Um Modelo Baseado em Regras para a Detecção de *Bots* no Twitter

Maria Alice G. L. Leite, Marcus Vinícius C. Guelpeli, Caroline Queiroz Santos

[1]Programa de Pós-Graduação em Educação (PPGEd)
Universidade Federal dos Vales do Jequitinhonha e Mucuri (UFVJM)
Diamantina, MG - Brasil

maria.leite@ifnmg.edu.br, {marcus.guelpeli, caroline.queiroz}@ufvjm.edu.br

Abstract. *The increasing use of online social networks has made them important studies' sources in several fields, from the stock market and forecasting elections to human behavior. However, nowadays, bots' accounts activity in these networks has affected data shared and information dissemination, and these data have become vulnerable. This work proposes a supervised approach to extract knowledge from a specific literature database, using techniques that aim not only to classify but also to describe the main characteristics of bots and genuine Twitter accounts. To this, a rule-based classification model was generated. This model could contribute to building a framework to collect Twitter data with little interference from malicious accounts. The results were considered satisfactory when compared to other related works.*

Resumo. *O crescimento do uso das redes sociais online pela sociedade as tornou importantes fontes de estudos em vários campos, desde o mercado de ações e previsão de eleições até o comportamento humano. No entanto, amostras de dados extraídas dessas redes tornaram-se vulneráveis à atividade de contas bots. Por isso, este trabalho propõe uma abordagem supervisionada para extração de conhecimento a partir de uma base de dados da literatura, utilizando técnicas que visam não apenas classificar, mas também descrever as principais características dos bots e das contas genuínas no Twitter. O modelo de classificação baseado em regras foi gerado com o objetivo de contribuir para a construção de um framework para coletar dados do Twitter com pouca interferência de contas maliciosas. Os resultados foram considerados satisfatórios, se comparados a outros trabalhos relacionados.*

1. Introdução

As redes sociais *online* são estruturas que inter-relacionam empresas e pessoas, permitindo que seus usuários recebam e compartilhem informações sobre diversos assuntos [Naaman et al. 2010]. A sua capacidade de difundir de maneira imediata qualquer informação, possibilitando, por exemplo, detectar e subsidiar a construção de modelos de previsão de comportamento, torna possível correlacionar conteúdos virtuais com acontecimentos da vida real [Arias et al. 2014].

No entanto, essas características das redes sociais, somadas à credibilidade que os usuários possuem em relação aos seus amigos, impulsionam também o crescimento

de usuários mal-intencionados. Por isso, muitas coletas de dados extraídos dessas redes tornaram-se vulneráveis a ataques de robôs (*bots*), que são contas automatizadas criadas, muitas vezes, para a disseminação de desinformação, como rumores, *spam* e notícias falsas. Embora os *bots* existam em todas as redes sociais, é no Twitter que eles ganham destaque, devido ao número de usuários, de *tweets* (cerca de 500 milhões por dia [Saeed et al. 2019]) e à agilidade no compartilhamento de informações.

Não se sabe, exatamente, quantas contas do Twitter são *bots*, mas estima-se que 50% das contas sejam, de alguma maneira, automatizadas [Chu et al. 2012]. Contudo, este tipo de conta representa apenas 9% a 15% dos seus 330 milhões de usuários ativos [Varol et al. 2017]. Os trabalhos relacionados a detecção de *bots* no Twitter, em sua maioria, pertencem a dois grupos: *i)* os que se concentram na construção de modelos baseados em aprendizagem de máquina [Taigman et al. 2014, Ferrara 2017]; e *ii)* os que, de maneira qualitativa, identificam e caracterizam o comportamento dos *bots* [Bessi et al. 2014]. O primeiro grupo apoia-se, principalmente, na construção de modelos de alta performance para a produção de escores, negligenciando a interpretabilidade dos resultados, o que pode ser prejudicial na detecção de *bots* [Yang et al. 2019].

A questão que norteia este trabalho pode ser descrita como: em que medida é possível reconhecer e identificar os padrões de comportamento dos *bots* no Twitter, para otimizar o trabalho de analistas de dados em mídias sociais? Para tentar responder à questão, propomos um modelo baseado em regras para a detecção de *bots* no Twitter. O modelo será capaz de gerar regras que identifiquem e expliquem o comportamento das contas *bots*. Este trabalho é parte de outro, maior, que propõe a construção de um *framework* de coleta, tratamento e visualização de dados, reunindo um conjunto de funcionalidades que permitirão aos usuários coletarem dados no Twitter e extraírem informação e conhecimento desses dados. Espera-se que os resultados trazidos aqui possam contribuir com a construção da camada de filtro de *bots* do *framework* Oráculo [1].

O artigo está organizado da seguinte forma: na seção 2 é feita uma revisão da literatura, com os principais conceitos e contexto do trabalho, seguida da seção de Metodologia. Na seção 4 são discutidos os resultados encontrados e, por fim, as considerações finais são apresentadas na seção 5.

