Revista de Sistemas de Informação da FSMA
n. 24 (2019) pp. 31-39

SISTEMAS DE INFORMAÇÃO
http://www.fsma.edu.br/si/sistemas.html

TRILHA PRINCIPAL

Detecção automática de bots em redes sociais: um estudo de caso no segundo turno das eleições presidenciais brasileiras de 2018

Matheus de Oliveira Lêu, *Graduando em Sistemas de Informação, Escola de Artes, Ciências e Humanidade, USP*,
Daniel Marques Gomes de Morais, *Doutorando em Sistemas de Informação, Escola de Artes, Ciências e Humanidade, USP*,
Fernando Xavier, *Doutorando em Engenharia Elétrica, Escola Politécnica, USP* e
Luciano Antonio Digiampietri, *Professor Doutor, Escola de Artes, Ciências e Humanidade, USP*

***Resumo*—Bots sociais são usuários automatizados que utilizam redes sociais e aplicativos de troca de mensagens para interagir com usuários reais. Estes bots podem servir para o compartilhamento de notícias importantes, como informações sobre o clima ou situações de emergência. Porém, podem também ser utilizados com objetivos maliciosos, como a propagação de notícias falsas. Este trabalho apresenta um estudo de caso sobre a penetração de bots nas discussões eleitorais no Twitter durante o segundo turno das eleições presidenciais brasileiras de 2018. Identificou-se que a quantidade de bots participando das discussões eleitorais pode ser superior a 10% do total de usuários. O uso de características gerais das contas dos usuários se mostrou bastante específico na identificação dos humanos (estima-se que mais de 97% dos usuários humanos foram classificados corretamente como humanos), porém não foram identificados mais do que 52,3% dos bots.**

***Palavras-chave*—Identificação de bots, Redes sociais, Análise de Redes Sociais.**

Automatic bot detection in social networks: a case study in the second round of the 2018 Brazilian presidential elections

***Abstract*—Social bots are automated users who use social networks and messaging applications to interact with real users. These bots can be used for sharing important news such as weather information or in emergency situations. However, they can also be used for malicious purposes, such as spreading fake news. This paper presents a case study on bot penetration in electoral discussions during the second round of the 2018 Brazilian presidential elections. It was found that the number of bots participating in the electoral discussions may have exceeded 10% of the total users. The use of general user account characteristics proved to be very specific in identifying humans (it is estimated that over 97% of human users were correctly classified as humans), but bot recalls did not exceed 52.3%.**

Autor correspondente: Luciano Digiampietri, *digiampietri@usp.br*

***Index Terms*—Bot identification, Social Networks, Social Network Analysis**

I. Introdução

As redes e mídias sociais, utilizadas por centenas de milhões de usuários diariamente, são excelentes mecanismos de comunicação e divulgação, permitindo que o usuário tenha um poder de comunicação que antes estava restrito apenas a grandes meios de comunicação, podendo se expressar e, ao mesmo tempo, atingir um número muito grande de pessoas. Porém, dadas as características destas redes, há um incentivo real para a utilização de mecanismos para a manipulação dos fluxos de comunicação, de forma a reforçar (ou denegrir) imagens, ou manipular a opinião pública acerca de determinados assuntos.

Dentre os mecanismos utilizados para este fim, destaca-se a criação de algoritmos que têm por objetivo se passar por usuários destas redes. A estes algoritmos é dado o nome de *bots* sociais, ou apenas *bots*. A atuação destes *bots* tem sido observada em diversos eventos e situações, por exemplo a atuação em processos eleitorais, de forma a angariar (ou destruir) apoios para determinados candidatos ou correntes políticas [1]–[3], terrorismo, ou ainda propagação de notícias falsas com os mais diversos fins [4], [5].

Tendo em vista o potencial destrutivo destes *bots*, é necessária a criação de medidas de contenção, de forma a inibir seu poder de atuação e, consequentemente, garantir lisura e transparência das comunicações realizadas por meio das redes sociais. A possibilidade de influência por parte destes mecanismos em processos eleitorais é um dos usos mais destrutivos destes agentes, uma vez que podem influenciar rumos de nações inteiras, de forma a atender interesses de pequenos grupos, minando o que é esperado em um Estado democrático.

Com base em situações observadas em processos eleitorais em diversos países e a atuação de *bots* nestes

processos [1]–[3], este trabalho tem por objetivo analisar a penetração destes agentes no processo eleitoral brasileiro, em especial no segundo turno das eleições presidenciais de 2018.

Para tanto, o restante deste artigo está organizado da seguinte forma. A seção II descreve alguns trabalhos correlatos. Já a seção III apresenta os materiais e métodos utilizados. A seção IV contém a apresentação e descrição dos resultados. Por fim, a seção V contém as considerações finais e trabalhos futuros.

II. Trabalhos Correlatos

Entre as técnicas mais populares para a detecção de *bots* em redes sociais está a utilização de mecanismos de aprendizado supervisionado, dentre os quais se destacam a utilização de *Random Forest* [1], [6]–[10], *Adaboost* [11]–[14] e ainda *Gradient-Boosted Trees* [15]. Ainda, é possível observar na literatura a utilização de árvores de decisão [16], [17], abordagens estas aplicadas quando há o interesse em se obter modelos interpretáveis, em geral de forma a compreender como a atuação de um *bot* se deu em determinado cenário estudado.

