA et al. 2014). Metodologia com tendência semelhante também pôde ser vista com Trumper, Meira e Almeida (2011). Os pesquisadores realizaram pequenas, mas significativas mudanças na metodologia de Tumasjan *et al.*, (2010). Ao invés de considerar todas as menções de todos os perfis analisados, os cálculos das projeções foram feitos considerando apenas uma menção a um termo de cada (TRUMPER, MEIRA e ALMEIDA, 2011). No final da pesquisa, obtiveram resultados satisfatórios de previsão eleitoral com um erro absoluto médio de 4,07 pontos percentuais.

Outros pesquisadores (BENEVENUTO et al. 2010; LUMEZANU et al. 2012) optaram por debruçar-se sobre a identificação de perfis pessoais de usuários nas redes sociais. Di Grazia, *et al.* (2013), por sua vez, mostrou em pesquisa sobre o cenário político dos Estados Unidos que a porcentagem de menções a candidatos republicanos no Twitter durante parte do ano de 2010 se correlacionam com a margem de votos que os mesmos obtiveram nas eleições daquele ano. Os autores também trazem o debate sobre a falta de ênfase que os recentes estudos da área dão a variáveis como a incumbência dos candidatos, o partido ao qual eles pertencem, cobertura midiática e a composição sociodemográfica do eleitorado, mostrando que, mesmo levando em consideração essas variáveis, a correlação se mantinha (DI GRAZIA, *et al.*, 2013). As redes sociais, segundo os autores, podem ser um indicador válido do eleitorado americano. São muitos os estudos sobre as características das contas de eleitores, que assiduamente interagem com conteúdo político de campanha. Assim como são inúmeros e extensos os trabalhos que abordam o comportamento específico dos perfis voltados para a disseminação de informação política em período eleitoral (LUMEZANU et al. 2012), identificando padrões de uso e mensagens (GHOSH, 2012).

Há também os estudos analíticos quantitativos focados em comprovar se as formas conhecidas de interação entre eleitor e campanha eleitoral nas redes sociais, como por exemplo o compartilhamentos, a curtidas e o *retweets*, entre outras modalidades, podem ser aceitas como índices relevantes e capazes de alimentar pesquisas de intenção de voto em um sistema multipartidário de eleições democráticas (KRISTENSEN, *et al.*, 2017). Kristensen produziu notório levantamento sobre o comportamento dos eleitores nas redes sociais durante as eleições locais na Dinamarca (2017). Para tal, realizou pesquisas tradicionais de intenções de voto e as correlacionou com as informações obtidas em bancos de dados coletado a partir das publicações em páginas públicas de partidos e políticos dinamarqueses em campanha entre janeiro de 2015 e janeiro de 2017. O objetivo era não apenas comprovar se havia alguma relação direta entre os resultados da pesquisa tradicional com os dados digitalmente coletados, como também comprovar se é possível identificar as preferências políticas do eleitor a partir da forma como estes interagiram com as campanhas eleitorais (KRISTENSEN, *et al.*, 2017). Em posse de novos métodos para

previsões eleitorais, o autor afirmou ser possível não somente separar público por perfis e preferências políticas como também generalizar as conclusões para níveis nacionais e até além da fronteira. Assim, Kristensen desafiou a tendência contemporânea de buscar métodos baseados em *big data* para atingir conclusões científicas ao mostrar que uma curtida a uma publicação ou página no Facebook pode ter grande importância científica (KRISTENSEN, *et al.*, 2017).

Além disso, argumentou que a maioria dos estudos de previsão de comportamento individual de eleitores consideram contextos bipartidários, como no caso da política americana. O contexto eleitoral em questão, portanto, é uma variável determinante neste caso, já que o sistema de partidos define qual a melhor metodologia a ser empregada em cada cenário. Um esquema metodológico para sistemas multipartidários também foi apresentado por Tumasjan, *et al.*, (2013). Os pesquisadores concluem não só que o número de menções ao partido refletiu o resultado das eleições, como também que o sentimento das mensagens corresponde às preferências políticas dos eleitores (TUMASJAN, *et al.*, 2013).

