DETEÇÃO DE BOTS NAS REDES SOCIAIS: MÉTODOS, DESAFIOS E EVOLUÇÃO

David Franco (110733); damfo1@iscte-iul.pt
Diogo Aqueu (110705); drcau@iscte-iul.pt
João Dias (110305); jprds@iscte-iul.pt
Rafael Cerqueira (110860); rrcao2@iscte-iul.pt

Afiliação dos Autores
Iscte-IUL
Av. Forças Armadas, 1649-026 Lisboa, Portugal

RESUMO

O artigo de divulgação visa abordar a evolução dos mecanismos de deteção de *bots* nas redes sociais de uma perspetiva metodológica, com principal ênfase em *botnets* e em *social bots*, discutindo as suas principais implicações e consequências na manipulação da opinião pública. São apresentados exemplos relacionados à COVID-19 e à guerra entre Rússia-Ucrânia para explicar o impacto destes automatismos. As técnicas de deteção existentes incluem métodos de aprendizagem automática supervisionada e não supervisionada, todavia ainda são apontados inúmeros desafios, como a falta de estudos em mais redes sociais além do Twitter, a natureza reativa dos modelos atuais, que faz com que estejam sempre um passo atrás dos *bots* e, por fim, a fraca capacidade de adaptação dos algoritmos às constantes mutações dos *social bots*. O objetivo deste artigo de divulgação foi realizar um levantamento metodológico e histórico sobre a deteção de *bots* nas redes sociais e enfatizar a importância de desenvolver mecanismos mais robustos e complexos para detetar *bots*, a fim de prevenir possíveis ataques maliciosos.

Palavras-chave **—** deteção de *bots*, redes sociais, cibersegurança, aprendizagem automática, *social bots*, *botnets*

1. INTRODUÇÃO

Numa era cada vez mais digital, as redes sociais têm assumido um papel preponderante no que respeita à disponibilização de informação [4]. A par e passo, o crescimento do número de utilizadores destas plataformas tem-se tornando cada vez mais evidente. Só durante o ano de 2022 foi registado um aumento de 137 milhões no número de utilizadores ativos destas plataformas [6]. O crescimento da popularidade destes recursos, quando combinado com as vastas informações partilhadas pelos utilizadores, vem despertar o interesse de pessoas mal-intencionadas [1]. Assim, a forma mais predominante de *malware* conhecida nas redes sociais pode ser associada aos *bots* [3]. Embora se reconheça a existência de *bots* benignos, as atenções centram-se nos que são utilizados para a elaboração de atividades maliciosas como o desenvolvimento de contas e interações falsas, *phishing*, *spaming* e manipulação da opinião pública através de boatos e rumores que são espalhados nestas plataformas digitais [4]. Como tal, os últimos anos têm sido assinalados por desenvolvimentos crescentes na investigação de mecanismos de deteção deste tipo de *malware*, que recorrem a técnicas de aprendizagem automática supervisionada e não supervisionada [8].

2. TIPOS DE *BOTS* NAS REDES SOCIAIS

O termo "*bot*", da palavra "*robot*", refere-se a um programa de computador capaz de automatizar tarefas na internet [11]. Por outras palavras, trata-se de um programa criado para repetir tarefas, de forma estruturada, a uma velocidade muito mais elevada do que a dos humanos [11].

2.1. *Social bots* e *botnets*

No contexto de redes sociais, os *bots* podem ser categorizados como *social bots*, quando se tratam de programas únicos com um objetivo concreto de aplicação, ou como *botnets*, quando se tratam de redes de *bots* que atuam entre si para alcançar um determinado objetivo [16].

Os *social bots* procuram reproduzir comportamentos humanos nas redes sociais e podem assumir inúmeros formatos [11]. De acordo com Khaund [11], os mais populares podem ser categorizados consoante o seu propósito, destacando-se os *bots* de:

1. Propaganda - responsáveis por influenciar opiniões políticas através do afastamento de opiniões divergentes.
2. Influência - desenhados para influenciar comportamentos ou opiniões.
3. Promocionais - desenvolvidos por áreas como a de marketing digital para melhorar a experiência do utilizador.
4. *Spam* – caracterizados pelo envio de *spam* aos restantes utilizadores.
5. *Hackers* – desenhados para distribuir *malware*, atacar websites ou redes.