Um os principais problemas observados acerca da utilização de aprendizado supervisionado é a especialização do classificador em um determinado conjunto de dados. Tal especialização reduz a eficiência da abordagem na construção de mecanismos de detecção de propósito geral que sejam resilientes, uma vez que os responsáveis pelos *bots* tendem a ajustar seus algoritmos para burlar detectores, reduzindo a relevância de características consideradas distintivas pelos detectores. Isto foi observado na análise da *botnet StarWars* [18], que expõe uma rede bem estabelecida destes agentes, em atuação por vários anos. Esta *botnet* recebe este nome pois seus *bots* costumam citar trechos de episódios da série de filmes Star Wars. Algumas das características de seus *bots* que os ajudam a não ser identificados como *bots* são: frequência baixa de postagens (menos de 12 mensagens por semana) e pequena quantidade de amigos e seguidores (no máximo, poucas dezenas).

Todavia, para trabalhos cujo objetivo é a análise de um conjunto fechado de dados, a abordagem supervisionada mostra-se efetiva, uma vez que o classificador será usado apenas no contexto dos dados analisados. Esta abordagem é observada, com ênfase na análise de processos eleitorais, nos trabalhos [1]–[3]. Eles utilizam *Random Forest*, SVM combinado com Redes Neurais e Agrupamento Hierárquico, respectivamente, com resultados adequados ao objetivo pretendido. A construção de árvores de decisão, quando há a necessidade de interpretatibilidade do modelo [16], [17], apesar da alta especialização resultante (que pode ser contornada com podas, mas com eventual perda de acurácia), também é uma abordagem que mostra-se adequada quando restrita ao conjunto de dados a ser analisado.

Cabe ressaltar ainda a popularidade do *Twitter* como principal plataforma para este tipo de análise. Todos os trabalhos aqui citados se baseiam em análise de dados desta rede social em particular. Dentre as principais razões para este fato, além da popularidade e volume de informação disponível, está a acessibilidade da informação, por meio da disponibilização de APIs públicas para a obtenção destes dados.

III. Materiais e Métodos

Os materiais utilizados neste trabalho consistem de *tweets* relacionados ao segundo turno das eleições presenciais brasileiras ocorridas em 2018.

Os dados foram coletados entre os dias 08 e 28 de outubro de 2018, período entre a confirmação oficial dos dois candidatos mais bem votados no primeiro turno e a realização do segundo turno da eleição presidencial. Para a identificação dos *tweets* a serem coletados, usou-se o sobrenome de cada candidato, bem como, em cada dia, as *hashtags* referentes a cada candidato que estavam melhor posicionadas nos *Trends Topics* Brasil do Twitter, que é uma lista dos assuntos mais comentados no momento.

Para a coleta dos dados, foi desenvolvido um *script* em Python com armazenamento dos dados no banco de dados NoSQL MongoDB. Os *tweets* foram acessados por meio de uma *Application Programming Interface* (API) disponibilizada pelo Twitter, cujo acesso é feito de maneira autenticada e mediante criação de aplicação na plataforma para uso da API.

O conjunto de dados minerado contém informações de 635.957 usuários. Ao todo, os usuários do conjunto de dados postaram 12.037.347.994 *tweets* ao longo de todo o tempo de vida da de suas contas no Twitter, uma média de 18.928 postagens por usuário. Ao se considerar apenas o 1% mais ativo destes usuários, observa-se um total de 1.717.691.584 postagens (mais de 14% do total de postagens).

Neste trabalho optou-se pela extração de um conjunto de características gerais de cada um dos usuários do conjunto de dados coletado. Estas características foram selecionadas devido a seu uso relativamente frequente e por apresentarem bons resultados em trabalhos correlatos [19], [20]. Um total de 22 atributos ou características foram extraídos: *acc_age_days* que corresponde a idade da conta, em dias; *default_pic* e *default_prof* são características binárias que informam se o usuário, respectivamente, utiliza a foto padrão do Twitter em seu perfil e se alterou ou não o perfil padrão inicial da conta. Os atributos *friends*, *likes*, *tweets* e *followers* correspondem ao número de amigos, curtidas, *tweets* postados e seguidores, respectivamente. Os atributos *verified* e *geo_enabled* são binários e indicam, respectivamente, se um usuário tem conta verificada e se sua localização está ativada. Já os atributos *tweets_tag* e *retweets_tag* correspondem aos números de *tweets* e *retweets* do usuário dentro das *tags* políticas a partir das quais os usuários foram extraídos, sendo *tag_tweet_freq* a proporção de *tweets* por *retweets*. Os demais atributos são derivados dos atributos já apresentados, a saber: *tweets_per_day* que corresponde ao número de *tweets* dividido por *acc_age_days* e *likes_followers_ratio*, que é definido pela divisão de *likes* por *followers*.

Para possibilitar o treinamento e validação de modelos, 642 usuários foram selecionados aleatoriamente e classificados manualmente como *humano* ou *bot*, resultando em 577 humanos (89,9%) e 65 *bots* (10,1%). Estes números podem indicar (posto que a nossa amostragem foi aleatória) uma penetração potencial de *bots* nas eleições em torno de 10% dos usuários, destacando-se que, frequentemente, estes *bots* são muito ativos.

As características extraídas foram utilizadas como entrada para algoritmos de aprendizado de máquina supervisionados considerados interpretáveis (ou explicáveis) de forma a indicar quais razões levaram os modelos a considerar um usuário do Twitter como um *bot* ou não. Foram selecionados dois algoritmos, um baseado em árvores de decisão, especificamente, Árvores Aleatórias (*Random Trees*) e outro baseado em regressão linear [21]. O único parâmetro que foi variado durante o treinamento das Árvores Aleatórias foi a altura das árvores, de forma a se gerar modelos de mais simples interpretação (com poucos nós) até mais complexos e, potencialmente, mais precisos. Para a produção do modelo baseado em regressão linear, a classe *bot* foi convertida para o valor numérico 1.000.000 e a classe *humano* para o valor numérico -1.000.000.