Pesquisas como essas geraram muitos testes metodológicos (JUNGHERR, 2012; GAYO-AVELLO, 2010) sob a premissa de buscar uma confirmação sobre algumas das hipóteses anteriormente mencionadas. Muitos autores acabaram não comprovando a afirmativa de que é possível prever um resultado eleitoral apenas com índices quantitativos como, por exemplo, a contagem de menções. Foram apontadas inúmeras limitações básicas como a falta de regras bem fundamentadas para a coleta dos dados e o recorte temporal ideal para esse tipo de análise de previsão. Entre os testes metodológicos mais relevantes podemos destacar a análise feita sob tweets coletados durante as eleições federais na Alemanha em 2009 (JUNGHERR, 2012) e para o congresso norte americano em 2010 (GAYO-AVELLO, 2010). Ambos não observaram nenhuma relação entre os dados coletados e o resultado das urnas. Gayo-Avello chegou a realizar uma grande pesquisa bibliográfica (2014) sobre trabalhos que objetivassem previsões eleitorais a partir de dados coletados em plataformas sociais digitais, focando sua análise em levantar erros, apontar falhas e limitações metodológicas até então empregadas.

Outra grande deficiência recorrente em pesquisas de monitoramento de rede e previsões eleitorais (MUKHERJEE et al. 2013) é a falta de mecanismos utilizados para

distinguir, por exemplo, uma simples menção a um candidato de um voto em si. Essa lacuna evidenciou, entre tantas outras coisas, à importância da compreensão sobre o que é publicado textualmente por esses usuários. Assim foi introduzida a análise de sentimento sobre o conteúdo produzido por usuários de redes sociais durante períodos eleitorais como sendo uma determinante que não pode ser ignorada por quem busca associar dados coletados nas mídias digitais com o resultado obtido em disputas eleitorais (BORA, 2014).

Está claro e evidente, portanto, que relativizações são necessárias nesse campo de estudo. Pesquisadores apontam, também, o foco em entretenimento e expressões emocionais do fluxo de comunicação de todas as redes como um desses problemas (DI GRAZIA, et al., 2013). O chamado *digital divide*, um fenômeno que ocorre porque parte da população mundial não tem (ou tem pouco) acesso pleno à internet ou mesmo a recursos que possibilitam uma conexão online no meio digital, também é mencionado, gerando fortes questões sobre a real presença da população do planeta, de forma representativa, no ambiente online. Esse problema específico pode ser atenuado quando a metodologia do estudo trabalha com *big data* (FAN e BIFET, 2013). Porém, a amostra de pessoas em determinada rede social pode ser enviesada, já que todo serviço online, por sua natureza, atrai determinados segmentos populacionais como consumidores. Por consequência, podemos identificar perfis de público distintos e que podem não representar amostras adequadas para estudos científicos (DI GRAZIA, et al., 2013).

Todos os fatores anteriormente mencionados levam a uma ideia de que a escolha da mídia para a realização de estudos de previsão eleitoral pode ser determinante, principalmente considerado as diferenças técnicas e estruturais de cada uma, o que por vezes pode impor mudanças ou formas de funcionamento de metodologias de estudo ou mesmo influenciar seus resultados. Kristensen (2017), por exemplo, defende que estudos na área de previsão eleitoral que se baseiam no Facebook e no Twitter mostram melhores resultados do que estudos que usam outras mídias. No campo, há pesquisas que mostram o potencial de previsão eleitoral baseados em dados digitais de plataformas como o YouTube, Google e até o Wikipedia (YASSERI e BRIGHT, 2016).

3 Metodologia

O presente estudo analisou, via métricas, a comunicação política no Facebook realizada por concorrentes ao cargo de presidente da República durante o período de campanha, para identificar se há um modelo de previsão eleitoral baseado nas interações do público com os candidatos no Facebook. Assim, temos a nossa hipótese (H_1):

H_1: O grau de engajamento ao candidato no Facebook prevê a intenção de voto no mesmo, no 1º turno das eleições presidenciais do Brasil em 2018.

Com o fim de entender se o engajamento ao candidato no Facebook é consistente como preditor eleitoral, mediante a plataforma Atlas Político, foram duas fontes de dados. A primeira se refere ao engajamento diário do candidato no Facebook, que será a nossa variável independente, obtida mediante a *Dashboard* de *Big Data* do Atlas político, que trabalha com a API da rede social. A segunda fonte de dados, como variável dependente, é a pesquisa sobre intenção de voto, realizado diariamente mediante a plataforma Atlas Tracking.

No que se refere ao Facebook foram coletadas cerca de 10 mil postagens, provenientes das páginas dos candidatos à Presidência da República, entre 1 de junho e 7 de outubro de 2018. Os candidatos monitorados foram: Jair Bolsonaro (PSL), Fernando Haddad (PT), Ciro Gomes (PDT), Geraldo Alckmin (PSDB), João Amoedo (Novo), Henrique Meirelles (MDB), Marina Silva (Rede), Alvaro Dias (Podemos) e Guilherme Boulos (PSOL).