6. Conversação – desenvolvidos através de métodos de texto e/ou áudio para conduzir conversas com os utilizadores.
7. Notícias – responsáveis por facilitar a disseminação de notícias.

Por sua vez, as *botnets* apresentam-se como uma das maiores ameaças à cibersegurança por permitirem a expansão dos ataques dos *social bots* para campanhas de *spam*, espionagem, *phishing* ou *ransomware* a larga escala [16]. A figura 2.1 ilustra o diagrama de funcionamento de uma *botnet* que lança um ataque para o preenchimento e obtenção de credenciais de um determinado utilizador. Contudo, a sua forma de funcionamento pode ser generalizada para diferentes aplicações e objetivos.

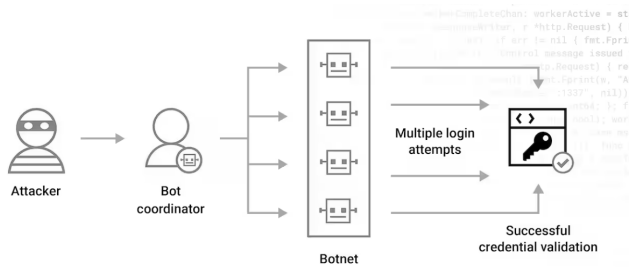

Figura 2.1: Diagrama de funcionamento de um ataque de uma *botnet* [2]

3. PRINCIPAIS IMPLICAÇÕES

Existem inúmeros utilizadores que recorrem às plataformas digitais como fonte primária de acesso à informação, fazendo com que os *social bots* possam contribuir para a manipulação da opinião pública [3]. Tomemos alguns exemplos práticos para consolidar as reais implicações deste problema.

3.1. O caso da COVID-19

A pandemia de COVID-19 ficou marcada pelo termo "infodemia", devido ao aumento de *social bots* que participavam ativamente no debate sobre as medidas de saúde para o controlo pandémico e outros tópicos relacionados durante este período [15]. A complexidade do ecossistema das plataformas sociais, aliado aos *bots*, contribuiu para a onda de desinformação, que permitiu a amplificação de mensagens de qualidade dúbia, ou polarizou opiniões em temas controversos, como as vacinas [9]. Consequentemente, assistiu-se a um aumento significativo do número de casos de infetados pelo coronavírus [9]. Ainda assim, as implicações dos *bots* não se ficam pelo setor da saúde, única e exclusivamente.

3.2. O caso da guerra entre a Rússia e a Ucrânia

As ferramentas para a discussão política mudaram radicalmente com a crescente popularidade das redes sociais, que apesar de terem acelerado processos e metodologias para o debate de ideias, trouxeram também alguns perigos para a opinião pública, especialmente quando se colocam em perspetiva os *social bots* [7]. Contudo, é importante considerar que estes também podem desempenhar um papel chave para gerar *awareness* em torno de um determinado problema político [7].

No caso específico da guerra entre a Rússia-Ucrânia, verificou-se uma posição política clara nos *tweets* de *bots* [14]. A percentagem de publicações no Twitter de *bots* a favor do lado ucraniano era quase nove vezes maior do que a do lado russo [14]. Os *social bots* amplificaram a voz do lado ucraniano nas conversas da plataforma digital, através da condenação dos atos da Rússia e da manifestação de apoio à Ucrânia [14].

Do lado russo, os *tweets* de *bots* que publicavam com mais frequência focavam-se essencialmente em conteúdo controverso como a expansão da NATO para leste [14]. Além disso, verificaram-se poucas contas de *bots*, identificadas como provenientes do país, o que pode ser explicado pela suspensão levada a cabo pelo Twitter de *bots* russos após o início da guerra, ou pelo bloqueio imposto pelo governo russo aos cidadãos, que os impedia de aceder à rede social [14].