Os modelos foram avaliados usando validação cruzada com dez subconjuntos (*10-fold cross-validation*). As principais medidas utilizadas foram taxa total de acertos para o problema de classificação usando árvores de decisão e coeficiente de correlação para a regressão linear. Apesar do conjunto de dados ser desbalanceado, neste trabalho optou-se por não realizar balanceamento no conjunto de treinamento. Isto poderia ser utilizado para aumentar a identificação de prováveis *bots*, porém também acarretaria na diminuição da especificidade na identificação, por isto, este balanceamento não foi adotado.

Os modelos produzidos foram aplicados no conjunto total de dados de forma a se estimar a quantidade de *bots* durante a eleição. Com este propósito, a regressão linear foi utilizada para a atribuição de uma pontuação (*score*) e um limiar foi estabelecido, com base nos dados de treinamento, para a classificação como *bot* ou *humano*.

Adicionalmente, um seletor de atributos (ou características) foi utilizado para identificar os atributos mais importantes (ou informativos) para a classificação de usuários como humanos ou *bots*. O seletor utilizado foi o *ChiSquaredAttributeEval*, o qual avalia o valor de um atributo com base na estatística qui-quadrado em relação à classe. Neste projeto, as implementações dos classificadores, bem como do seletor de atributos utilizados foram as disponíveis no arcabouço Weka [21].

IV. Resultados e Discussão

ESTA seção apresenta e discute os resultados deste trabalho. Inicialmente foram produzidas cinco árvores de decisão de tamanhos diferentes (variando da menor árvore com altura um, até uma árvore com altura cinco). Por fim, são apresentados e discutidos os resultados do uso de regressão linear.

A precisão dos modelos é discutida em relação à quantidade de instâncias corretamente classificadas e os modelos apresentados são aqueles produzidos com base no uso de todo o conjunto de treinamento, composto por 642 instâncias.

Destaca-se que o conjunto de dados é desbalanceado (cerca de 89,9% das instâncias é formada por humanos). Assim, um possível *baseline* para a acurácia da classificação é a taxa de elementos na classe majoritária, isto é, a proporção de humanos.

A. Árvores de Decisão

A figura 1 apresenta a árvore de decisão produzida com altura igual a um. Nesta árvore há apenas um nó de decisão, indicando que se o usuário postou menos do que 48,49 *tweets* por dia (*tweets_per_day*) então ele é humano (e esta condição classifica corretamente 564 usuários como humanos e incorretamente 43 *bots* como humanos). Caso contrário, o usuário será considerado um *bot*. Observa-se que 35 usuários satisfazem essa condição e, destes, 13 são humanos e 22 são *bots*.

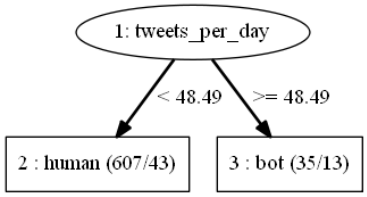


Fig. 1. Modelo de Árvore de Decisão com altura um. Os números entre parênteses representam, respectivamente, o número de usuários classificados nesta categoria e a quantidades destes que está falsamente classificada.

As figuras 2, 3 e 4 apresentam, respectivamente, as árvores de decisão produzidas de altura dois, três e quatro.

Na figura 2 é possível observar que as duas folhas do lado esquerdo contêm dados de usuários classificados como humanos. Isto quer dizer que independente do valor do atributo de decisão (no caso, *tweets_tag*) o usuário será classificado como humano. O que muda entre os elementos do nó à esquerda ($tweets\_tag < 3,5$) e os à direita ($tweets\_tag \geq 3,5$) é a taxa de acertos da regra em relação ao conjunto de treinamento: 96,7% para os usuários à esquerda e 81,6% para os usuários à direita.

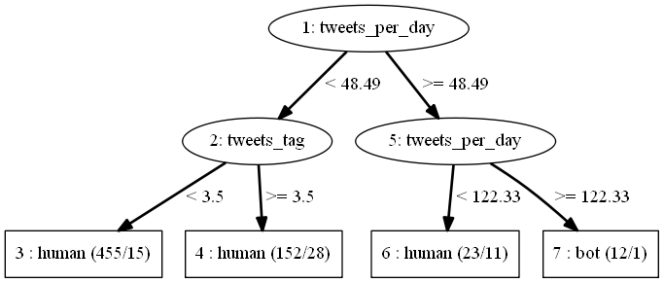


Fig. 2. Modelo de Árvore de Decisão com altura dois

Nas diferentes árvores é possível encontrar caminhos que levam a nós com taxas de acertos bastante altas. Por

exemplo, na árvore apresentada na figura 3, o caminho percorrido à direita da raiz (isto é, *tweets* por dia maior ou igual a 48,49) e então à esquerda do próximo nó de decisão (isto é, idade da conta inferior à 1117 dias) leva a um conjunto de instâncias composto por 14 usuários dos quais 13 são *bots* (92,9%). Já a folha desta mesma árvore rotulada com o número 5 contém 440 humanos e 14 *bots* (96,9% de humanos).

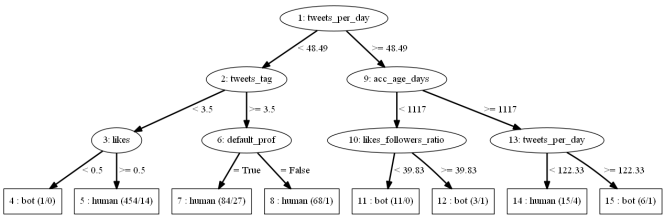


Fig. 3. Modelo de Árvore de Decisão com altura três

Apesar da existência de diversas folhas com regras com alta taxa de acerto, vale ressaltar que algumas folhas contêm poucos usuários (menos de 5% do conjunto de treinamento), o que gera regras que podem ter dificuldade de generalização para novas instâncias. Por exemplo, na figura 4 há sete folhas que classificaram, cada uma, menos de 10 usuários (folhas indicadas com os números 4, 14, 19, 20, 23, 26 e 27).

