Detecção de Bots Sociais: Uma Discussão sobre o Tempo de Vida de Abordagens Tradicionais

Erick N. da Mata[1], Gabriela M. S. Dias[1], Ronaldo M. Salles[1]

[1]Instituto Militar de Engenharia (IME) 22.290-270 – Rio de Janeiro – RJ – Brazil

{erick,gabriela,salles}@ime.eb.br

Abstract. *Social bot detection algorithms seem to aim a common thread: accuracy. However, if accuracy can be lost as bots evolve, the lifetime of these solutions must also be discussed. Using datasets from 2010 to 2018 to represent this evolution, three traditional models were evaluated to understand this lifetime. Looking at the performance changes, it was possible to see a loss of accuracy over the years and that it reflects a gradual change in the behavior of bots. Other factors such as heterogeneity of data and the ability to remain capable of detecting human accounts over the years (true negatives) are also discussed.*

Resumo. *Trabalhos de detecção de bots sociais parecem focar em um ponto comum: a acurácia. Mas, se a acurácia pode ser perdida com a evolução dos bots, o tempo de vida útil destas soluções também deve ser discutido. Utilizando datasets de 2010 até 2018, para representar essa a evolução, três modelos tradicionais foram avaliados para compreensão desse tempo de vida. Observando as alterações de desempenho, foi possível perceber uma perda de acurácia ao longo dos anos e que ela reflete uma alteração gradual do comportamento dos bots. Outros fatores, como a heterogeneidade dos dados e a capacidade de manter-se capaz de detectar contas humanas ao longo dos anos (verdadeiros negativos) também são discutidos.*

1. Introdução

Bots sociais são uma realidade, permeiam as redes sociais online (RSO) e interagem com usuários humanos constantemente [Ferrara et al. 2016]. Podem produzir conteúdo legítimo, divulgam notícias verídicas ou apenas automatizam tarefas. Mas, como seus algoritmos podem imitar comportamentos realizados por contas humanas [Ferrara et al. 2016], outras ações podem ocorrer; por exemplo, a propagação de notícias intencionalmente inverídicas (*fake news*) [Wang et al. 2018], ataques de ódio ou phishing [Latah 2020]; além disso a comercialização de tais atividades como serviço pode ocorrer, facilitando crimes e o abuso de políticas das RSO [Cresci et al. 2015].

Diversos trabalhos apontam uma constante evolução em seu padrão de atividade e isto pode ser observado na própria visão de sua classificação. Em [Lee et al. 2011] *bots* sociais eram percebidos como contas que produziam conteúdo automatizado, gerando poluição nas RSO. [Ferrara et al. 2016] apontou a mimetização do comportamento humano ao interagir com outros usuários. Dois anos após, [Wang et al. 2018] adiciona a visão desses *bots* como *softwares* que coabitam as RSO e interagem mediante uma designação. Logo, pode se dizer que há um "mestre" controlador destas

contas. Evoluindo e se organizando, elas hoje devem ser vistas como potenciais comunidades [Cresci 2020]. Esta nova geração de *bots* é vista em [Latah 2020] como perfis automatizados que utilizam as RSO como canais de comando e controle. Em [Santos et al. 2021] é demonstrado que tais contas agravam a influência sobre a opinião pública. Nas eleições presidenciais dos Estados Unidos em 2016 e do Brasil em 2018, foi percebidas atuação de contas automatizadas nas discussões realizadas nas RSO [Santos et al. 2020, Santos et al. 2021]. [Egli et al. 2021]

Não é simples detectar estas contas, pois possuem formas e intenções diversas [Latah 2020; Sayyadiharikandeh et al. 2020]. O modelo construído com tal finalidade irá depender de sua capacidade de generalização para classificar corretamente exemplos nunca vistos, que podem ser contas ainda não detectadas ou padrões inéditos [Echeverría et al. 2018]. Assim, o treinamento é baseado em exemplos conhecidos, enquanto sua tarefa (detecção de *bots*) será classificar contas não conhecidas. Essa é uma dificuldade natural para todos os modelos, mas nem todo objeto classificado muda com o tempo. *Bots* precisam evoluir para sobreviver a detecção e, caso não se adaptem a isso, é esperado que a eficiência dos classificadores decaia com o tempo. Isto pode tornar sua vida útil mais breve e ameaçar a segurança dos sistemas que confiam nos modelos de detecção. Discutir sobre seu tempo de vida é necessário.

Assim, o presente trabalho avalia a capacidade de três abordagens tradicionais em classificar corretamente *bots* mais avançados que os utilizados em seu treinamento. Para isso, foram utilizados conjuntos de dados de quatro artigos científicos. Distribuídos entre 2010 e 2018, que contêm dados sobre o passado de contas humanas e *bots* sociais, com intervalo médio de 2 anos entre *datasets*. Atualizando dados das contas ainda ativas, um quinto conjunto de dados foi produzido. Esses cinco conjuntos foram organizados pela data de atualização em ordem crescente para compor seis intervalos (ou janelas) de tempo e representar a evolução dos exemplos utilizados nos artigos.

Para cada janela de tempo (ex.: 2010), três modelos tradicionais foram treinados para detecção de *bots* e suas alterações de desempenho ao longo do tempo foram avaliadas ao apresentar dados inéditos e capturados no futuro (ex.: 2012, 2014, ...). Como medida de avaliação, foi utilizada a acurácia, que evidenciou melhor balanceamento dos resultados e apresenta uma medida ponderada entre erros e acertos na classificação. Notou-se que *bots* sociais de um momento futuro possuem alguma semelhança com exemplos conhecidos, mas que essa reduz gradativamente. Foi percebido também que a heterogeneidade dos dados pode impactar sua capacidade de generalização do modelo, e que diferentes abordagens tradicionais possuem problemas semelhantes ao classificar dados de janelas de tempo futuras ao seu treinamento.