2. Revisão de Literatura

Muito se tem discutido sobre as abordagens de detecção da nova geração dos *bots* no Twitter, chamada de Social *Bots*. Alguns estudos apontam para a ineficiência da própria plataforma em detectá-los e ainda a dificuldade dos humanos em distinguí-los [Varol et al. 2017, Lee et al. 2011]. Uma das peculiaridades dos social *bots* é a capacidade de evolução contínua, adotando sofisticadas técnicas de burlar as abordagens automatizadas de detecção existentes, como as baseadas no conteúdo textual das mensagens compartilhadas, nos padrões de postagens e nas relações sociais [Cresci et al. 2017].

Uma ferramenta que protagonizou a aplicação científica de detecção de *bots* foi a *Botometer* [Varol et al. 2017], que, como a maioria das abordagens existentes, analisa os perfis de maneira individual, conta a conta. A *Botometer* utiliza a técnica *Random Forest* [Xia et al. 2016] para a classificação, tratando mais de mil atributos das contas. Em

[1] O projeto do *framework* Oráculo está em desenvolvimento no grupo de pesquisa MTPLNAM (`http://mtplnam.com.br/site/`)

um teste da *Botometer*, com a mesma base de dados utilizada neste trabalho, os resultados obtidos foram considerados insatisfatórios, pois o algoritmo demonstrou uma tendência em classificar os *bots* mais evoluídos como contas genuínas [Cresci et al. 2017].

Esses erros de generalização são comuns entre os modelos supervisionados de detecção de *bots*, devido a ausência de bases rotuladas de boa qualidade. Uma ferramenta de aprendizagem de máquina supervisionada é tão boa quanto os dados utilizados para treiná-la [Yang et al. 2019]. Outro ponto que pode acometer os sistemas de Inteligência Artificial (IA) que utilizam técnicas de aprendizado profundo (*deep-learning*) é o alto viés dos modelos produzidos. A utilização de metodologias de aprendizado mais sofisticadas tendem a produzir modelos complexos e enviesados [Kirkpatrick 2016].

Diferente dos trabalhos anteriores, [Miller et al. 2014] criticam a abordagem supervisionada quando aplicada na detecção de *spam bots*, e propuseram uma técnica baseada em detecção de anomalias. O autores modificaram dois algoritmos de *streaming clustering* encontrando resultados satisfatórios. Novas implementações desses algoritmos também foram testadas por [Cresci et al. 2017], obtendo os piores resultados entre as técnicas que utilizaram a base de dados desse projeto como teste. Para as abordagens que analisam as contas individualmente, a identificação de *bots* que agem em uma rede coordenada (*botnets*) é complexa, pois utiliza a comparação da série temporal de amostras da API do Twitter [Chavoshi et al. 2016] .

Assim, é possível verificar a existência de várias abordagens de detecção de *bots* nas redes sociais, principalmente baseados em IA. Todavia, pouco tem se investido na explicação dos modelos. Esse, ainda, é um grande desafio dos métodos de *deep-learning* quando aplicados a domínios de tomada de decisão. Mas, a interpretabilidade dos modelos pode facilitar na correção do viés, melhorando capacidade de generalização em novas amostras [Goodman and Flaxman 2016]. Outro grande benefício dos modelos explicáveis, inerentes ao domínio desse projeto, é a possibilidade de atestar quais variáveis, de fato, inferem os resultados, contribuindo para a evolução das discussões acerca do comportamento dos *bots*.

3. Metodologia e resultados

As etapas desta pesquisa (Figura 1) foram baseadas no processo de extração de conhecimento denominado *Cross-Industry Standard Process for Data Mining* - CRISP-DM [Chapman et al. 2000]. Na primeira fase foi feita a seleção da amostra da base de dados e a análise descritiva (pré-processamento). Na fase seguinte (processamento), os dados foram preparados, construindo-se o conjunto final de dados para a extração de conhecimento. Nesta fase ocorreram as seleções de variáveis, criação de novas variáveis, transformações e limpeza dos dados. Por último, na fase de pós-processamento, foram aplicadas as técnicas de extração de conhecimento para a produção das regras.

3.1. Fase 1: Pré-processamento

Na fase 1, foram definidos os objetivos e os requisitos do trabalho e desenvolvido um plano das ações a serem tomadas [Silva Filho and Adeodato 2019]. A coleta inicial dos dados foi realizada, seguida da seleção amostral e da compreensão dos dados. Para o desenvolvimento supervisionado, duas alternativas eram possíveis: a coleta e rotulação de uma base exclusiva para este estudo (*ad-hoc*), ou a utilização de uma base já rotulada.

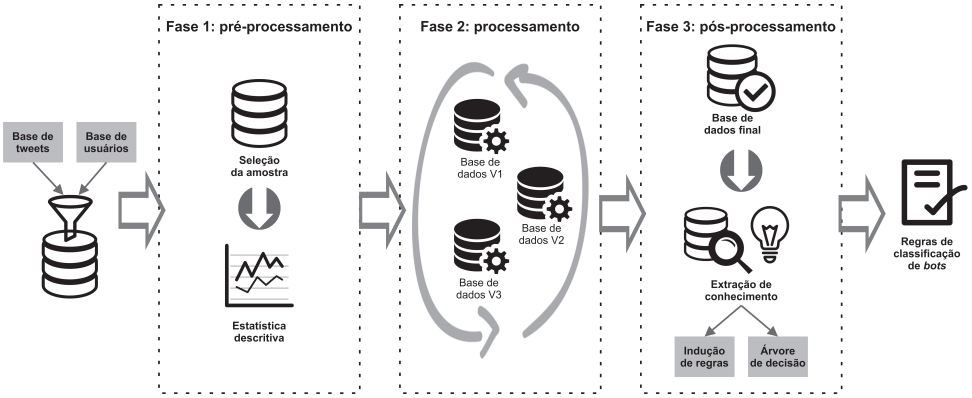

Figura 1. Etapas da pesquisa.