Cada postagem foi classificada pelo seu índice de engajamento, dado pela soma das reações, comentários e compartilhamentos. Em seguida, foi obtido o engajamento total diário de cada candidato, realizando a somatória do engajamento de cada sua postagem no período estudado.

Respeito às pesquisas de intenção de voto, o Atlas Tracking é uma plataforma que realiza *surveys* diários de opinião online. O questionário é publicado e recolhe entre 800 e 2000 respostas diariamente em todo o território nacional. Uma vez encerrado o tempo de resposta, e realizada a coleta definitiva dos dados, é feita uma pós-estratificação com balanceamento amostral, baseada em sete variáveis chave,

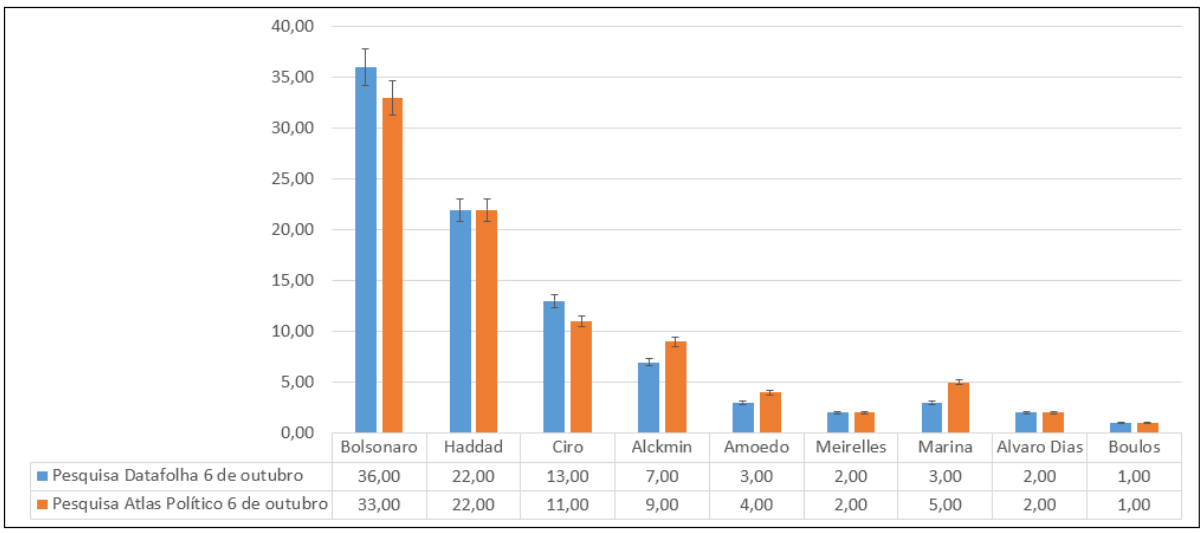
além da verificação geográfica mediante endereço IP como forma de controle adicional.

No que tange mais especificamente o desenho amostral, trata-se, portanto, de um estudo pós-estratificado, representativo da população eleitoralmente ativa, com seleção aleatória em todos os estados do Brasil. As respostas finais têm um peso amostral ajustado por um algoritmo de ranqueamento para garantir a representatividade mediante sete variáveis chave de estratificação, como gênero, idade, renda mensal familiar, local de residência, uso das redes sociais e voto no 1º e 2º turno das últimas eleições para presidente da república. A margem de erro é de ± 2% e o intervalo de confiança é de 95%.

Se compararmos o Atlas Tracking com o Datafolha, teremos somente uma discrepância de até 2% em 95% dos dados, o que, de certa forma, valida os dados utilizados no presente artigo. A comparação pode ser observada no GRAF. 1 a seguir:

GRÁFICO 1

Comparação entre Atlas Tracking e Datafolha em 6 de outubro de 2018.



	Bolsonaro	Haddad	Ciro	Alckmin	Amoedo	Meirelles	Marina	Alvaro Dias	Boulos
Pesquisa Datafolha 6 de outubro	36,00	22,00	13,00	7,00	3,00	2,00	3,00	2,00	1,00
Pesquisa Atlas Político 6 de outubro	33,00	22,00	11,00	9,00	4,00	2,00	5,00	2,00	1,00

FONTE: Elaboração própria.