Após a seção de introdução, o presente artigo está organizado da seguinte forma: a Seção 2 apresenta alguns trabalhos na área de detecção de *bots* sociais e alguns cuja discussão se assemelha a proposta neste trabalho, a Seção 3 dispõe sobre o método, os conjuntos de dados utilizados e as abordagens de *machine learning* escolhidas, a Seção 4 possui as análises e discussões sobre os resultados, organizados por abordagem e a Seção 5 concluí com as considerações finais.

2. Trabalhos Relacionados

A detecção de contas automatizadas, ou *bots* sociais, é uma atividade que completou uma década desde o primeiro artigo publicado em 2010, tratando da detecção de spam na rede social Twitter [Cresci 2020].

Para a construção de um modelo de detecção, é necessário decidir se sua abordagem será geral (classificação independente da finalidade específica da conta) ou especializado em reconhecer algum tipo de atividade de interesse, como a propagação de fake news, disseminação de spam, etc [Cresci et al. 2017]. Dentre as possibilidades de construção de um modelo para a tarefa de classificação, [Ferrara et al. 2016] organiza as abordagens em quatro principais categorias baseadas em grafos, *crowdsourcing*, comportamentais e híbridas. [Latah 2020] categoriza essas abordagens em grandes grupos baseados em grafos, *machine learning* e abordagens em evolução.

[Stein et al. 2011] demonstrou o Facebook Immune System. Tratando a rede social como um grafo composto por conexões legítimas, utilizou diferentes abordagens de classificação, observando suas interligações, para detecção de contas ilegítimas. Uma solução que exige um grande processamento dos dados.

[Cresci et al. 2017] utilizou *crowdsourcing* em um experimento para classificação de *bots* tradicionais e da nova geração. A capacidade humana em identificar esses novos *bots* demonstrou-se baixa. Tal abordagem exige o emprego de mão de obra ou envolvimento ativo de uma larga comunidade, e pode não ser a melhor abordagem para analisar grandes volumes de dados [Ferrara et al. 2016].

[Jr et al. 2018] combinou abordagens comportamentais e análise textual para potencializar a detecção, analisando o tema do texto. Considerou inclusive *bots* legítimos e suas intenções. Mesmo robustos, esses modelos são sensíveis às características de treinamento e tornam-se influenciados pelo padrão aprendido, o que pode levar a uma maior facilidade na evasão por parte de *bots* sociais [Latah 2020].

[Chavoshi et al. 2016] propôs solução para detecção ao correlacionar atividades de contas com intervalos dinâmicos de tempo, calculando a distância entre os comportamentos, para realizar uma classificação observando a atividade em grupos. É preciso avaliar sua capacidade de lidar com as evoluções comportamentais.

Todos são modelos com boas taxas de detecção, mas terão sua eficiência posta a prova diante de novas abordagens de construção de *bots* sociais [Cresci 2020]. Mesmo conhecendo esse fato, não há muitos trabalhos que busquem compreender a relação entre as taxas de acurácia dos modelos e seu tempo de vida útil. [Cresci et al. 2017] apresenta evidências da evolução dessas contas e demonstra a queda de capacidade de detecção de modelos, mas não estabelece um tempo para os fatores de queda. [Cresci et al. 2019] introduz um algoritmo para simulação da evolução comportamental, algo extremamente útil para melhorar a capacidade dos modelos, mas não discute como o tempo pode afetar esses fatores. [Sayyadiharikandeh et al. 2020; Yang et al. 2020] discutem muito bem os fatores que influenciam diferenças entre conjuntos de dados e os problemas inerentes a evolução dos *bots* sociais percebidos entre *datasets*, mas não os fatores de tempo. [Rauchfleisch and Kaiser 2020] incluem a mudança de classificação no tempo em sua discussão, mas seu foco principal é avaliar o problema dos falsos positivos nas ferramentas de detecção.

Em contraste, este trabalho não propõe solução geral de alta eficiência ou abordagem inovadora para melhor classificação dos *datasets*. Tais soluções são importantes, e é possível notar em [Latah 2020] que o tema é bem discutido. A contribuição deste artigo é adicionar ao tema uma discussão sobre o comportamento dos modelos tradicionais em função do tempo, verificando suas alterações de desempenho ao propor um futuro simulado.

3. Método Proposto

Mesmo com grande número de trabalhos publicados sobre o tema com foco na RSO Twitter, [Samper-Escalante et al. 2021] demonstra que há poucos conjuntos de dados disponíveis. Boa parte dos *datasets* são privados. Um possível motivo para isto é a preocupação que a exposição de dados sensíveis viole termos da rede social.

O Botometer[1] é um dos poucos repositórios dedicados à *bots* sociais. Alguns dos conjuntos de dados disponíveis contém informações do passado, obtidas pela API da rede social Twitter, o que tornou possível a representação do passado de maneira consistente. Foram selecionados quatro *datasets* com essas características, detalhados na Tabela 1: caverlee-2011 [Lee et al. 2011], cresci-2015 [Cresci et al. 2015], cresci-2017 [Cresci et al. 2017] e midterm-2018 [Yang et al. 2020].