Para o primeiro cenário, as contas seriam selecionadas aleatoriamente e classificadas por meio de um detector de *bots* pré-existente, ou por meio de uma avaliação humana.

A classificação automatizada foi descartada devido às limitações do número de requisições da API do principal detector de *bots* disponível (*Botometer*) e, principalmente, pela condição de enviesamento que estava sendo estabelecida, uma vez que seriam consideradas as saídas da ferramenta como entradas para o modelo deste trabalho. Além disso, baseado em um projeto de *crowdsourcing* em que foi identificada uma dificuldade de humanos em identificar os *bots*, apenas baseado em seus *tweets* [Cresci et al. 2017], a anotação manual foi considerada lenta e ineficiente.

Com isso, realizou-se uma busca nos principais conjuntos de dados já rotulados e disponíveis na literatura, com o objetivo de encontrar uma base heterogênea, no que diz respeito aos diferentes tipos de *bots*, com uma boa descrição do método de anotação e, sobretudo, com volume de dados suficiente para extração do conhecimento. Entre as bases encontradas, destacou-se uma base que reutiliza outra base [Lee et al. 2011] e adiciona novos dados anotados manualmente [Varol et al. 2017]. Essa base de dados conta com os códigos identificadores de 2.573 usuários e de suas classes, sendo 32% deles classificados como *bots*. A API do Twitter não permite a coleta de metadados de usuários e dos seus *tweets* quando estão desativados, situação de muitos usuários dessa base, principalmente dos *bots*, o que impossibilitou seu uso para o trabalho proposto.

Assim, optou-se pela base de dados do projeto *Fake Project* [Cresci et al. 2017], por ser mais recente e considerar a característica evolutiva dos *bots*. A base possui dezenas de metadados de mais de 14 mil contas de milhões de *tweets*. Os *bots* são de diversos tipos e, relativamente, recentes, quando comparados às outras bases de dados. No intuito de adequar a base de dados aos objetivos deste trabalho, optou-se em utilizar uma amostra dos mais de 6 milhões de *tweets* disponíveis. A seleção se mostrou necessária após a observação do grande desbalanceamento da quantidade de *tweets* por usuários (Tabela 1).

3.2. Fase 2: Processamento

Nesta fase foi realizada a principal tarefa relacionada a extração do conhecimento. O entendimento dos dados guiaram as etapas deste processo, que consistiu na exclusão de dados irrelevantes e, *a posteriori*, na criação e transformação de *features* em dados/informações que subsidiem a tomada de decisão. No processo de transformação das *features*, para o tratamento de valores ausentes, duas estratégias foram utilizadas: *i)* a

Tabela 1. Comparação das Bases de Dados Original e Extraída

NOME DO GRUPO	USUÁRIOS			TWEETS			
	Quantidade Esperada	*Quantidade Extraída*	*Número de atributos*	*Quantidade Esperada*	*Quantidade Extraída*	*Número de atributos*	*Usuários Únicos*
Genuínas	3.474	3.474	42	8.377.522	2.839.362	25	1.084
Social Spambots #1	991	991	41	1.610.176	1.610.034	25	991
Social Spambots #2	3.457	3.457	40	428.542	428.542	25	3.457
Social Spambots #3	464	464	41	1.418.626	1.418.557	25	464
Tradicional Spambots #1	1.000	1.000	40	145.094	145.094	25	1.000
Tradicional Spambots #2	100	100	40	74.957	0	25	0
Tradicional Spambots #3	433	403	40	5.794.931	0	25	0
Tradicional Spambots #4	1.128	1.128	40	133.311	0	25	0
Seguidores Falsos	3.351	3.351	40	196.027	196.027	25	196.027

Fonte: Elaborada pelos autores.

exclusão dos atributos que já haviam sido descontinuados pela API ou que possuíam mais de 90% dos valores ausentes; e *ii)* o preenchimento com 0 para alguns atributos, como os *booleanos*, que possuíam todos os valores diferentes de 1 iguais a vazio. Para os atributos de texto, o preenchimento foi feito com uma *string* vazia como, por exemplo, a descrição do perfil do usuário. Esse tipo de atributo ainda foi transformado, posteriormente, em *booleano*, o que não permite saber se há existência ou não da descrição. Antes da seleção amostral, foram também excluídos atributos com variâncias próximas de zero.