Com uma série final de 115 unidades de análises, relativas a cada dia estudado, temos os dados diários de engajamento ao candidato e intenção de voto ao mesmo. Com ditos dados, foi realizada uma matriz de correlação de Pearson. Em seguida, concentrando-nos exclusivamente nos coeficientes de correlação relativos ao

engajamento e intenção de voto de cada candidato, foram realizadas regressões lineares múltiplas para entender se é possível afirmar que há um modelo de previsão eleitoral baseado no engajamento no Facebook. O software utilizado para a análise foi o SPSS.

Antes de realizar as regressões lineares, foram garantidas todas as condições prévias para o uso de modelos paramétricos como:

- ✓ distribuição normal e simétrica das variáveis dependentes;
- ✓ relação linear entre as variáveis dependentes e independentes;
- ✓ n > 30;
- ✓ independência dos valores;
- ✓ distribuição normal dos resíduos;
- ✓ homocedasticidade (independência dos resíduos).

Uma vez garantidas todas as condições ideais para a realização de estatísticas paramétricas, foram testados ao todo 90 modelos de previsão e selecionados o que oferecia um valor de R^2 ajustado maior para cada variável dependente.

Já que se buscou entender como a performance dos candidatos no facebook impactou a opinião pública, gerando um efeito sobre a intenção de voto, optou-se por estudar o impacto mútuo entre diversas campanhas entre si. Por esta razão foram realizadas regressões lineares múltiplas em lugar das lineares simples.

4 Apresentação dos resultados

A seguir pode-se ler a TAB. 1 que mostra a matriz de correlação de Pearson. Como pode-se observar, há valores altos, acima de 0,7 e valores baixos, de 0,2. Tratou-se de não descartar os valores pequenos (inferiores entre -0,3 e 0,3), pois somente a regressão dirá se as correlações são significativas ou não. Assim, por exemplo, poder-se-ia ter um uma correlação de 0,1 com maior significância que outra de 0,7.

TABELA 1

Matriz de correlação de Pearson

		Engajamento no Facebook								
		Bolsonaro	Haddad	Ciro	Alckmin	Amoedo	Meirelles	Marina	Alvaro Dias	Boulos
Intenção de Voto	Bolsonaro	0,772668	0,82148	0,690386	0,468526681	0,613643	0,559523	0,580131	0,4024948	0,43104
	Haddad	0,771551	0,79389	0,669228	0,502908484	0,608433	0,573569	0,558526	0,3858573	0,44742
	Ciro	0,677502	0,76349	0,616228	0,448435305	0,434198	0,514023	0,488581	0,3408383	0,31619
	Alckmin	0,524813	0,58916	0,487984	0,337442535	0,357502	0,384803	0,393336	0,204047	0,30749
	Amoedo	0,588843	0,59382	0,434798	0,488296189	0,70473	0,425705	0,526545	0,269933	0,26031
	Meirelles	0,435283	0,57391	0,418292	0,313698742	0,248332	0,401896	0,32302	0,286139	0,17791
	Marina	-0,46807	-0,43935	-0,449121	-0,108871222	-0,21669	-0,357207	-0,343242	-0,288283	-0,3803
	Alvaro Dias	-0,65136	-0,6455	-0,579714	-0,581487091	-0,648525	-0,488304	-0,534023	-0,26238	-0,31694
	Boulos	0,116386	0,13737	0,063556	0,075855564	0,11612	0,103853	0,006013	0,08412	0,03963

FONTE: Elaboração própria.

Já que correlações são somente indicativos iniciais que não indicam necessariamente um fenômeno de causação, realizamos regressões múltiplas mediante modelagem linear no SPSS. Utilizamos como preditores de entrada todas as variáveis sobre engajamento, com nível de confiança de 95%, com método de seleção dos modelos em base aos melhores subconjuntos, usando como parâmetro comparativo, de entrada e saída, o R^2 ajustado. Ao todo foram testados 90 modelos preditivos e selecionados os que obtiveram os mais altos índices para cada variável.

Tem-se a TAB. 2, de sumarização dos modelos de regressão para a variável de intenção de voto por candidato. Na TAB. 2 reportamos somente o melhor modelo para cada variável dependente. Na horizontal vemos as variáveis dependentes “intenção de voto” e na vertical temos as variáveis independentes “data” e “engajamento aos candidatos no Facebook”. As marcações indicam a presença ou ausência de cada variável independente no modelo de previsão da dependente. A linha relativa à R^2 ajustado indica o grau de previsibilidade de cada modelo, expresso nas colunas. As marcações em verde mostram as variáveis que têm mais peso no modelo ($p < 0,05$ e na maioria dos casos $p < 0,01$). As marcações em vermelho indicam variáveis de engajamento que têm um peso muito pequeno no modelo e se se considerarmos somente graus de significância de $p > 0,05$.