A figura 5 apresenta a árvore de decisão produzida com altura igual a cinco. Destaca-se que a primeira condição, presente na raiz da árvore, é a mesma das demais árvores (envolvendo a característica *tweets* por dia), mas, ao longo da árvore, observa-se o uso de cerca de metade das características consideras neste estudo. Há uma quantidade significativa de folhas com o número zero como o segundo número entre parênteses. Isto representa o número de elementos incorretamente classificados no nó (apesar de algumas destas folhas representarem um número pequeno de usuários).

A classificação de usuários como *bots* atingiu alta especificidade (a maioria dos humanos foi classificada corretamente como humanos), mas sensibilidade relativamente baixa (isto é, muitos *bots* foram classificados como humanos). A tabela I apresenta a especificidade (ou taxa de verdadeiros negativos) e a sensibilidade (ou taxa de verdadeiros positivos) das classificações considerando as árvores de decisão com diferentes alturas.

TABELA I
Resultados de Especificidade e Sensibilidade

Altura da Árvore de Decisão	Especificidade	Sensibilidade	Classificações Correta
Um	97,7%	33,8%	91,3%
Dois	99,8%	16,9%	91,4%
Três	99,7%	29,2%	92,5%
Quatro	99,8%	46,2%	94,4%
Cinco	100,0%	52,3%	95,2%

Os resultados da tabela I indicam que estes modelos podem ser utilizados com sucesso para separar os humanos, no sentido de que grande quantidade dos humanos será de fato classificada como *humano* (variando de 97,7% a 100%, para o conjunto de treinamento utilizado), porém não conseguem identificar uma grande quantidade de *bots* (entre 16,9% e 52,3%). Destaca-se que os modelos mais complexos (com maior altura) obtiveram os melhores resultados. A única exceção observada é que o modelo de altura um é significativamente mais sensível do que o de altura dois.

Detalhes sobre os quatro resultados de classificação (classificações corretas e incorretas para humanos e *bots*) para as cinco árvores podem ser observados na tabela II que apresentam as matrizes confusão.

TABELA II
Matrizes Confusão - Árvores de Decisão de Altura de Um a Cinco

Altura	Classificado como Humano	Classificado como Bot	
Um	564	13	Humano
	43	22	Bot
Dois	576	1	Humano
	54	11	Bot
Três	575	2	Humano
	46	19	Bot
Quatro	576	1	Humano
	35	30	Bot
Cinco	577	0	Humano
	31	34	Bot

Os cinco modelos de árvore de decisão produzidos com base no conjunto de treinamento foram aplicados ao conjunto total de dados, de forma a se calcular a porcentagem de *bots* prevista por cada um dos modelos. A tabela III apresenta estes resultados.

Observa-se que uma quantidade relativamente pequena de *bots* foi prevista, entre 0,25% e 2,84% dos usuários, valores significativamente inferiores à taxa observada no conjunto de treinamento (10,1%). Este fato ocorre principalmente por dois motivos relacionados. O primeiro é o desbalanceamento do conjunto de dados, que leva a produção de modelos que, ao maximizarem a taxa total de acertos, acabam por privilegiar a classificação na classe majoritária. O segundo é que, conforme já discutido, os modelos são bastante específicos apesar de pouco sensíveis. Neste contexto, isto significa que os modelos classificam menos usuários como *bots* e conseguem identificar os humanos de forma bastante precisa (conforme visto na especificidade de 97,7% a 100% para o conjunto de treinamento).

Desta forma, considera-se que os modelos de árvores de decisão produzidos são bastante eficientes (em termos de especificidade) em detectar *bots* reais, porém muitos *bots* não são detectados. Destaca-se que, conforme pode ser observado nas árvores de decisão, os usuários classificados