Tabela 1. Composição dos *datasets*

Dataset	Ref.	Descrição	Ano	Contas
caverlee-2011	[1]	*Bots* sociais que interagiram com *honeypots*.	2010	22.223
	[2]	Contas humanas legítimas	2009	19.276
cresci-2015	[3]	Seguidores falsos comercializados pela Fast Followerz	2014	1.169
	[4]	Seguidores falsos comercializados pela Intertwitter	2014	1.337
	[5]	Seguidores falsos comercializados pela Twitter Technology	2014	845
	[6]	Contas humanas envolvidas em momento político	2014	1.481
	[7]	Contas humanas verificadas	2014	469
cresci-2017	[8]	*Bots* utilizados para *retweets* de um candidato italiano	2016	991
	[9]	*Spambots* para publicidade de aplicativos	2016	3.457
	[10]	*Spam* de produtos (venda online pela Amazon.com)	2016	464
	[11]	*Spambots* de oferta de emprego	2016	433
	[12]	Outro conjunto de *spambots* de oferta de emprego	2016	1.128
	[13]	Seguidores falsos comercializados como serviço	2013	3.351
	[14]	Contas humanas verificadas	2016	3.474
midterm-2018	[15]	*Bots* anotados manualmente	2018	42.446
	[16]	Contas humanas anotadas manualmente	2018	8.902

[1]Projeto conjunto do Observatory on Social Media (Universidade de Indiana - EUA): https://botometer.osome.iu.edu/bot-repository/datasets.html

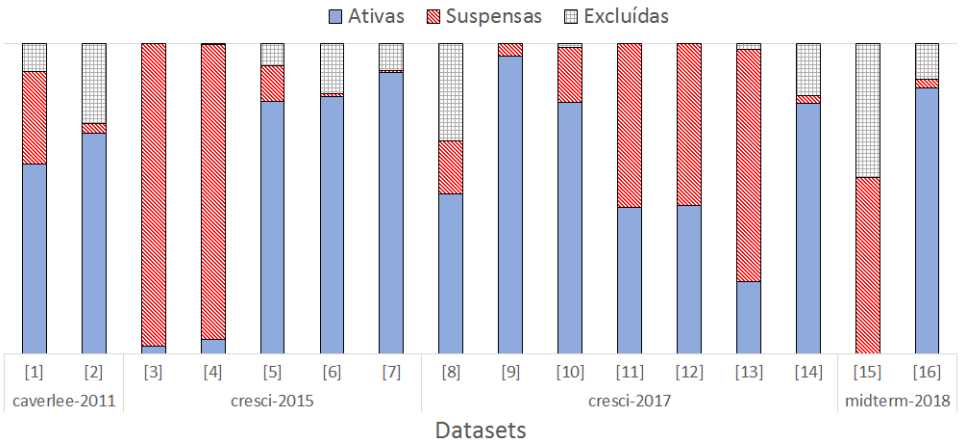

Figura 1. Estado atual das contas registradas nos *datasets*

Há grande pluralidade entre as contas e em seus estados. A Figura 1 apresenta a diversidade de estados por cada *dataset*. É possível perceber uma grande quantidade de contas já suspensas e excluídas, vide o conjunto [15], onde há somente 86 contas ativas (na elaboração da figura). Nos conjuntos [3], [4] e [13], compostos por perfis falsos comercializados, bem como em [11] e [12], que são *bots* tradicionais de oferta de empregos, há predominância de contas suspensas [Cresci et al. 2017]. Nota-se que grande parte das contas classificadas como *bots* sociais por esses trabalhos continuam ativa na RSO.

Como não é possível ter precisão sobre o momento exato de alteração do estado da conta, e para simular da melhor forma o passado, todas as contas (ativas, suspensas e excluídas) foram mantidas nos conjuntos de treino, validação e teste.

Assim, os dados foram agrupados por ano de referência e data de atualização, simulando sua data de captura. A seleção de atributos considerou a presença da informação em todos os *datasets*, pois diferentes artigos utilizam diferentes atributos. Ainda, a API do Twitter deixou de informar algumas das variáveis ao passar dos anos. Isto é um problema para um modelo tradicional, construído para receber um conjunto fixo de parâmetros. Para evitar essas inconsistências, os conjuntos foram padronizados com os mesmos atributos. O esquema dos dados é demonstrado na Tabela 2.

Tabela 2. Esquema dos dados

Atributo	Descrição
name	total de caracteres no nome.
username	total de caracteres no nome de usuário.
description	total de caracteres no campo descrição (bio).
statuses_count	total de *tweets* publicados.
friends_count	total de contas seguidas pelo perfil.
followers_count	total de contas que seguem perfil.
avg_tweet	média de *tweets* publicados por dia no intervalo *acc_age*
name_similarity	percentual de similaridade entre os campos *name* e *username*
description_variety	total de caracteres (sem espaço, ponto ou repetição) no campo descrição.
acc_age	total de dias entre a criação da conta e a data de atualização original dos dados.
bot	classificação da conta como *bot* ou humano.

3.1. Métodos de Aprendizagem

Três métodos de aprendizado supervisionado foram escolhidos: k-Nearest Neighbors (kNN), Support Vector Machines (SVM) e Multilayer Perceptron (MLP). Suas implementações foram realizadas em linguagem R (RStudio versão 1.4.1106) e em Python (Google Colaboratory). A motivação e hipóteses para cada método é discutida nesta subseção, enquanto os parâmetros para construção dos modelos estão dispostos na seção 4.

O kNN é um método que avalia similaridades no espaço de dados, utilizando uma constante *k*. Ele calcula a proximidade do registro avaliado a uma quantidade *k* de vizinhos, possibilitando classificação de registros por semelhança [Kramer 2013]. Esse modelo já foi muito utilizado em tarefas de visualização de dados, como realizado por [Yang et al. 2020], onda nota-se que diferentes *datasets* geram espaços de dados desiguais. É esperado que isto afete a classificação do kNN.