TABELA 2

Sumarização dos modelos de regressão múltipla

		Intenção de Voto								
		Bolsonaro	Haddad	Ciro	Alckmin	Amoedo	Meirelles	Marina	Alvaro Dias	Boulos
R^2 ajustado		,653	,521	,624	,641	,600	,342	,262	,814	,101
Engajamento	Bolsonaro	✓	✓	✓		✓	✓	✓		✓
	Haddad	✓	✓	✓				✓	✓	✓
	Ciro	✓	✓	✓	✓	✓	✓	✓	✓	
	Alckmin	✓						✓	✓	
	Amoedo		✓	✓	✓	✓	✓	✓		✓
	Meirelles	✓				✓		✓		✓
	Marina		✓			✓				✓
	Alvaro Dias				✓	✓			✓	
	Boulos	✓	✓	✓		✓		✓		✓
Data		✓	✓	✓	✓	✓	✓		✓	✓

FONTE: Elaboração própria.

A seguir, da TAB. 3 à TAB. 11, pode ser contemplada a sumarização dos modelos de previsão por coeficientes. A previsão da intenção de voto aos candidatos em base às variáveis de engajamento. Na seguinte tabela é onde vemos o peso que cada variável tem no modelo. Foi usado um parâmetro de significância de 1,000 para a formulação de cada modelo, mas, como pode ser visto evidenciado em cor laranja, consideramos mais relevantes somente os que contêm uma significância de até 0,05. Observa-se que há variáveis com muito peso no modelo e outras com menor peso.

TABELA 3
Bolsonaro

Termo Modelo	Coeficiente ▶	Sig.	Importância
Intercepto	-6,593	,001	
jair_bolsonaro_fb_engagement_transformed	0,000	,001	0,346
Data_months	0,012	,001	0,332
guilherme_boulos_fb_engagement_transformed	-0,000	,058	0,101
ciro_gomes_fb_engagement_transformed	0,000	,118	0,068
haddad_fb_engagement_transformed	0,000	,121	0,067
henrique_meirelles_fb_engagement_transformed	0,000	,172	0,052
alckmin_fb_engagement_transformed	-0,000	,259	0,035

TABELA 4
Haddad

Termo Modelo	Coeficiente ▶	Sig.	Importância
Intercepto	12,488	,000	
Data_months	-0,021	,000	0,304
joao_amoedo_fb_engagement_transformed	0,000	,000	0,190
guilherme_boulos_fb_engagement_transformed	0,000	,001	0,169
ciro_gomes_fb_engagement_transformed	-0,000	,002	0,140
jair_bolsonaro_fb_engagement_transformed	-0,000	,024	0,074
haddad_fb_engagement_transformed	-0,000	,026	0,071
marina_silva_fb_engagement_transformed	0,000	,057	0,052

TABELA 5
Ciro

Termo Modelo	Coeficiente ▶	Sig.	Importância
Intercepto	-6,306	,000	
Data_months	0,011	,000	0,388
joao_amoedo_fb_engagement_transformed	-0,000	,000	0,199
jair_bolsonaro_fb_engagement_transformed	0,000	,002	0,139
guilherme_boulos_fb_engagement_transformed	-0,000	,003	0,128
ciro_gomes_fb_engagement_transformed	0,000	,004	0,124
haddad_fb_engagement_transformed	0,000	,200	0,023

TABELA 6
Alckmin

Termo Modelo	Coeficiente ▶	Sig.	Importância
Intercepto	-8,111	,000	
Data_months	0,014	,000	0,740
joao_amoedo_fb_engagement_transformed	-0,000	,000	0,176
ciro_gomes_fb_engagement_transformed	0,000	,002	0,063
alvaro_dias_fb_engagement_transformed	-0,000	,079	0,020

TABELA 7
Amoedo

Termo Modelo	Coeficiente ▶	Sig.	Importância
Intercepto	-0,851	,073	
joao_amoedo_fb_engagement_transformed	0,000	,000	0,560
ciro_gomes_fb_engagement_transformed	-0,000	,004	0,147
jair_bolsonaro_fb_engagement_transformed	0,000	,042	0,071
Data_months	0,001	,068	0,057
guilherme_boulos_fb_engagement_transformed	-0,000	,074	0,055
henrique_meirelles_fb_engagement_transformed	0,000	,104	0,045
alvaro_dias_fb_engagement_transformed	0,000	,163	0,033
marina_silva_fb_engagement_transformed	0,000	,173	0,032