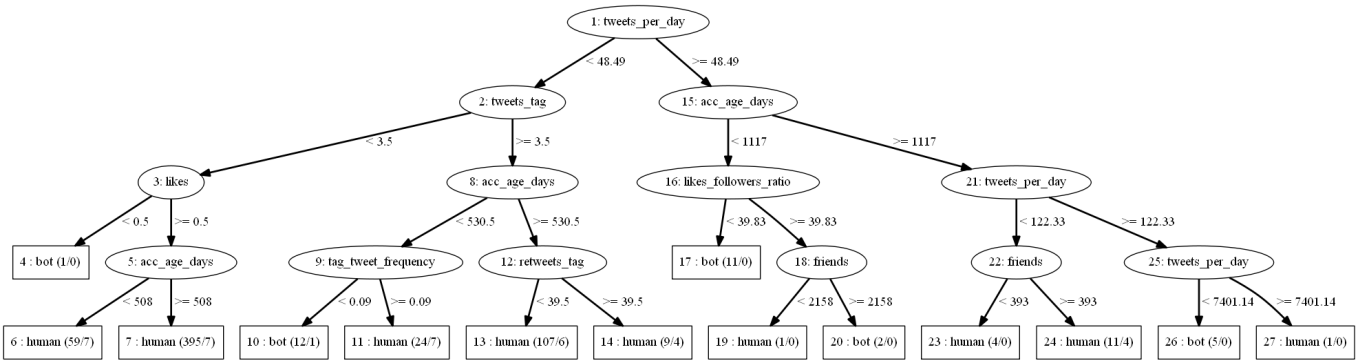


Fig. 4. Modelo de Árvore de Decisão com altura quatro

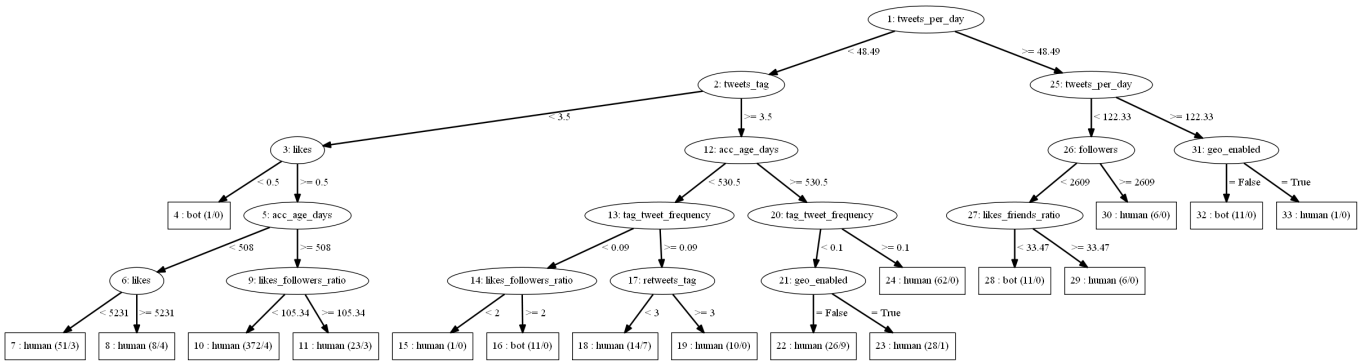


Fig. 5. Modelo de Árvore de Decisão com altura cinco

TABELA III
Resultado da Classificação - Conjunto Total de Dados

Classificações	Humanos	Bots	Porcentagem de Bots
Altura Um	617.921	18.036	2,84%
Altura Dois	634.346	1.611	0,25%
Altura Três	626.988	8.969	1,41%
Altura Quatro	622.554	13.403	2,11%
Altura Cinco	622.053	13.904	2,19%

como *bots* são justamente alguns dos usuários que mais publicaram mensagens podendo, assim, serem grandes propagadores de, por exemplo, notícias falsas.

O uso de Árvores de Decisão permite três análises interessantes para problemas como os tratados neste artigo. Em primeiro lugar, os resultados da classificação considerando diferentes aspectos (como taxa de acerto, especificidade e sensibilidade), que são diretamente relacionados ao objetivo da classificação automática. Porém, um segundo resultado interessante é a observação individual dos nós folha das árvores, os quais podem indicar classificações mais ou menos precisas de acordo com as regras/decisões que levaram a cada um dos nós. Por fim, a árvore permite também realizar uma análise de quais são os atributos considerados mais relevantes pelo modelo utilizado.

A análise da importância dos atributos foi complementada com o uso do seletor de atributos *ChiSquaredAttributeEval*, o qual ordena os atributos mais relevantes/informativos de acordo com a classe do problema. Os sete atributos melhor ranqueados, em ordem da maior para a menor importância, foram: *tweets_per_day*, *tweets_tag*, *tweets*, *retweets_tag*, *acc_age_days*, *default_prof* e *geo_enabled*. Destaca-se nas três primeiras posições: *tweets_per_day*, isto é, a quantidade de tweets por dia; *tweets_tag* que corresponde à quantidade de *tweets* publicados pelo usuário com as *tags* políticas utilizadas neste estudo; e *tweets* que representa o total de *tweets* publicados pelos usuários.

B. Regressão Linear

Uma regressão linear foi utilizada para produzir uma pontuação (*score*) para cada um dos usuários. A hipótese considerada é que esse sistema de pontuação pode ser utilizado para identificar usuários com maior ou menor chance de serem *bots*. A correlação entre a classe (convertida para um valor numérico) e o resultado da regressão linear foi de 0,4964 para o conjunto de treinamento, indicando uma correlação moderada.

A equação 1 apresenta a função de pontuação produzida pela regressão linear. Observa-se que seis atributos foram considerados pela regressão, dois binários e quatro numéricos. Os dois maiores pesos da regressão foram dados os atributos binários: $default\_prof = True$, significando que se o usuário utiliza o perfil padrão (isto é, não fez modificações em seu perfil) maior será sua pontuação e $geo\_enabled = False$, significando que se o usuário não habilitou a geolocalização então sua pontuação será

maior. Os dois atributos numéricos com pesos positivos são *retweets_tag*, ou seja, quanto maior a quantidade de *retweets* com as *tags* políticas utilizadas, maior será a pontuação do usuário e *tweets*, isto é, quanto maior o número de *tweets* postados maior será a pontuação. Por fim, os dois atributos com pesos negativos são *likes*, isto é, quanto maior a quantidade de *likes* menor será a pontuação e *acc_age_days*, significando que quanto mais velha for a conta do usuário menor será a pontuação.

$$\begin{aligned}\text{Pontuação} = \quad & 77458,6767 \times default\_prof = True + \\ & 63044,9069 \times geo\_enabled = False + \\ & 5129,7994 \times retweets\_tag + \\ & 1,3396 \times tweets + \\ & -0,5348 \times likes + \\ & -35,1782 \times acc\_age\_days + \\ & 41029,2316\end{aligned} \tag{1}$$

A tabela IV apresenta a quantidade de *bots* bem como as respectivas pontuações para dez intervalos. Os intervalos foram definidos por valores de pontuação de acordo com a quantidade de *bots*, dos de maior pontuação para os de menor pontuação, contendo sempre 6 ou 7 *bots* em cada intervalo. Estes valores foram escolhidos para manter uma distribuição equilibrada de *bots* em cada intervalo.