As técnicas utilizadas neste trabalho são robustas à presença de valores discrepantes (*outliers*). Mesmo com os *outliers* presentes em vários atributos, todos os valores foram mantidos, pois faziam parte do problema de domínio. Para indução de regras, apesar de os algoritmos *PART* e *J48* lidarem bem com atributos numéricos, o algoritmo *APRIORI* se mostrou capaz de gerar regras por meio de atributos categóricos. Para utilização desse algoritmo, todos os atributos numéricos tiveram seus valores discretizados por frequência ou por intervalos que fizessem mais sentido na interpretação das regras.

Baseado nos trabalhos que propuseram uma série de atributos que pudessem ser explicativos quanto ao comportamentos dos *bots* [Lee et al. 2010, Stringhini et al. 2010], 19 novos atributos foram gerados (Tabela 2), assim como sua descrição, tipo e tabela de origem. Com a transformação e criação desses novos atributos, uma única base de dados foi formada, por meio da junção das tabelas de usuários com as de *tweets*, utilizando o código identificador do usuário. A base encontra-se no grão usuário e possui 2.682 usuários, sendo 61% de *bots*, caracterizados por 32 atributos.

A validação da base de dados, como etapa preliminar à extração das regras, é necessária para obtenção de evidências de que a amostra selecionada seja capaz de validar algumas das características já conhecidas na literatura que diferenciam os *bots* das contas genuínas. Posto isso, foram formuladas algumas hipóteses a partir de outros trabalhos [Stringhini et al. 2010, Varol et al. 2017, Dong and Liu 2018], que afirmam que os *bots:*

- Tendem a receber menos menções;
- Tendem a receber menos respostas a seus *tweets*;
- Tendem a receber um menor número de *retweets* de suas postagens;
- Tendem a serem contas mais jovens;
- Tendem a ter o nome mais longo.

Embora o simples cálculo dos valores médios, em alguns dos casos citados, já sinalizem a veracidade das afirmações, a grande dispersão das amostras reforçaram a necessidade de um teste estatístico. Assim, definiu-se, para os testes de hipóteses, todas as hipóteses alternativas como unilaterais, de modo que a rejeição das hipóteses nulas,

Tabela 2. Novos atributos

Atributo	Descrição	Tipo	Tabela
n_retweet_mean	Média dos retweets	numérica	tweet
n_favorite_mean	Média de favoritos nos tweets	numérica	tweet
n_hashtags_mean	Média de hashtags dos tweets	numérica	tweet
n_urls_mean	Média de urls dos tweets	numérica	tweet
n_mentions_mean	Média de menções nos tweets	numérica	tweet
size_twitter	Tamanho do tweet	numérica	tweet
media_size_twitter	Média do tamanho dos tweets	numérica	tweet
dp_size_twitter	Desvio padrão do tamanho do tweet	numérica	tweet
in_tweets_reply	Se tweet é uma resposta	binária	tweet
max_count_tweet_dia	máximo de tweets em um dia	numérica	tweet
count_dias_tweet	Dias que tweetou	numérica	tweet
in_tweet_dia	200/(count_dias_tweet)	numérica	users
in_following_followers	Amigos/Seguidores	numérica	users
age	Idade da conta	numérica	users
n_alfa_sname	Quantidade de letras screen name	numérica	users
n_number_sname	Quantidade de números screen name	numérica	users
size_sname	Tamanho do screen name	numérica	users
size_description	Tamanho da descrição	numérica	users
size_name	Tamanho do nome	numérica	users

Fonte: Elaborada pelos autores.

quando não há diferenças entre as classes, trará fortes evidências para as afirmações dos autores. O nível de significância utilizado foi de 5% para o erro tipo I. Para todas as amostras utilizadas não foi possível mostrar evidências de normalidade, então foi utilizado o teste não paramétrico *Mann-Whitney*. Somente a hipótese 5 não aceitou a hipótese alternativa, o que demonstra a congruência da amostra extraída para essa pesquisa, com os relatos da literatura. Nesse sentido, espera-se que os resultados extraídos aqui possam ser mais facilmente generalizados.

3.3. Fase 3: Pós-Processamento

Essa etapa contou com um processo cíclico e iterativo, principalmente nos ajustes de parâmetros dos algoritmos aplicados, visando a obtenção de resultados que possam ser melhor interpretados por humanos e descrever de forma mais eficiente o comportamento dos *bots*. Para todas as técnicas e algoritmos utilizados, foi realizada a fase de treino em apenas 66% do conjunto. O restante foi utilizado para testes e avaliação do modelo. Para a divisão, preocupou-se em manter a proporção das classes entre os conjuntos, uma vez que a concentração da classe de *bots* nos dois conjuntos é de 61%.

Uma árvore de decisão foi gerada por meio do algoritmo *J48*, que se baseia no algoritmo C4.5 [Quinlan 1993], em que o conhecimento é extraído baseado em condições do tipo "se-então". O principal objetivo foi a estruturação do conhecimento da base de dados de maneira sequencial e de fácil interpretação, não deixando de considerar as métricas da capacidade de classificação do modelo construído. A Figura 2 ilustra a árvore extraída do conjunto de treinamento com a opção de poda. Foi utilizado o ganho de informação como critério de divisão, com o número mínimo de elementos por folha igual a 20. No topo da árvore estão os atributos mais importantes. Os caminhos da raiz à folha formam uma regra, sendo cada folha uma condição dessa regra.