TABELA 8
Meirelles

Termo Modelo	Coeficiente ▶	Sig.	Importância
Intercepto	-1,666	,000	
Data_months	0,003	,000	0,477
joao_amoedo_fb_engagement_transformed	-0,000	,001	0,348
ciro_gomes_fb_engagement_transformed	0,000	,035	0,138
jair_bolsonaro_fb_engagement_transformed	0,000	,271	0,037

TABELA 9
Marina

Termo Modelo	Coeficiente ▶	Sig.	Importância
Intercepto	0,094	,000	
jair_bolsonaro_fb_engagement_transformed	-0,000	,004	0,321
alckmin_fb_engagement_transformed	0,000	,011	0,257
joao_amoedo_fb_engagement_transformed	0,000	,073	0,124
ciro_gomes_fb_engagement_transformed	-0,000	,134	0,087
henrique_meirelles_fb_engagement_transformed	-0,000	,158	0,077
haddad_fb_engagement_transformed	-0,000	,170	0,073
guilherme_boulos_fb_engagement_transformed	0,000	,206	0,061

TABELA 10
Alvaro Dias

Termo Modelo	Coeficiente ▶	Sig.	Importância
Intercepto	5,200	,000	
Data_months	-0,009	,000	0,944
alckmin_fb_engagement_transformed	-0,000	,021	0,029
ciro_gomes_fb_engagement_transformed	-0,000	,119	0,013
haddad_fb_engagement_transformed	0,000	,194	0,009
alvaro_dias_fb_engagement_transformed	0,000	,293	0,006

TABELA 11
Boulos

Termo Modelo	Coeficiente	Sig.	Importância
Intercepto	0,489	,026	
haddad_fb_engagement_transformed	0,000	,005	0,347
Data_months	-0,001	,028	0,208
marina_silva_fb_engagement_transformed	-0,000	,068	0,143
guilherme_boulos_fb_engagement_transformed	-0,000	,069	0,142
henrique_meirelles_fb_engagement_transformed	0,000	,243	0,058
joao_amoedo_fb_engagement_transformed	0,000	,253	0,056
jair_bolsonaro_fb_engagement_transformed	0,000	,298	0,046

5 Análise dos dados

A metodologia de analise consistiu em realizar dois tipos de coleta de dados. O índice de engajamento diário no facebook e o tracking diário de intenção de voto.

No que tange à correlação entre engajamento e intenção de voto, foi evidenciado que há correlações discrepantes entre os dados de cada candidato. De modo geral há graus de correlação suficientemente altos, como pode-se contemplar na TAB. 1.

Respeito às regressões, o primeiro ponto a salientar é o relativo ao teste de significância total e dos coeficientes do modelo.

O R^2 ajustado indica o quanto cada modelo de regressão linear múltipla consegue prever a variável dependente, ou seja, quais variáveis de engajamento no facebook conseguem explicar a variável dependente intenção de voto.

De modo geral, vemos pela TAB. 12 a seguir, que se obtiveram resultados que confirmam a hipótese de previsão eleitoral mediante o engajamento. Não obstante a percentual pequena de previsibilidade para Boulos (10%), Marina (26%) e Meirelles (34%), os dados possuem alta significância. Todos os demais candidatos apresentam índices elevados de previsibilidade eleitoral pelo engajamento no Facebook, a exceção de Alvaro Dias, que obteve uma forte presença do elemento “data” no seu modelo, o que o invalida quase por completo neste caso específico.

TABELA 12

Grau de previsibilidade dos modelos por candidato

	Candidatos								
	Bolsonaro	Haddad	Ciro	Alckmin	Amoedo	Meirelles	Marina	Alvaro Dias	Boulos
R^2 ajustado (%)	65,3%	52%1	62,4%	64,1%	60%	34,2%	26,2%	81,4%	10,1%

FONTE: Elaboração própria.

Por outro lado, pode-se observar que não sempre a variável de engajamento do próprio candidato explica de forma mais adequada o crescimento o decrescimento de sua intenção de voto. Como é possível ver no caso de Boulos, o engajamento dos públicos ao Haddad explica boa parte da performance eleitoral do pessolista. Este indício faz muito sentido, pois ambos compartilham uma parcela de eleitorado à esquerda. Quando se compartilha os mesmos públicos, o crescimento de um significa o decrescimento do outro.