Conforme a premissa, nos intervalos de maior valor de pontuação encontram-se taxas mais elevada de *bots* (100% para o primeiro intervalo, 77,78% para o segundo e assim por diante), lembrando-se que, no conjunto de treinamento, apenas 10,1% dos usuários são *bots*.

TABELA IV
PORCENTAGEM DE BOTS ENTRE OS USUÁRIOS CONSIDERANDO DIFERENTES INTERVALOS DE PONTUAÇÃO

Número de bots no intervalo	Revocação dos bots	Usuários no intervalo	Menor pontuação no intervalo	Porcentagem de bots no intervalo
6	9,2%	6	861.000	100,00%
7	20,0%	9	549.000	77,78%
6	29,2%	12	327.000	50,00%
7	40,0%	13	266.000	53,85%
6	49,2%	18	198.000	33,33%
7	60,0%	27	178.000	25,93%
6	69,2%	46	170.000	13,04%
7	80,0%	96	136.000	7,29%
6	89,2%	50	104.000	12,00%
7	100,0%	286	- 28.000	2,45%

A pontuação obtida pode ser utilizada para a classificação, estabelecendo-se um limiar a partir do qual todos os usuários são classificados como *bots*. O limiar ótimo, considerando a maximização dos usuários corretamente classificados para o conjunto de treinamento, foi de 266.000 pontos e obteve uma taxa de acerto de 91,74%.

A pontuação também pode ser utilizada para um controle mais refinado da especificidade e da sensibilidade do classificador, dependendo dos objetivos de quem está realizando a classificação.

A figura 6 apresenta a curva de precisão e revocação em relação ao limiar, já a figura 7 contém a curva ROC. Estas duas figuras destacam os desafios da identificação de *bots*: É possível observar que a classificação se inicia bastante precisa, porém rapidamente a precisão cai ao passo que a revocação sobe (característica comum em problemas desbalanceados). A área sobre a curva ROC apresentada é de 0,819, correspondendo a um resultado satisfatório para o problema, mas com margem para melhorias.

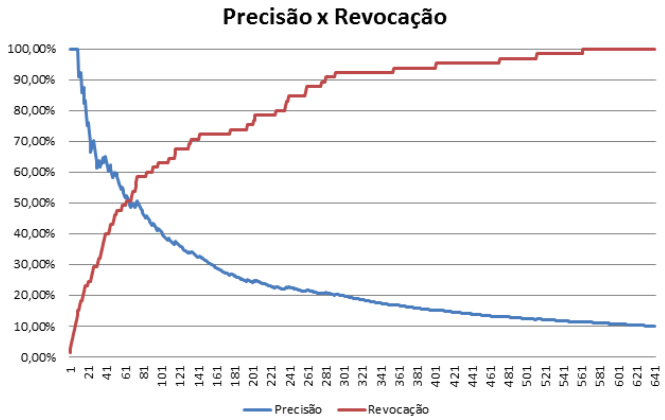


Fig. 6. Precisão versus Revocação da classificação utilizando regressão linear

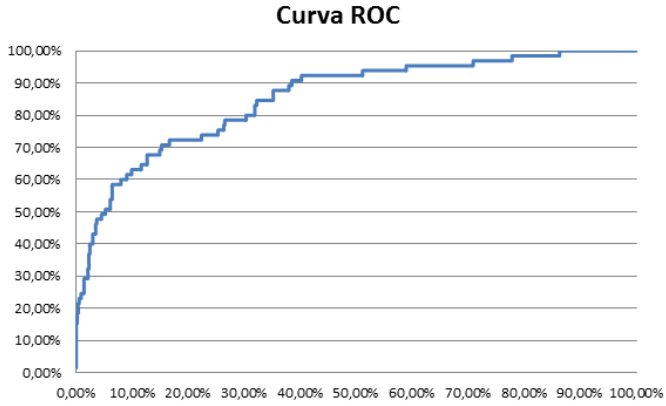


Fig. 7. Curva ROC da classificação utilizando regressão linear

A pontuação também pode ser utilizada numa primeira etapa da classificação, separando, por exemplo, um conjunto no qual se estime que 50% dos usuários sejam *bots* e, em seguida, utilizar outro método de classificação para estes usuários (por exemplo, classificação manual caso a identificação de *bots* seja um problema crítico). Essa abordagem foi utilizada no presente trabalho, combinada com o uso de uma árvore de decisão. O limiar de pontuação 181.000 foi utilizado, pois, no conjunto de treinamento, o subconjunto formado por usuários com esta pontuação ou uma pontuação superior ficou igualmente dividido entre *bots* e humanos. Adicionalmente, este conjunto contém 43,08% do total de *bots* do conjunto de treinamento.

Os usuários do conjunto de treinamento que atingiram uma pontuação maior ou igual a 181.000 foram utilizados para o treinamento de uma nova árvore de decisão de altura cinco (esta altura foi selecionada por ser a mesma altura da árvore que apresentou os melhores resultados apresentados anteriormente). Essa árvore foi capaz de classificar corretamente 63,29% destes usuários. Vale ressaltar que se trata de um conjunto balanceado formado por *bots* e humanos com as maiores pontuações.

Considerando o conjunto total de usuários, 34.986 (5,5% do total) atingiram pontuações maiores ou iguais a 181.000. Destes, 15.338 (2,4% do total de usuários) foram considerados *bots* de acordo com as regras da árvore de decisão produzida. Estes resultados são compatíveis com os resultados do uso direto de árvores de decisão, apresentados anteriormente.