4. TÉCNICAS PARA A DETEÇÃO DE *BOTS*

4.1. Aprendizagem automática supervisionada

O primeiro trabalho desenvolvido especificamente para a deteção de contas automáticas em redes sociais surgiu em janeiro de 2010 [5]. Desde então, o tema tem registado um crescimento progressivo, conforme se pode ver na figura 4.1.

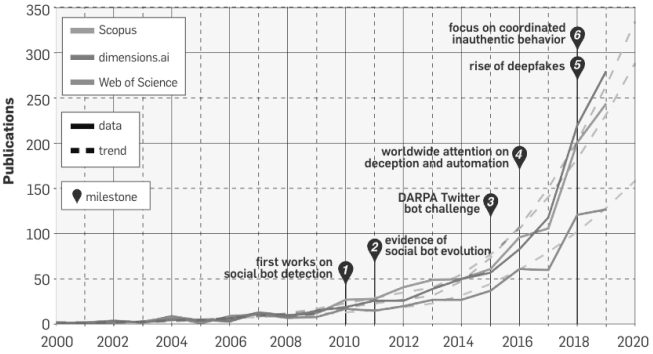

Figura 4.1: Publicações por ano sobre a caracterização, deteção e impacto estimado de *social bots* [5]

Os primeiros modelos para a deteção de *bots* baseavam-se na ideia de que os *bots* e os humanos eram facilmente separáveis em dois grupos distintos com características próprias [5]. Dessa forma, recolhiam-se os dados através das APIs das

respetivas redes sociais, com identificadores de *bots* e de utilizadores legítimos, e fazia-se o seu pré-processamento para depois treinar um modelo de classificação que viria a separar os utilizadores através de um rótulo binário de *bot*/não *bot* [8]. Ainda que seja a particularização de um caso específico de um modelo de aprendizagem profunda, a figura 4.2 pode ser generalizada para a ilustração do *workflow* dos mecanismos de aprendizagem automática supervisionada.

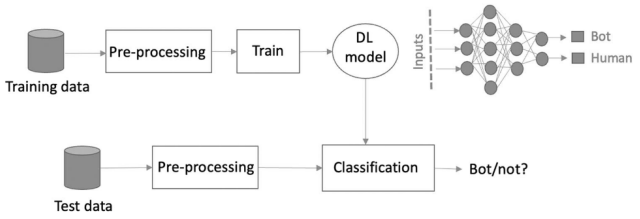

Figura 4.2: *Workflow* generalista de um modelo de classificação de um detetor de *bots* [8]

As técnicas de aprendizagem supervisionada continuam a ser as mais predominantes na atualidade, no que diz respeito à investigação sobre a deteção de *bots* [17]. Os algoritmos de classificação utilizam características como a taxa de publicação de URLs em publicações, fatores relacionados com as taxas de interação, a dimensão do nome dos utilizadores, o rácio de seguidores-amigos e algumas análises de sentimento mais detalhadas para identificar *bots* nas redes sociais [17]. As aplicações de [13] mostraram que os *social bots*, independentemente do seu propósito, tendem a apresentar um comportamento menos diversificado do que o de utilizadores legítimos.

A análise comparativa dos diferentes algoritmos mostra que as *Random Forest* têm o melhor desempenho no que diz respeito à *accuracy* para identificar *social bots*, apresentando também o maior número de aplicações em *papers* [4]. A figura 4.3 ilustra os algoritmos mais aplicados por investigadores neste contexto.