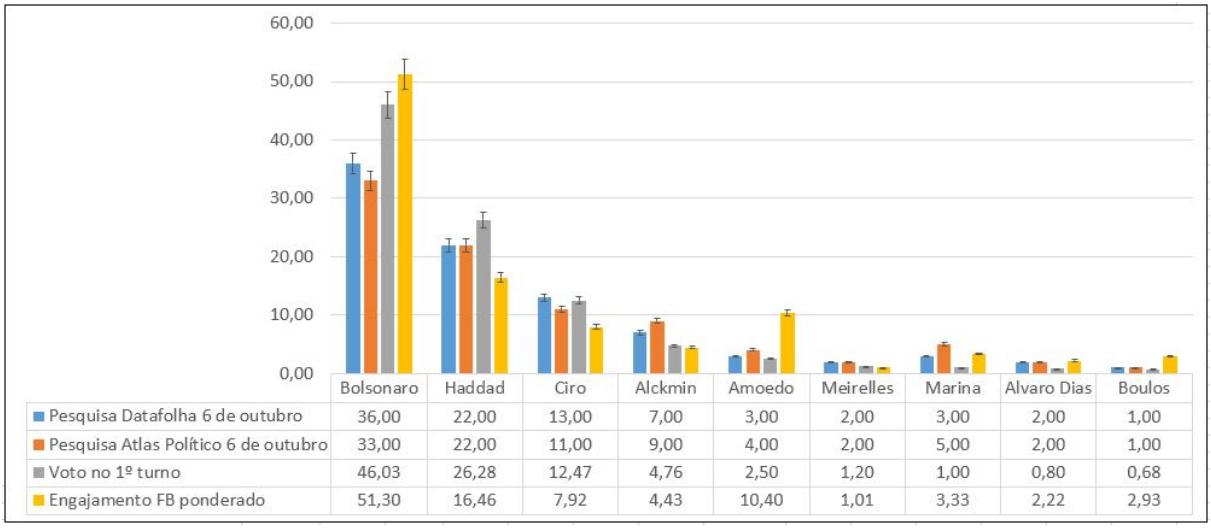
Outro elemento interessante a ser levado em consideração é o peso que a variável Data possui para diversos candidatos. Esta variável foi testada como independente em todos os modelos para entender se o fator de proximidade ao dia da votação mostraria algum impacto sobre a intenção de voto. Em alguns casos a data teve um peso enorme sobre o modelo, como no caso de Alckmin e Alvaro Dias, em outros o peso foi moderado ou marginal. Este fator também é importante na hora de considerar o engajamento um bom preditor.

Por outro lado, pode-se apresentar outra abordagem aos dados para realizar uma previsão eleitoral. Não obstante para tal abordagem não seja possível realizar uma regressão, pode-se observar claramente no GRAF. 3, a seguir, como o total ponderado (a 100%) dos valores também representam indícios de que o engajamento no Facebook pode ser usado como abordagem indicativa à previsão eleitoral.

O GRAF. 3 representa o resultado do 1º turno, comparando-o ao Datafolha, ao Atlas Tracking e à média de engajamento total ponderado no Facebook. O gráfico demonstra que o engajamento consegue oferecer resultados bastante parecidos ao resultado da votação e às pesquisas de opinião, inclusive mais precisos em alguns casos.

GRÁFICO 3

Comparação entre preditores eleitorais (%)



	Bolsonaro	Haddad	Ciro	Alckmin	Amoedo	Meirelles	Marina	Alvaro Dias	Boulos
Pesquisa Datafolha 6 de outubro	36,00	22,00	13,00	7,00	3,00	2,00	3,00	2,00	1,00
Pesquisa Atlas Político 6 de outubro	33,00	22,00	11,00	9,00	4,00	2,00	5,00	2,00	1,00
Voto no 1º turno	46,03	26,28	12,47	4,76	2,50	1,20	1,00	0,80	0,68
Engajamento FB ponderado	51,30	16,46	7,92	4,43	10,40	1,01	3,33	2,22	2,93

FONTE: Elaboração própria.

6 Conclusões

O presente estudo teve como objetivo identificar, no ambiente online, índices capazes de reforçar teorias que defendem a influência direta de dados digitais no processo eleitoral. Mais precisamente, a pesquisa busca identificar nas métricas de engajamento no Facebook algum potencial capaz de auxiliar na previsão de votos em disputas eleitorais no Brasil, que pudesse ser comparado com os meios tradicionalmente usados para este tipo de pesquisa, realizadas mediante *survey* ou modelos estruturais, mais comuns nos Estados Unidos. Com base na leitura dos dados digitais extraídos, analisados e dos resultados demonstrados nos gráficos e tabelas acima, o estudo concluiu que embora a média dos valores pudesse refletir ótimos indícios sobre o resultado final do pleito em questão e, em alguns casos, até mesmo o percentual aproximado alcançado individualmente por cada campanha, as regressões realizadas ajudam a reforçar as teorias que corroboram com a existência deste tipo de modelo de previsão eleitoral.