O modelo foi implementado em linguagem Python com uso da biblioteca Scikit Learn. Por não haver processamento de textos, é possível que o espaço de dados reflita a disposição dos registros nos *datasets*. Assim, o método foi escolhido para observar o quanto o kNN é sensível às diferenças nos dados ao longo das janelas de tempo. Como a acurácia do modelo é um indicador afetado pela disposição dos dados, como hipótese, espera-se que ela decaia ao longo dos anos.

SVM também observa os registros como pontos no espaço, mas usa outro critério de separação, o que a torna tão poderosa quanto as redes neurais [Cortes and Vapnik 1995]. O método utiliza vetores de suporte e calcula um hiperplano para segmentar os dados. Mesmo para conjuntos onde essa divisão é muito difícil, ainda é possível segregá-los com o truque de Kernel, uma distorção temporária no espaço de dados [Cortes and Vapnik 1995]. Modelos baseados em SVM são muito presente na detecção de *bots* sociais, como demonstrado em [Latah 2020].

O modelo foi implementado em linguagem R, por meio da biblioteca e1071. É esperado que a alteração comportamental e evolução dos *bots* sociais cause dispersões dos registros, impactando a disposição dos pontos no espaço. Com a evolução da mimetização do comportamento humano [Ferrara et al. 2016], pode haver uma tendência natural de aproximação dos dados no plano. Logo, a classificação de conjuntos não linearmente separáveis pode ser afetada por esta mudança. Espera-se que o SVM lide melhor com estas alterações, entretanto, reduza sua acurácia com o tempo.

MLP, uma abordagem em redes neurais artificiais, é largamente utilizada em tarefas de classificação. Trata-se da criação de uma rede composta por nós (neurônios ou perceptrons) com diversas camadas. Em cada camada, o nó estará totalmente interligado com os nós da camada seguinte, propagando sinais unidirecionais. Cada ligação (ou aresta) terá um peso, que pode potencializar ou reduzir o sinal. Ao receber o sinal, o nó realiza um processamento sobre ele e propaga o resultado por suas ligações, até a camada seguinte. O sucessivo processamento por camadas é capaz de promover a separação de conjuntos complexos, tornando-o um método altamente eficaz e generalizável [Popovic 2000]. Esta técnica foi escolhida em [Braz and Goldschmidt 2018] para detecção de *bots* sociais.

O modelo foi implementado em linguagem Python, com auxílio da biblioteca Tensorflow. Sua alta capacidade de generalização e a habilidade de segmentar conjuntos não linearmente separáveis [Popovic 2000] foram os critérios de escolha, mas é esperado que tais características tenham um limite. Como hipótese, espera-se que sua acurácia seja mantida sobre dados futuros em período maior que as abordagens anteriores.

3.2. Conjuntos de Dados

Os dados foram agrupados em 6 *datasets*, com 4.000 exemplos cada, balanceados com 2.000 contas humanas e 2.000 contas de *bots* sociais. Seu agrupamento seguiu a data de atualização dos registros. A Tabela 3 apresenta os conjuntos finais, e seus índices referenciam a Tabela 1.

Tabela 3. Composição dos *datasets*

	Ano de representação do *dataset*					
Classificação	2010	2012	2014	2016	2018	2021
Humanas	[1]	[1]	[6] [7]	[14]	[16]	Todos atualizados
Bots sociais	[2]	[13]	[3] [4] [5]	[8] [9] [10] [11] [12]	[15]	Todos atualizados

Para *datasets* (como cresci-2017) onde a composição se dava por múltiplas tabelas, os percentuais da composição original foram mantidos na amostragem. A cada modelo foi treinado, validado e testado somente dados de sua janela de tempo. Em seguida, o melhor modelo de cada abordagem tradicional (kNN, SVM e MLP) foi sequencialmente testado com dados inéditos de janelas de tempo futuras. Para cada janela, três testes foram realizados com 800 amostras aleatórias, igualmente balanceadas (400 de cada classe), sem reposição. Assim, seu comportamento foi observado.

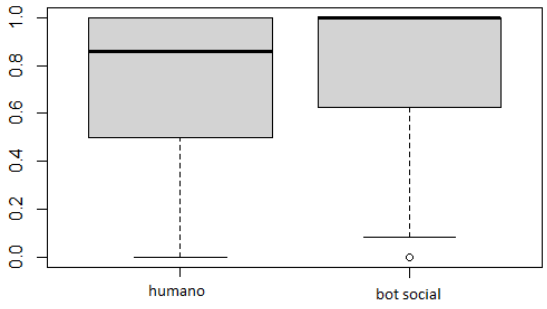

Figura 2. Box-plot do atributo name_similarity

Para enriquecer as relações entre as variáveis, foi criado um atributo para o cálculo de similaridade entre o nome de usuário (*username*) e o nome do usuário (*name*). Um problema nesse cálculo é o uso de caracteres estilizados e não pertencentes ao alfabeto

inglês ou latino. Por exemplo, o alfabeto Mathematical Markup Language (MathML)[2]. Usuários percebem a letra 𝔅 como "B", mas seus códigos Unicode são diferentes e a comparação computacional resultaria em "Falso". Para lidar com esse problema foram realizadas buscas automatizadas para obter o caractere de referência no site Compart[3]. Então, foram retirados espaços, pontuações e letras maiúsculas foram alteradas para minúsculas. A seguinte fórmula foi aplicada $[(A \cap B)/menor\{A,B\}]$, na qual A será o atributo "name" e B, "username". A Figura 2 demonstra que existe diferença no grau de similaridade calculado e que contas humanas apresentaram menor similaridade.