Após a aplicação do modelo treinado ao conjunto de testes, a árvore teve seu desempenho medido em cada nó, como mostrado na Tabela 3. As métricas avaliadas foram as de Cobertura, Confiança e *Lift*, que é a razão da confiança pela taxa total de contas genuínas. A fim de avaliar o poder de classificação de todo o modelo em todos os limiares de decisão, calculou-se a Área sob a Curva ROC[2]

[2]Área sob a curva ROC (AUC-ROC - *Area Under the Receiver Operating Characteristic curve*) é uma

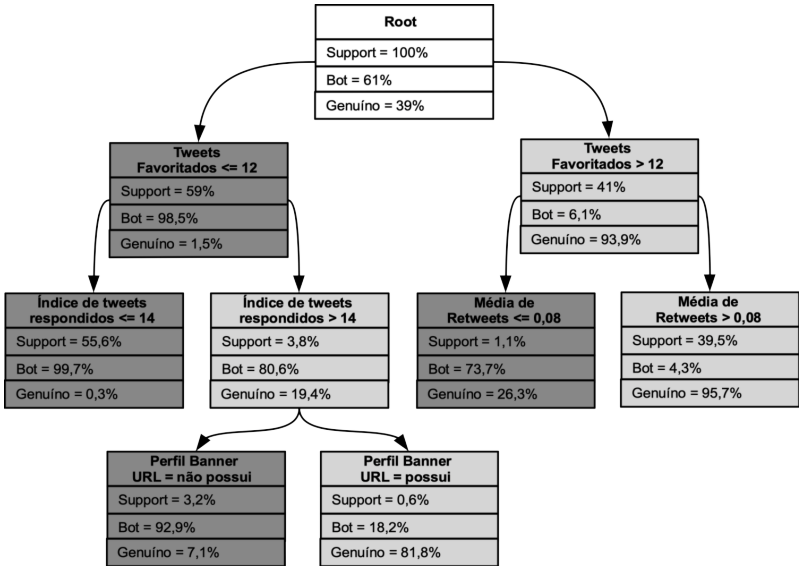

Figura 2. Árvore de decisão.

Tabela 3. Regras da Árvore de Decisão

REGRAS DA ÁRVORE	ALVO	COBERTURA	CONFIANÇA	LIFT
Tweets favoritados <= 12	1	59%	1%	0,04
Tweets favoritados <= 12 e Índice de *tweets* respondidos <= 14	1	55%	0%	0,01
Tweets favoritados >12	0	41%	94%	2,40
Tweets favoritados >12 e Média de *retweets* >0.08	0	39%	95%	0,05

Fonte: Elaborada pelos autores.

As regras induzidas, ao contrário da árvore, não dividem seu espaço de entrada pelo suporte, possibilitando encontrar nichos de conhecimento que podem passar despercebidos pelas árvores. As regras foram induzidas por meio dos algoritmos *PART, JRIP* e *APRIORI*. Para as regras geradas pelo algoritmo *APRIORI*, apenas as com até 3 condições e com a classe alvo do lado direito da regra foram analisadas. Na Tabela 4 são apresentadas as principais regras que foram induzidas nesse trabalho, com os valores da Cobertura, da Confiança e o *Lift* de cada regra.

4. Discussão dos resultados

Examinando as regras produzidas pelos modelos, percebe-se, além do alto poder de classificação dos *bots*, a abertura de novas discussões acerca das principais características que os diferem das contas genuínas, ressaltando características do perfil do usuário e do seu comportamento. No caso da árvore de decisão, as características com maior poder de classificação são as mais próximas do topo da árvore, as quais são automaticamente ramificadas em limiares numéricos que melhor distinguem os *bots* das contas genuínas. Ou seja: o algoritmo divide o atributo em um melhor ponto de separação. No primeiro nó da árvore, a quantidade de vezes que o usuário favorita os tweets de outras contas *(favorites_count)* é capaz de classificar corretamente a maioria da amostra.

Esse atributo também aparece em 4 regras (Tabela 4) que foram induzidas por 2 diferentes algoritmos. Apesar do atributo ter sido dividido em diferentes limiares para

das métricas utilizadas para a avaliação de um modelo de aprendizado de máquina [Metz 1978], obtendo o valor de 0,97.