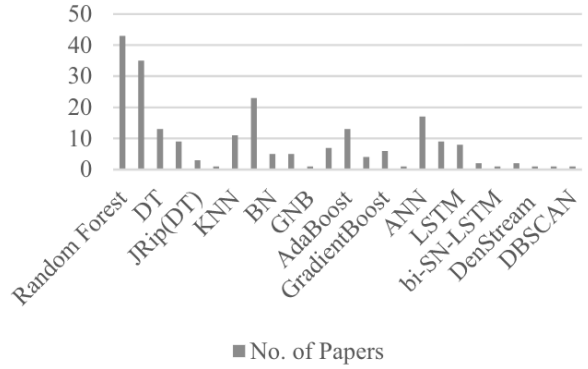

Figura 4.3: Principais algoritmos aplicados em *papers* sobre a deteção de *bots* em redes sociais [4]

Ainda que os métodos supervisionados tenham mostrado resultados muito promissores nos estágios iniciais de desenvolvimento e embora continuem a dominar a investigação na atualidade, é importante realçar os desafios e problemas que têm associados [5]. O primeiro está relacionado com o *dataset* utilizado para a validação dos dados previstos pelo modelo, que depende, em grande parte dos casos, de rótulos (*bot*/não *bot*) atribuídos por operadores humanos que procedem a uma análise manual, limitada e enviesada [12]. Além disso, a produção de contramedidas por parte de quem desenvolve *social bots* tem vindo a dificultar esta tarefa de classificação [5]. Atualmente, os *bots* são planeados de forma a aumentar a sua credibilidade, dispondo de fotografias de perfil, nomes e informações roubadas, vários seguidores reais e onde as mensagens maliciosas passam despercebidas entre inúmeras neutras, que são espalhadas através das *botnets*, o que dificulta a tarefa de os distinguir de utilizadores legítimos [5].

4.2. Aprendizagem automática não supervisionada

Os métodos de aprendizagem automática não supervisionados que utilizam os dados da cronologia de uma determinada rede social podem oferecer uma performance comparável ou até melhor que os métodos supervisionados, com menos enviesamento por parte de operadores humanos e com um custo de complexidade mais reduzido [17]. É importante destacar os resultados de [17] que confirmaram a possibilidade de criar *clusters* entre contas legítimas e de *bots*, através dos algoritmos *K-means* e *Agglomerative clustering*, que se mostraram capazes de manter a performance funcional do programa, sem ter de recorrer a enviesamentos humanos na introdução de rótulos na informação para a validação das previsões do modelo. Ainda assim, poucos investigadores optam por seguir a via não supervisionada na abordagem ao problema da deteção de *social bots* [4]. A figura 4.4 ilustra o domínio dos métodos supervisionados desde 2010, no que diz respeito ao número de novos mecanismos de deteção de *bots*.

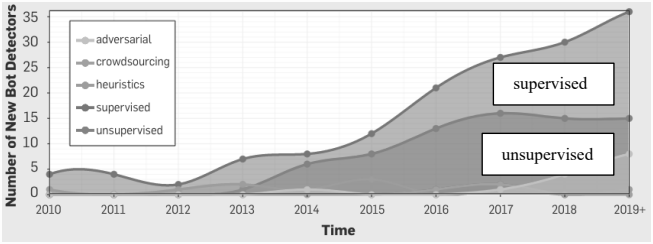

Figura 4.4: Número de novos detetores de bots [5]

4.3. Incorporação de mecanismos de deteção em ferramentas de escuta ativa

A escuta ativa é uma prática de monitorização dos canais das redes sociais para menções a uma determinada marca, a um concorrente ou a *keywords* relacionadas [10]. Distingue-se da monitorização dos canais nestas plataformas, por ter uma abordagem mais proativa e menos reativa [10].

De acordo com uma entrevista realizada a João Santos Silva, consultor de Assuntos Públicos da empresa LLYC, a escuta ativa nas redes sociais em torno de determinado tema é sem dúvida essencial para identificar mais eficazmente os *pain points* sobre o qual surge o conflito e os principais *opinion makers* que se manifestam sobre o tema, auxiliando na antecipação de possíveis impactos negativos e contribuindo, desta forma, para a elaboração de uma melhor estratégia de assuntos públicos. Porém, dado o potencial uso pernicioso da Inteligência Artificial na produção de conteúdos massificados para fins de manipulação, como se viu nas eleições norte-americanas de 2016, onde um quinto de todos os tweets sobre as eleições presidenciais de 2016 foram publicados por *bots*, será necessário que as plataformas de escuta ativa - assim como as próprias redes sociais - sinalizem rápida e eficazmente quais os conteúdos produzidos por *bots*. Esta identificação é determinante não apenas para a área de assuntos públicos, por forma a evitar a formulação de pressupostos alheios à realidade e às preocupações reais das pessoas, mas também para as nossas sociedades e democracias, para garantir que indivíduos, grupos e comunidades não são feitos reféns de narrativas infundadas, irreais e polarizantes.