Outro problema encontrado pela escassez de *datasets* públicos com informações do passado foi a ausência de registros específicos para o período de 2012. A solução proposta para o preenchimento desta lacuna foi a composição igualmente balanceada dos dados dos *datasets* [1] e [13], com atualizações em 2010 e 2013. Buscando compreender como os atributos quantitativos sofrem alterações, foram comparados os dados originais com dados atualizados (2021), resultando em 92 contas ainda ativas. A Figura 3 demonstra que a grande maioria das contas sofreu redução significativa nesses indicadores, enquanto os demais atributos (omitidos na figura) sofreram pouco ou nenhuma variabilidade. Assim, em caráter experimental os dados foram mantidos.

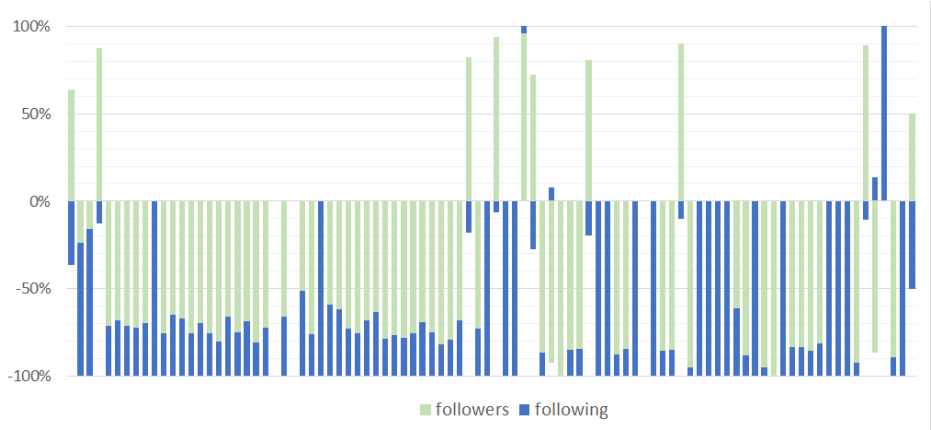

Figura 3. Crescimento e redução das conexões

4. Discussão e Análises

Esta seção apresenta os resultados referentes aos três métodos de aprendizado supervisionado, kNN, SVM e MLP, respectivamente nas seções 4.1, 4.2 e 4.3. Com base nos resultados, são apresentadas análises e discussões. Para cada método, é apresentado um gráfico para as taxas médias de acurácia e um retângulo vermelho claro representa o limiar inferior de acurácia (75%). O primeiro ponto de cada linha representa a acurácia do modelo em seu *dataset* (marco) e os próximos pontos indicam a acurácia média nos três testes de marcos futuros. Deste modo, a linha representa a oscilação da acurácia de cada modelo ao longo dos anos, ao avaliar amostras compostas por registros de um período futuro, e cada ponto inclui uma barra de erro de 5%. Assim,

[2] https://www.w3.org/TR/MathML2/overview.html

[3] https://www.compart.com/en/home

o modelo treinado com dados de 2016 inicia sua linha nesse ano e sua acurácia média nos testes de 2018 e 2021 determinará os pontos seguintes no gráfico.

4.1. kNN

Ao todo foram criados mais de 500 modelos. Como hiperparâmetros, foram testados valores de *k* entre 1 e 19, funções de peso uniforme e por distância, métricas de distância Euclidiana, Manhattan, Chebyshev, Minkowski e Hamming. Sua combinação possibilitou avaliar a resposta do modelo aos dados de validação em cada janela de tempo, sendo escolhido o de melhor acurácia. Apesar da função disponível na biblioteca Scikit Learn possuir um grande número de parâmetros possíveis, o modelo não foi muito sensível às alterações e manteve-se com bom nível de classificação para as métricas Euclidiana e Hamming. Os melhores valores de *k* foram 5 e 9. A função peso por distância produziu resultados ligeiramente melhores.

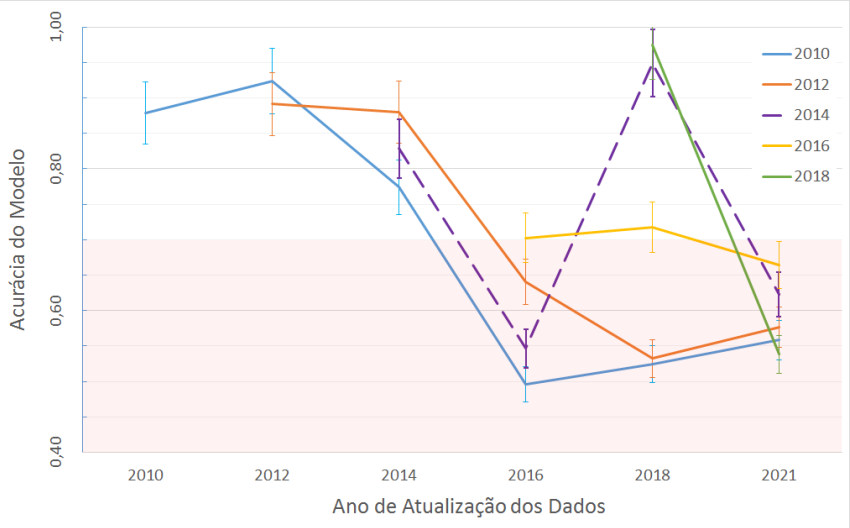

Figura 4. kNN - Taxas médias de acurácia

A Figura 4 apresenta os cinco melhores modelos kNN de cada janela de tempo, cada curva no gráfico inicia em seu marco e é seguida pela representação da sua acurácia ao longo dos anos. É possível notar que:

- 2010 e 2012: as curvas possuem quedas mais suaves na acurácia em relação às demais, mas que diferem entre si. Possivelmente pela prevalência de características semelhantes até o marco de 2014, as taxas médias de acurácia permaneceram acima do limiar estabelecido. A partir de então, as curvas ultrapassaram o limiar e mantiveram-se abaixo do nível mínimo de acurácia.
- 2014: houve uma queda rápida na acurácia já no marco seguinte. Isso pode estar ligado à diversidade apresentada pelo *dataset*, que introduz novos padrões de *bot* social em [Cresci et al. 2017]. Em contraste, houve uma grande oscilação positiva do modelo no marco de 2018. Isso pode estar relacionado a homogeneidade do *dataset* em [Yang et al. 2020].

- 2016: o modelo foi sensível à distribuição heterogênea dos dados e não convergiu para melhores taxas de acurácia, mantendo-se estável em 2018, decaindo em seguida.
- 2018: é possível observar uma queda brusca no marco seguinte, o que corrobora com a possibilidade da homogeneidade dos dados afetar o modelo diante de novos dados para classificação.

4.2. SVM

Foram criados mais de 100 modelos com o auxílio da função *tune* (pacote e1071), que avalia e extrai o melhor modelo ao variar parâmetros de teste. Como parâmetros, foram utilizados *kernel trick* linear, radial, polinomial e sigmoide; custos entre 0.5, 1.0, 2.0, 3.0 e 5.0; e função gamma com intervalo entre 0.5 a 2.0, com incremento de 0.5.

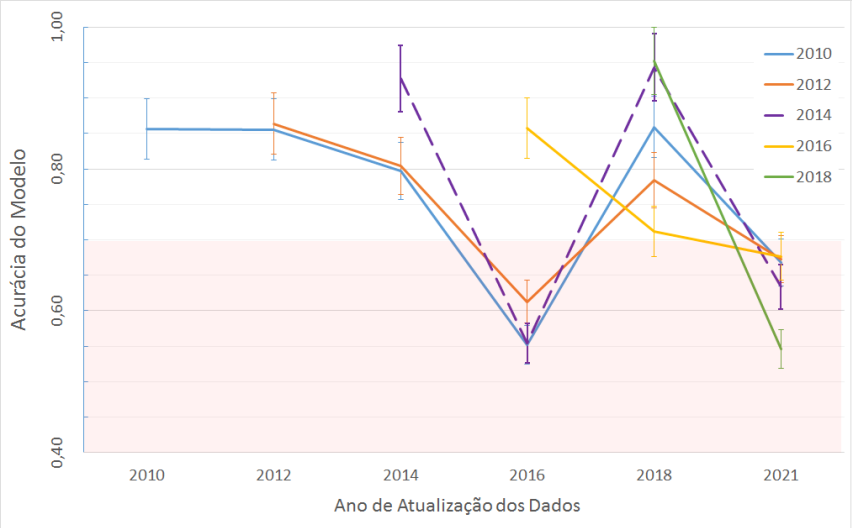

Figura 5. SVM - Taxas médias de acurácia

Modelos SVM apresentaram certa similaridade em suas capacidades de classificação. Com exceção da curva iniciada em 2016, obtiveram melhor acurácia ao avaliar conjuntos com maior homogeneidade e apresentaram dificuldades em grupos heterogêneos. Na Figura 5 é possível perceber que:

- 2010 e 2012: houveram oscilações semelhantes, tendo o modelo de 2012 apresentado mais estabilidade, tanto na queda, quanto na recuperação de sua acurácia. É possível que isso tenha ocorrido por influência de dados ligeiramente mais heterogêneos que o conjunto de 2010.
- 2014: a curva apresentou a mesma queda e recuperação para o marco seguinte, potencialmente pelos mesmos motivos dos modelos anteriores.
- 2016: a curva exibiu comportamento contrário aos demais, apresentando queda no marco 2018. Esse é um fator relevante, uma vez que seu conjunto de dados de treino possui maior diversidade que todos os outros. Pode representar uma baixa capacidade de generalização.

- 2018: apresentou a pior acurácia no marco de 2021, possibilitando inferir que a homogeneidade dos dados foi um potencial problema.

4.3. MLP

Ao todo foram criados mais de 40 modelos. Três topologias diferentes foram utilizadas para criação dos modelos[4]: (10→8→4→1); (10→6→3→1); (10→12→8→1). Para seus hiperparâmetros, foram utilizados os algoritmos de otimização RMSProp e Adagrad; função de perda Entropia Cruzada Binária. O treinamento foi realizado considerando um período de 100 épocas (mínimo) até 500 épocas. A partir de 100 épocas, uma nova validação foi realizada a cada 10 novas épocas. Conforme o resultado da validação do modelo, parâmetros adicionais foram variados, quais sejam a taxa de aprendizagem em 0,001 e 0,01, e o momento em 0,1 e 0,9. A manipulação destes parâmetros no início do treinamento prejudicou a variabilidade de resultados do modelo. Predominantemente, o RMSProp apresentou melhores resultados em todas as topologias.