Tabela 4. Indução de Regras

ID	REGRAS	ALVO	COBERTURA	CONFIANÇA	LIFT
1	Tweets favoritados <= 10 e Idade da conta <= 1405 dias	1	58%	0%	0,01
2	Índice de tweets respondidos >11	0	41%	88%	2,25
3	Foto no banner do perfil = 1	0	41%	88%	2,26
4	Tweets favoritados >= 14 e Índice de tweets respondidos >= 26	0	32%	97%	2,48
5	Média de retweet >= 10.81 e Tweets favoritados >=603	0	25%	98%	2,52
6	Tweets favoritados =17 e Idade da conta >=1055 dias e Média de retweets =0.83	0	27%	98%	2,50
7	Foto no banner do perfil = 0 e Índice de tweets respondidos = 1º quartil	1	55%	2%	0,05
8	Foto no banner do perfil = 0 e Foto no perfil =1 e Quantidade de número no nome = 1º quartil e Índice de tweets respondidos = 1º quartil	1	52%	2%	0,06
9	Foto no banner do perfil = 0	1	59%	5%	0,13

Fonte: Elaborada pelos autores.

cada regra, em todos os casos os *bots* estão associados às contas que interagem com menor frequência, ou seja, favoritam menos *tweets* alheios. Em um trabalho relacionado [Lundberg et al. 2019],que utilizou uma base de dados mais recente, elaborada exclusivamente para o estudo em 2019, esse comportamento foi apontado no topo da árvore. Todavia, os autores utilizaram a relação quantidade de favoritos pela idade em dias das contas, relacionando os menores índices aos *bots*.

Além dos atributos inerentes ao próprio comportamento dos *bots*, que podem ser facilmente manipulados pelos seus programadores, evitando a captura pelos detectores, são apresentadas características que são extrínsecas ao seu comportamento, ou seja, não dependem unicamente das suas atividades. Como exemplo, a regra 2 da Tabela 4, que relaciona a maior quantidade de vezes que a conta tem seus *tweets* respondidos às contas genuínas. Já na regra 7, a maior quantidade de vezes que o usuário tem seus *tweets* reproduzidos por outros usuários *(retweet)*, aparece associada a uma característica controlada pelos usuários, possuir foto no *banner* do perfil, explicando bem as contas genuínas.

A regra com melhor *Lift* para a classe dos *bots* associa a quantidade de *tweets* favoritados a uma idade menor que 4 anos. Os *bots* são conhecidos por serem mais jovens, pois muitas vezes são criados com um propósito específico, sendo destruídos após sua utilização, pelos donos ou pela plataforma. Para a melhor regra das contas genuínas, a alta contagem de favoritos foi relacionada à quantidade de vezes que os usuários têm seus *tweets* reproduzidos *(in_tweets_mean)*, novamente demonstrando a importância dos atributos que representam a capacidade de interação e engajamento das contas na plataforma.

Essas condições, que refletem o relacionamento das contas no Twitter, demonstram a crescente preocupação da programação dos *bots* em grupos, de forma que trabalhem em rede de cooperação, para que os *bots* possam contribuir uns com os outros, de maneira que se assemelhem cada vez mais com as contas genuínas, pois assim elevam sua percepção de confiabilidade na rede. Este comportamento foi estudado por [Chavoshi et al. 2016], que propuseram em seu estudo uma abordagem de detecção focada no relacionamento das contas. Porém essas abordagens costumam ser mais custosas computacionalmente para serem implementadas quando comparadas a abordagem desta pesquisa. Intuitivamente, alguns atributos requerem menos dados para serem calculados enquanto outros requerem um grande volume. Outro fator importante é a quantidade de

requisições necessárias na API do Twitter para a obtenção dos dados.

Os atributos destacados pelos modelos não aparecem na lista dos principais atributos de detecção de *bots*, apontadas por [Cresci et al. 2015], que testaram, em uma versão anterior à mesma base de dados utilizada neste estudo, várias técnicas de classificação em diferentes conjunto de atributos. As diferenças desses resultados, bem como de outros estudos que também descreveram os *bots* [Stringhini et al. 2010, Ahmed and Abulaish 2013], reforçam a sensibilidade das técnicas às amostras. Todavia, ressalta-se que em nenhum desses trabalhos são propostos os melhores limiares para os atributos que distinguem os *bots* das contas genuínas como feito nesta pesquisa.

4.1. Métricas de avaliação das regras induzidas

O principal objetivo deste trabalho é a proposição de regras que possam explicar o comportamento dos *bots* e das contas genuínas no Twitter, a fim de identificá-los. Para isso, é importante que as métricas de avaliação das regras induzidas sejam capazes de indicar as melhores regras, de maneira ordenada, para ambas as classes. Poucos estudos avaliaram regras explicativas sobre o comportamento das contas no Twitter. Dentre esses, destaca-se um que avaliou regras propostas pelo mercado de agências e mídias sociais, utilizando as métricas de Acurácia, *F-measure, Recall, Precision* e *ROC curve* [Cresci et al. 2017]. Essas métricas, comumente utilizadas para a avaliação de classificadores binários, trazem confusão no esclarecimento das regras que melhor descrevem e classificam ambas as classes, pois uma regra pode ter um bom desempenho em uma métrica e não ter em outra.

Com isso, optou-se, neste trabalho, pela utilização de um única medida, o *Lift*, para o ordenamento das regras. O *Lift* é calculado por meio do deslocamento (para cima ou para baixo) da Confiança em relação à média populacional, independente da cobertura da regra. Ele permite encontrar pequenos nichos de conhecimento, que não são percebidos pela árvore de decisão, devido ao baixo ganho de informação que o peso da porcentagem da população da regra pode representar. A Confiança quantifica a frequência relativa do alvo da regra, na amostra da população selecionada pela Cobertura [Han et al. 2012].