De maneira geral podemos identificar nos dados vários aspectos interessantes a serem aprofundados em futuros trabalhos, como por exemplo, a clara presença de mutua influência entre candidatos que compartilham o mesmo eleitorado, ou que buscaram se apropriar dos mesmos temas na agenda ao longo da campanha eleitoral.

Embora os valores de R^2 ajustado não ficassem acima de 80% para todos os candidatos, justificando o caráter focal e experimental dessa pesquisa, o mesmo torna-se relevante para futuros trabalhos com intuito de comprovar a existência de índices que, de fato, venham a constituir uma nova fase de estudos de previsão eleitoral, tão ou mais precisos que os modelos tradicionalmente utilizados.

Referências

BIFET A., FAN W.: Mining Big Data: Current Status, and Forecast to the Future. SIGKDD Explorations 14(2), 1-5, 2013.

BORA, N.N. (2014) Summarizing public opinions in tweet. In Journal Proceedings of CICLing, 2014.

DIGRAZIA J, MCKELVEY K, BOLLEN J, ROJAS F. More Tweets, More Votes: Social Media as a Quantitative Indicator of Political Behavior. PLoS ONE 8(11): e79449. https://doi.org/10.1371/journal.pone.0079449, 2013.

GAYO-AVELLO, D.(2014). “I wanted to predict elections with twitter and all I got was this lousy paper” – A balanced survey on elections prediction using twitter data. CoRR – http://arxiv.org/abs/1204.6441.

GAYO-AVELLO, D., METAXAS, P. AND MUSTAFARAJ, E. (2011). Limitis of election predictions using twitter. In Int. Conf. on Weblogs and Social Media (ICWSM), pages 490-493.

GHOSH, S., VISWANATH, B., KOOTI, F., SHARMA, N.K., KORLAM, G., BENEVENUTO, F., GANGULY, N., AND GUMMADI, K.P. (2012). Understanding and combating link farming in the twitter social network. In Int. Conf. on World Wide Web, WWW’ 12, pages 61-70.

JUNGHERR, A., JÜRGENS, P. AND SCHOEN, H. (2012). Why the pirate party won the german election of 2009 or the trouble with predictions: A response to Tumasjan, A., Sprenger, T.O., Sander, P.G., & Welpe, I.M.”Predicting elections with twitter: What 140 characters reveal about political sentiment”. Soc. Sci. Comput. Rev.,30(2):229-234.

KRISTENSEN JB, ALBRECHTSEN T, DAHL-NIELSEN E, JENSEN M, SKOVRIND M, BORNAKKE T. Parsimonious data: How a single Facebook like predicts voting behavior in multiparty systems. PLoS ONE 12(9): e0184562, 2017.

LUMEZANU, C., FEAMSTER, N., AND KLEIN, H. (2012). #bias: Measuring the tweeting behavior of propagandist. In Int. conf. on Weblogs and Social Media (ICWSM), pages 210-217.

MUKHERJEE, S., MALU, A., A.R., B., AND BHATTACHARYYA, P. (2013), Twisent: a multistage system for analizing sentiment in twitter. In Int. conf. on Information and knowledge management, CIKM' 12, pagens 2531-2534.

TRUMPER, D. S., MEIRA, W., & ALMEIDA, V. From Total Hits to Unique Visitors Model for Election's Forecasting. Proceedings of the ACM WebSci'11. Koblenz, 2011.

TRUMPER, D. S., MEIRA, W., & ALMEIDA, V. From Total Hits to Unique Visitors Model for Election's Forecasting. Proceedings of the ACM WebSci'11. Koblenz, 2011.

TUMASJAN A, SPRENGER T. O., SANDNER P. G., AND WELPE I. M.. Election Forecasts With Twitter: How 140 Characters Reflect the Political Landscape. Social Science Computer Review, 2010.

TUMASJAN, A., SPRENGER, T.O., SANDER, P.G., & WELPE, I.M.(2010). Predicting elections with twitter: What 140 characters reveal about political sentiment. Int. conf. on Weblogs and Social Media (ICWSM), pages 178-185.

YASSERI T, BRIGHT J. Wikipedia traffic data and electoral prediction: towards theoretically informed models. EPJ Data Sci, 2016.