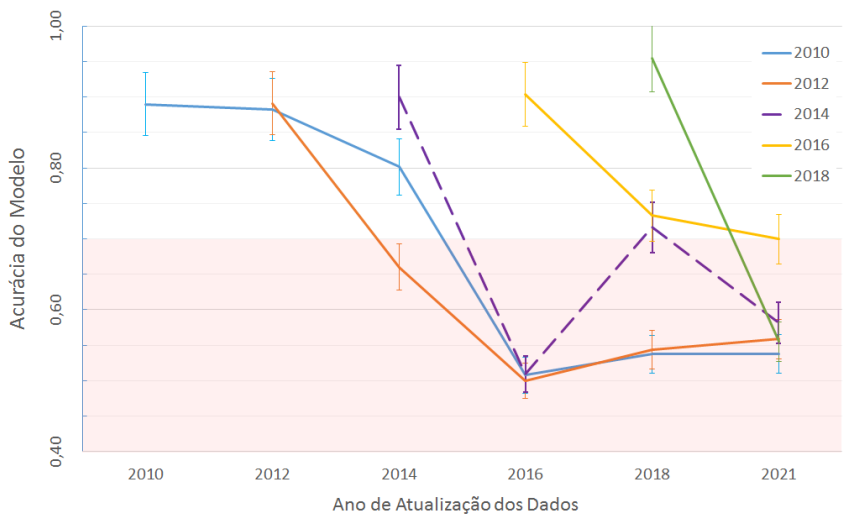

Figura 6. MLP - Taxas médias de acurácia

Na Figura 6, nota-se que cada modelo obteve bons níveis de acurácia em sua própria janela de tempo, mas, gradativamente, perdeu sua capacidade de classificação. Observando as curvas de cada modelo é possível se inferir que:

- 2010: o modelo declinou de maneira mais suave e foi capaz de manter-se acima do limiar por dois outros marcos seguintes. Após alcançar seu ponto mais baixo de acurácia em 2016, não houve recuperação.
- 2012: teve acurácia semelhante, mas declinou mais rápido que o modelo anterior, convergindo no mesmo ponto de queda e mantendo-se abaixo do limiar.
- 2014: a curva exibiu comportamento ainda mais drástico, alcançando seu ponto mais baixo no marco seguinte. Ao contrário dos modelos

[4] Todos possuem quatro camadas, com dez neurônios de entrada e apenas um como saída. A quantidade de neurônios das camadas escondidas é representada pelos números entre as setas.

anteriores, foi capaz de aumentar sua acurácia no marco de 2018, mas permaneceu abaixo do ponto de corte. Pode estar relacionado com uma distribuição espacial dos pontos mais uniforme em 2018 e mais confusa em 2016.

- 2016: o modelo manteve-se acima do limiar por período mais longo que o anterior e foi o único modelo a permanecer acima de 75% de acurácia até o último marco. É fator relevante que eu treinamento ocorreu com o conjunto de dados mais variado de todos, isso reforça a importância da heterogeneidade nos exemplos de treinamento.
- 2018: o resultado apresentou queda imediata sobre os dados apresentados em 2021. É possível que a mesma relação percebida em 2014 tenha ocorrido com este modelo, que foi treinado com um conjunto mais homogêneo de exemplos.

Por último, algo interessante foi observado como tendência nas matrizes de confusão dos modelos. A Tabela 4 apresenta a média harmônica dentre todos os verdadeiros positivos e verdadeiros negativos. É possível que haja uma maior facilidade para os modelos em classificar contas humanas (verdadeiros negativos) do que *bots* sociais (verdadeiros positivos), mas é preciso avaliar este fator isoladamente como hipótese, que está fora do escopo deste trabalho.

Algo semelhante foi discutido na solução proposta em [Rodríguez-Ruiz et al. 2020], que utiliza apenas uma classe para classificação de *bots* sociais, tratando a atividade de contas humanas como um padrão esperado e a atividade de *bots* como uma anomalia. Isso permite uma detecção mais rápida para contas recém-criadas e pode ser utilizado como meio de alerta para que outras abordagens de detecção possam aprofundar a investigação.

Tabela 4. Média harmônica dos verdadeiros positivos e negativos

Modelo	Verdadeiro Positivo	Verdadeiro Negativo
kNN	0,0638	0,8263
SVM	0,1729	0,8659
MLP	0,0434	0,8787

5. Considerações Finais

Existem diversos trabalhos em profusão, que buscam melhores técnicas para identificar *bots* sociais e atividades potencialmente maliciosas no uso de redes sociais. Tais atividades são muito valiosas, mas é preciso avaliar o tempo de vida útil das soluções.

Com base na análise dos resultados obtidos neste trabalho, pode-se concluir que diferentes modelos tradicionais terão o mesmo problema em com a classificação de dados no futuro e as mudanças de cenário e padrões de comportamento sempre serão um desafio. Por mais que haja grande esforço na construção de modelos ideais de detecção,

o ciclo de vida do detector como *software* não deve ser ignorado. Com o tempo, todo modelo será afetado.

A construção de conjuntos de dados com exemplos mais heterogêneos, que representam baixa linearidade, potencialmente podem melhorar seu tempo de vida. Porém, para os cenários observados, este tempo de vida não foi maior que quatro anos. Esse é um período potencialmente longo, pois esta pesquisa utiliza cenários sintéticos. Em cenário real, este tempo se reduzirá. Como esperado, o declínio da acurácia modelos se mostrou mais rápido ao longo do tempo, isto pode significar que o tempo de vida de modelos tradicionais reduziu e que tais soluções não apresentam boa durabilidade. Não são, portanto, boas escolhas para sistemas reais de detecção. Modelos atuais precisam antecipar a evolução dos *bots* sociais atuais, considerando-os como modelos ultrapassados, visto que tal a percepção é predominante em *softwares*.