V. Conclusões e Trabalhos Futuros

COM o grande aumento da interação entre pessoas ocorrendo por meio de redes sociais ou aplicativos de troca de mensagens, se torna cada vez mais importante a identificação se as informações compartilhadas são verdadeiras, bem como se as mensagens de fato estão sendo compartilhadas por pessoas.

Este trabalho apresentou um estudo de caso sobre a identificação automática de *bots* com base em postagens da rede social Twitter, analisando a penetração de *bots* nas discussões sobre o segundo turno das eleições presenciais brasileiras de 2018.

Nos dados anotados manualmente, mais de 10% dos usuários foram classificados como *bots*. Apesar de ser uma extrapolação possível, imaginar que 35.957 usuários do conjunto de dados analisado pode ser de *bots*, os resultados apresentados neste trabalho sugerem um percentual inferior de *bots* identificados automaticamente (até 2,84%).

A classificação automática dos usuários utilizou algoritmos de classificação consideradas explicáveis (ou interpretáveis). As árvores de decisão apresentaram resultados bastante específicos (menos de 3% dos humanos foram classificados como *bots*), porém apenas um dos modelos produzidos foi capaz de, no conjunto de treinamento, atingir uma revocação de *bots* acima de 50%.

Aplicando o modelo mais preciso (árvore de decisão de altura cinco) no conjunto total de dados, 13.904 usuários (2,19% do total de usuários) foram classificados como *bots*. Apesar de ser um número significativamente menor do que o estimado pela anotação manual no conjunto de treinamento, destaca-se que são usuários bastante ativos (conforme pode ser observado nos nós de decisão dos modelos produzidos).

O desbalanceamento do conjunto de dados é uma característica que deve ser levada em consideração ao se analisar esses dados. Modelos que objetivam a maximização da classificação correta total dos elementos costumam a tender para a atribuição de indivíduos na classe majoritária.

Observou-se tanto com o uso de árvores de decisão quanto de regressão linear que há nós de decisão ou intervalos de valor bastante precisos na correta identificação de *bots*, restando como desafio futuro classificar os *bots* que não se encontram nesses nós ou intervalos.

Como trabalhos futuros, objetivamos analisar o conteúdo textual das mensagens postadas e utilizar este conteúdo na classificação. Pretendemos também analisar as postagens do primeiro turno das eleições.

Referências

[1] J. Fernquist, L. Kaati e R. Schroeder, "Political Bots and the Swedish General Election", em *2018 IEEE International Conference on Intelligence and Security Informatics (ISI)*, nov. de 2018, pp. 124–129. DOI: 10.1109/ISI.2018.8587347.

[2] S. Khaled, N. El-Tazi e H. M. O. Mokhtar, "Detecting Fake Accounts on Social Media", em *2018 IEEE International Conference on Big Data (Big Data)*, dez. de 2018, pp. 3672–3681.

[3] S. Sadiq, Y. Yan, A. Taylor, M. Shyu, S. Chen e D. Feaster, "AAFA: Associative Affinity Factor Analysis for Bot Detection and Stance Classification in Twitter", em *2017 IEEE International Conference on Information Reuse and Integration (IRI)*, ago. de 2017, pp. 356–365. DOI: 10.1109/IRI.2017.25.

[4] V. S. Subrahmanian, A. Azaria, S. Durst, V. Kagan, A. Galstyan, K. Lerman, L. Zhu, E. Ferrara, A. Flammini e F. Menczer, "The DARPA Twitter Bot Challenge", *Computer*, v. 49, n. 6, pp. 38–46, jun. de 2016. DOI: 10.1109/MC.2016.183.

[5] E. Shaabani, R. Guo e P. Shakarian, "Detecting Pathogenic Social Media Accounts without Content or Network Structure", em *2018 1st International Conference on Data Intelligence and Security (ICDIS)*, abr. de 2018, pp. 57–64. DOI: 10.1109/ICDIS.2018.00016.

[6] C. A. S. d. Freitas, F. Benevenuto e A. Veloso, "Socialbots: Implications on the Safety and Reliability of Twitter-Based Services", em *2014 Brazilian Symposium on Computer Networks and Distributed Systems*, mai. de 2014, pp. 302–309. DOI: 10.1109/SBRC.2014.36.

[7] M. Singh, D. Bansal e S. Sofat, "A Novel Technique to Characterize Social Network Users: Comparative Study", em *Proceedings of the 6th International Conference on Communication and Network Security*, sér. ICCNS '16, Singapore, Singapore: Association for Computing Machinery, 2016, pp. 75–79, ISBN: 9781450347839. DOI: 10.1145/3017971.3017977. endereço: https://doi.org/10.1145/3017971.3017977.

[8] Z. Gilani, E. Kochmar e J. Crowcroft, "Classification of Twitter Accounts into Automated Agents and Human Users", em *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*, sér. ASONAM '17, Sydney, Australia: Association for Computing Machinery, 2017, pp. 489–496, ISBN: 9781450349932. DOI: 10.1145/3110025.3110091. endereço: https://doi.org/10.1145/3110025.3110091.

[9] B. Boreggah, A. Alrazooq, M. Al-Razgan e H. AlShabib, "Analysis of Arabic Bot Behaviors", em *2018 21st Saudi Computer Society National Computer Conference (NCC)*, abr. de 2018, pp. 1–6. DOI: 10.1109/NCG.2018.8592980.

[10] E. Van Der Walt e J. Eloff, "Using Machine Learning to Detect Fake Identities: Bots vs Humans", *IEEE Access*, v. 6, pp. 6540–6549, 2018. DOI: 10.1109/ACCESS.2018.2796018.