4.2. Aplicação dos resultados

Embora os resultados tenham sido satisfatórios e extraídos de uma amostra de dados capaz de validar muitas hipóteses da literatura, a reutilização das regras em ambiente de produção ainda deve ser vista com cautela. A detecção de *bots* no Twitter ainda é um problema complexo, devido à constante evolução da plataforma e do crescente interesse por parte da sociedade no conteúdo compartilhado. As regras apresentadas aqui são oriundas de uma base de dados que, embora heterogênea, não representa todos os tipos de *bots* existentes no Twitter. A identificação dos diferentes tipos e estratégias dos *bots*, bem como a construção de modelos específicos para cada tipo, podem trazer melhores resultados quando aplicados ao mundo real.

Outro importante ponto é a necessidade da rápida evolução dos modelos para adequá-los às mudanças no comportamento, tanto das contas genuínas como dos *bots*. Essa adaptação pode ocorrer com a incrementação das bases de treino com os novos comportamentos das contas que vem surgindo. Todavia, a metodologia empregada aqui pode, em princípio, ser facilmente adaptada para essas mudanças.

5. Considerações finais

A análise de dados de mídias sociais vem crescendo fortemente nas últimas décadas, principalmente entre pesquisadores que desejam capturar essa grande quantidade de dados, das mais variadas fontes, de forma rápida e fácil, e extrair informação e conhecimento desses dados. No entanto, muitos dos dados disponíveis nas mídias sociais são produzidos ou publicados por *bots*, muitas vezes com intenção de disseminar notícias falsas. Isso pode interferir na análise de dados e produzir um resultado que não represente, de fato, informações contidas naquele conjunto de dados coletados.

O propósito deste trabalho foi criar um conjunto de regras que possibilitem a descrição e a identificação de *bots*, e a possível desconsideração dos dados gerados por eles em suas análises. Para isso, foram geradas regras que possibilitam identificar os comportamentos dos *bots* no Twitter e classificá-los, por meio dos seus metadados disponíveis pela API da aplicação. O método empregado utiliza de algoritmos que extraem conhecimento de uma amostra de uma base de dados da literatura, descobrindo nichos de conhecimento que ajudaram na discussão sobre o comportamento dos *bots*, além de propor regras com alto poder discriminante. Apenas os atributos disponíveis na plataforma do Twitter foram utilizados. A partir deles, por meio da preparação dos dados, guiada pelo problema do domínio, utilizou-se informações da literatura para excluir, transformar e criar novos atributos para aumentar o ganho de informação da amostra. Campos textuais não foram utilizados para extração de valores semânticos, o que facilitou a utilização do ferramental apresentado de maneira universal, sem distinção de idioma ou localização.

Testes de hipóteses validaram as informações conhecidas que diferenciam as contas genuínas dos *bots*, no Twitter, legitimando também a base de dados e a estratégia da seleção amostral. Em uma abordagem contrária, técnicas de indução de regras e árvore de decisão extraíram novos nichos de conhecimento sobre o comportamento das contas genuínas e dos *bots*. Os modelos gerados, apesar do alto poder discriminante, primam pela interpretabilidade do conhecimento extraído. As regras foram avaliadas pelas métricas de Cobertura, Confiança e *Lift* e o modelo da árvore de decisão teve o seu desempenho preditivo avaliado também pela métricas AUC e acurácia, obtendo 0,97 em ambas.

O modelo gerado se mostrou capaz de classificar *bots* no Twitter, com métricas satisfatórias de performance. No entanto, essas métricas não foram correlacionadas com as de outros modelos existentes. A sensibilidade dos modelos às amostras aos quais são treinados [Stringhini et al. 2010, Ahmed and Abulaish 2013] requerem que, para uma justa comparação, eles sejam reimplementados, treinados e aplicados na mesma base de dados deste estudo ou treinados em suas bases originais e aplicados em uma base de dados neutra. Estudos com esse propósito foram realizados em 2015 e em 2017 [Cresci et al. 2015], tendo este último avaliado o desempenho dos principais trabalhos relacionados para duas amostras da mesma base de dados utilizada aqui.

Assim, as principais contribuições desta pesquisa são: as regras para identificação e descrição de *bots* no Twitter e a metodologia empregada para extração do conhecimento que, guiada pelas informações da literatura e do domínio, resultou em modelo de classificação que auxilia no entendimento do comportamento dos *bots*. O código fonte das fases de pré-processamento e processamento estão disponíveis online [3].

[3]https://github.com/aliceleite/BotsFilterBraSNAM.git

Como trabalhos futuros, pode-se mencionar a reprodução do ferramental produzido neste estudo em uma base de dados mais recente. Outro importante ponto é a complementação dos modelos. Espera-se que os resultados da extração de conhecimento apresentados possam contribuir com a implementação de outros modelos de decisão binária mais robustos para classificação dos *bots*. Para a validação da técnica, também é importante a sistematização de testes comparativos com outros modelos, sobretudo os supervisionados e de caráter descritivo. Para isso, será necessário a reprodução dos modelos e a comparação dos resultados em uma base neutra.