Para trabalhos futuros é interessante avaliar se modelos adaptativos podem manter maior acurácia por mais tempo. Inclusive, é ponto relevante testar se as abordagens híbridas podem manter-estabilidade diante de novos dados e incluir o processamento de textos como fator adicional da avaliação de desempenho. Outros *datasets,* contendo dados do passado da RSO Twitter, também podem ajudar a construir melhores conjuntos de testes e reproduzir um cenário mais detalhado com intervalos de tempo menores que dois anos. Também é relevante avaliar a detecção a partir de contas humanas, pois os modelos mantiveram bons resultados para verdadeiros negativos

Referências

Braz, P. A. and Goldschmidt, R. R. (2018). "Redes Neurais Convolucionais na Detecção de Bots Sociais: Um Método Baseado na Clusterização de Mensagens Textuais".

Chavoshi, N., Hamooni, H. and Mueen, A. (2016). "Identifying Correlated Bots in Twitter", *Social Informatics*. Lecture Notes in Computer Science. Cham: Springer International Publishing. v. 10047p. 14–21.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine Learning*, v. 20, n. 3, p. 273–297.

Cresci, S. (2020). "A decade of social bot detection". *Communications of the ACM*, v. 63, n. 10, p. 72–83.

Cresci, S., Di Pietro, R., Petrocchi, M., Spognardi, A. and Tesconi, M. (2015). "Fame for sale: Efficient detection of fake Twitter followers". *Decision Support Systems*, v. 80, p. 56–71.

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

Cresci, S., Petrocchi, M., Spognardi, A. and Tognazzi, S. (2019). "Better Safe Than Sorry: An Adversarial Approach to Improve Social Bot Detection".

Echeverría, J., De Cristofaro, E., Kourtellis, N., et al. (2018). "LOBO: Evaluation of Generalization Deficiencies in Twitter Bot Classifiers", *Proceedings of the 34th Annual Computer Security Applications Conference*. . ACM.

Egli, A., Rosati, P., Lynn, T. and Sinclair, G. (2021). "Bad Robot: A Preliminary Exploration of the Prevalence of Automated Software Programmes and Social Bots"

Ferrara, E., Varol, O., Davis, C., Menczer, F. and Flammini, A. (2016). "The rise of social bots". *Communications of the ACM*, v. 59, n. 7, p. 96–104.

Jr, S. B., Campos, G. F. C., Tavares, G. M., et al. (2018). "Detection of Human, Legitimate Bot, and Malicious Bot in Online Social Networks Based on Wavelets". *ACM Transactions on Multimedia Computing, Communications, and Applications*, v. 14, n. 1s, p. 1–17.

Kramer, O. (2013). K-Nearest Neighbors. In: Kramer, O.[Ed.]. . *Dimensionality Reduction with Unsupervised Nearest Neighbors*. Berlin, Heidelberg: Springer Berlin Heidelberg. p. 13–23.

Latah, M. (2020). "Detection of malicious social bots: A survey and a refined taxonomy". *Expert Systems with Applications*, v. 151, p. 113383.

Popovic, D. (2000). "Intelligent Control with Neural Networks". *Soft Computing and Intelligent Systems*. Elsevier. p. 419–467.

Rauchfleisch, A. and Kaiser, J. (2020). "The False positive problem of automatic bot detection", social science research. *PLOS ONE*, v. 15, n. 10, p. e0241045.

Rodríguez-Ruiz, J., Mata-Sánchez, J. I., Monroy, R., Loyola-González, O. and López-Cuevas, A. (2020). "A one-class classification approach for bot detection on Twitter". *Computers & Security*, v. 91, p. 101715.

Samper-Escalante, L. D., Loyola-González, O., Monroy, R. and Medina-Pérez, M. A. (2021). "Bot Datasets on Twitter: Analysis and Challenges". *Applied Sciences*, v. 11, n. 9, p. 4105.

Santos, B., Ferreira G., do Ó, M, Braz, R. and Digiampietri, L. (2020). "Comparação de algoritmos para detecção de bots sociais nas eleições presidenciais no Brasil em 2018 utilizando características do usuário". *Revista Brasileira de Computação Aplicada*, v. 13, n. 1.

Santos, J., Ituassu, A., Lifschitz, S., et al. (2021). "Das milícias digitais ao comportamento coordenado: métodos interdisciplinares de análise e identificação de bots nas eleições brasileiras". *2021: ANAIS DO X BRAZILIAN WORKSHOP ON SOCIAL NETWORK ANALYSIS AND MINING*, p. 187–192.

Sayyadiharikandeh, M., Varol, O., Yang, K.-C., Flammini, A. and Menczer, F. (2020). "Detection of Novel Social Bots by Ensembles of Specialized Classifiers".

Stein, T., Chen, E. and Mangla, K. (2011). "Facebook immune system", *Proceedings of the 4th Workshop on Social Network Systems - SNS '11*. . ACM

Wang, P., Angarita, R. and Renna, I. (2018). "Is this the Era of Misinformation yet: Combining Social Bots and Fake News to Deceive the Masses", *Companion of the The Web Conference 2018 on The Web Conference 2018 - WWW '18*. . ACM Press.

Yang, K.-C., Varol, O., Hui, P.-M. and Menczer, F. (2020). "Scalable and Generalizable Social Bot Detection through Data Selection", *Proceedings of the AAAI Conference on Artificial Intelligence*, v. 34, n. 01, p. 1096–1103.