[11] J. P. Dickerson, V. Kagan e V. S. Subrahmanian, "Using sentiment to detect bots on Twitter: Are humans more opinionated than bots?", em *2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2014)*, ago. de 2014, pp. 620–627. DOI: 10.1109/ASONAM.2014.6921650.

[12] V. Natarajan, S. Sheen e R. Anitha, "Multilevel Analysis to Detect Covert Social Botnet in Multimedia Social Networks", *The Computer Journal*, v. 58, n. 4, pp. 679–687, abr. de 2015. DOI: 10.1093/comjnl/bxu063.

[13] F. Morstatter, L. Wu, T. H. Nazer, K. M. Carley e H. Liu, "A new approach to bot detection: Striking the balance between precision and recall", em *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, ago. de 2016, pp. 533–540. DOI: 10.1109/ASONAM.2016.7752287.

[14] P. Andriotis e A. Takasu, "Emotional Bots: Content-based Spammer Detection on Social Media", em *2018 IEEE International Workshop on Information Forensics and Security (WIFS)*, dez. de 2018, pp. 1–8. DOI: 10.1109/WIFS.2018.8630760.

[15] M. Kantepe e M. C. Ganiz, "Preprocessing framework for Twitter bot detection", em *2017 International Conference on Computer Science and Engineering (UBMK)*, out. de 2017, pp. 630–634. DOI: 10.1109/UBMK.2017.8093483.

[16] O. Loyola-González, R. Monroy, J. Rodríguez, A. López-Cuevas e J. I. Mata-Sánchez, "Contrast Pattern-Based Classification for Bot Detection on Twitter", *IEEE Access*, v. 7, pp. 45 800–45 817, 2019. DOI: 10.1109/ACCESS.2019.2904220.

[17] E. Ferreira Dos Santos, D. Carvalho, L. Ruback e J. Oliveira, "Uncovering Social Media Bots: A Transparency-Focused Approach", em *Companion Proceedings of The 2019 World Wide Web Conference*, sér. WWW '19, San Francisco, USA: Association for Computing Machinery, 2019, pp. 545–552, ISBN: 9781450366755. DOI: 10.1145/3308560.3317599. endereço: https://doi.org/10.1145/3308560.3317599.

[18] J. Echeverria e S. Zhou, "Discovery, Retrieval, and Analysis of the "Star Wars" Botnet in Twitter", em *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*, sér. ASONAM '17, Sydney, Australia: Association for Computing Machinery, 2017, pp. 1–8, ISBN: 9781450349932. DOI: 10.1145/3110025.3110074. endereço: https://doi.org/10.1145/3110025.3110074.

[19] A. Bessi e E. Ferrara, "Social bots distort the 2016 U.S. Presidential election online discussion", *First Monday*, v. 21, n. 11, 2016, ISSN: 13960466.

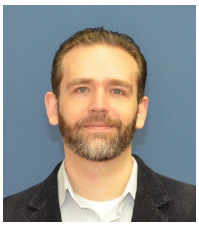
[20] J. Fernquist, L. Kaati e R. Schroeder, "Political Bots and the Swedish General Election", em *2018 IEEE International Conference on Intelligence and Security Informatics (ISI)*, nov. de 2018, pp. 124–129.

[21] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann e I. H. Witten, "The WEKA Data Mining Software: An Update", *SIGKDD Explor. Newsl.*, v. 11, n. 1, pp. 10–18, nov. de 2009, ISSN: 1931-0145.

Matheus de Oliveira Lêu é graduando no Bacharelado em Sistemas de Informação da Universidade de São Paulo. Tem experiência na área de Análise de Redes Sociais e, em particular, na detecção de bots sociais.

Daniel Marques Gomes de Morais possui graduação em Sistemas de Informação pela Universidade de São Paulo (2009) e mestrado em Programa de Pós-graduação em Sistemas de Informação pela Universidade de São Paulo (2014). Atualmente é doutorando no Programa de Pós-Graduação em sistemas de Informação pela Universidade de São Paulo e professor no Instituto Federal de São Paulo, Campus São Paulo. Tem experiência na área de Ciência da Computação, com ênfase em Ciência da Computação

Fernando Xavier possui graduação em Ciência da Computação pela Universidade Estadual de Campinas (2004), especialização em Gestão e Implementação de EaD pela Universidade Federal Fluminense (2015) e mestrado em Informática pela Universidade Federal do Rio de Janeiro (2016). Atualmente, é doutorando em Engenharia de Computação na Escola Politécnica da Universidade de São Paulo (Poli-USP). Tem experiência na área de Ciência da Computação, atuando como cientista de dados nos seguintes temas: saúde planetária, biodiversidade e recursos hídricos.



Luciano Antonio Digiampietri (autor correspondente) possui graduação em Ciência da Computação pela Universidade Estadual de Campinas (2002), doutorado em Ciência da Computação pela Universidade Estadual de Campinas (2007) e o título de Livre-docente em Informação e Tecnologia pela USP (2015). Desde abril de 2008 é professor pesquisador no Bacharelado em Sistemas de Informação na Escola de Artes, Ciências e Humanidades da Universidade de São Paulo (EACH-USP) e desde 2010 é professor permanente no Mestrado em Sistemas de Informação da EACH-USP. Tem experiência na área de Ciência da Computação, com ênfase em Biologia Computacional, Bancos de Dados e Inteligência Artificial, atuando principalmente nos seguintes temas: workflows científicos, bioinformática, composição automática de serviços, processamento de imagens e análise de redes sociais. E-mail: *digiampietri@usp.br*.