Por fim, uma importante etapa futura pode ser a inserção das regras ao *framework* Oráculo, mencionado na Introdução. O desenvolvimento desse dispositivo disponibilizará para a comunidade de pesquisadores e analistas de dados uma ferramenta de alta usabilidade, que realiza a coleta e o tratamento dos dados do Twitter, buscando contornar as limitações de números de requisições à API do Twitter e contando com métodos e técnicas de identificação de *tweets* advindos de bots.

Referências

Ahmed, F. and Abulaish, M. (2013). A generic statistical approach for spam detection in online social networks. *Computer Communications*, 36:1120–1129.

Arias, M., Arratia, A., and Xuriguera, R. (2014). Forecasting with twitter data. *ACM Trans. Intell. Syst. Technol.*, 5(1):8:1–8:24.

Bessi, A., Coletto, M., Davidescu, G., Scala, A., Caldarelli, G., and Quattrociocchi, W. (2014). Science vs conspiracy: Collective narratives in the age of misinformation. *PloS one*, 10.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., and Wirth, R. (2000). Crisp-dm 1.0 step-by-step data mining guide.

Chavoshi, N., Hamooni, H., and Mueen, A. (2016). Identifying correlated bots in twitter.

Chu, Z., Gianvecchio, S., Wang, H., and Jajodia, S. (2012). Detecting automation of twitter accounts: Are you a human, bot, or cyborg? *Dependable and Secure Computing, IEEE Transactions on*, 9:811–824.

Cresci, S., Pietro, R., Petrocchi, M., Spognardi, A., and Tesconi, M. (2015). Fame for sale: Efficient detection of fake twitter followers. *Decision Support Systems*, 80.

Cresci, S., Pietro, R., Petrocchi, M., Spognardi, A., and Tesconi, M. (2017). The paradigm-shift of social spambots: Evidence, theories, and tools for the arms race.

Dong, G. and Liu, H. (2018). *Feature Engineering for Machine Learning and Data Analytics*. CRC Press, Inc., Boca Raton, FL, USA, 1st edition.

Ferrara, E. (2017). Disinformation and social bot operations in the run up to the 2017 french presidential election. *First Monday*, 22.

Goodman, B. and Flaxman, S. (2016). Eu regulations on algorithmic decision-making and a "right to explanation". *AI Magazine*, 38.

Han, J., Kamber, M., and Pei, J. (2012). Data mining concepts and techniques, third edition.

Kirkpatrick, K. (2016). Battling algorithmic bias: how do we ensure algorithms treat us fairly? *Communications of the ACM*, 59:16–17.

Lee, K., Caverlee, J., and Webb, S. (2010). Uncovering social spammers: social honeypots + machine learning. pages 435–442.

Lee, K., Eoff, B., and Caverlee, J. (2011). Seven months with the devils: A long-term study of content polluters on twitter.

Lundberg, J., Nordqvist, J., and Laitinen, M. (2019). Towards a language independent twitter bot detector. In *Proceedings of the Digital Humanities in the Nordic Countries 4th Conference, Copenhagen, Denmark, March 5-8, 2019*, pages 308–319.

Metz, C. E. (1978). Basic principles of ROC analysis. *Seminars in Nuclear Medicine*, 8(4):283–298.

Miller, Z., Dickinson, B., Deitrick, W., Hu, W., and Wang, A. (2014). Twitter spammer detection using data stream clustering. *Information Sciences*, 260:64–73.

Naaman, M., Boase, J., and Lai, C.-H. (2010). Is it really about me?: Message content in social awareness streams. In *Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work*, CSCW '10, pages 189–192, New York, NY, USA. ACM.

Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Saeed, Z., Abbasi, R., Maqbool, O., Sadaf, A., Razzak, I., Daud, A., Aljohani, N., and Xu, G. (2019). What's happening around the world? a survey and framework on event detection techniques on twitter. *Journal of Grid Computing*.

Silva Filho, R. L. C. and Adeodato, P. J. L. (2019). Data mining solution for assessing the secondary school students of brazilian federal institutes. In *2019 8th Brazilian Conference on Intelligent Systems (BRACIS)*, pages 574–579.

Stringhini, G., Kruegel, C., and Vigna, G. (2010). Detecting spammers on social networks. In *Proceedings of the 26th Annual Computer Security Applications Conference*, ACSAC '10, pages 1–9, New York, NY, USA. ACM.

Taigman, Y., Yang, M., Ranzato, M., and Wolf, L. (2014). Deepface: Closing the gap to human-level performance in face verification.

Varol, O., Ferrara, E., Davis, C. A., Menczer, F., and Flammini, A. (2017). Online human-bot interactions: Detection, estimation, and characterization. *CoRR*, abs/1703.03107.

Xia, X., Shihab, E., Kamei, Y., Lo, D., and Wang, X. (2016). Predicting crashing releases of mobile applications. In *Proceedings of the 10th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement*, ESEM '16, New York, NY, USA. Association for Computing Machinery.

Yang, K.-C., Varol, O., Davis, C. A., Ferrara, E., Flammini, A., and Menczer, F. (2019). Arming the public with artificial intelligence to counter social bots. (December 2018):48–61.