Durante a entrevista, o consultor destacou que a equipa de *Deep Learning* da empresa recorre a uma ferramenta de escuta ativa – Brandwatch - para se inteirar das conversas que decorrem no espaço social do Twitter, mapear comunidades e identificar potenciais *Key Opinion Leaders* (KOLs). Para filtrar o ruído da rede social, a plataforma disponibiliza filtros de spam nas configurações de cada projeto e ferramentas de "*Bot Detection*" para a identificação de contas maliciosas.

5. DESAFIOS E OPORTUNIDADES

A maior parte da investigação conduzida nesta área é feita na rede social Twitter, conforme se pode ver na figura 5.1, o que se justifica pela facilidade na recolha dos dados através da API da plataforma e pelas várias coleções de *datasets* públicos disponíveis. A escassez de estudos sobre as restantes redes sociais deve-se em grande parte às políticas rigorosas de privacidade em vigor nas plataformas ou ao facto de serem muito recentes, como é exemplo o TikTok [4].

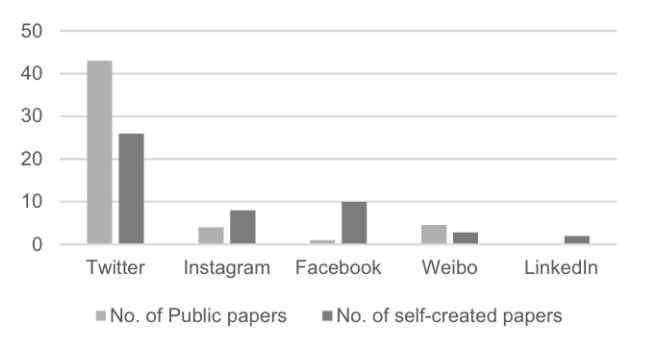

Figura 5.1: Distribuição de *datasets* por plataforma [4]

Além disso, a generalidade dos mecanismos de deteção de *bots* atuais segue um esquema reativo, estando por isso sempre um passo atrás das contas maliciosas desenvolvidas, o que explica, parcialmente, o facto da influência de *bots* e outros atores nas discussões *online* não ter diminuído, mesmo após ter-se verificado um aumento do número de técnicas de deteção [5].

Por fim, é importante realçar que a classificação de *bots* não é uma tarefa estática, já que a sua evolução faz com que estes exibam comportamentos e características diferentes ao longo do tempo. Além disso, os algoritmos não podem estar assentes no pressuposto da neutralidade, que é constantemente violado pelos criadores de *bots* na sua procura ativa para enganar os sistemas [5].

Neste contexto, o desenvolvimento de novos mecanismos de deteção de *bots* tem centrado o paradigma atual da investigação no estudo das vulnerabilidades dos sistemas para a deteção precoce e prevenção de possíveis ataques, o que contribui para o desenvolvimento de sistemas mais robustos [8].

6. PRINCIPAIS CONCLUSÕES

O artigo de divulgação procurou fornecer, de forma aprofundada, a categorização dos diferentes tipos de *bots* que podem surgir nas redes sociais, focando-se essencialmente nos *social bots* e nas *botnets*, discutindo as principais implicações que estes/as podem ter na manipulação da opinião pública, com recurso a casos práticos associados à COVID-19 ou à guerra entre a Rússia-Ucrânia.

De seguida, destacaram-se as principais técnicas de deteção e os respetivos *frameworks* gerais de utilização. Apesar das limitações, continuam a ser os métodos de aprendizagem automática supervisionada que dominam a investigação nos sistemas de deteção de *bots*.

Por fim, são elencados alguns dos principais desafios e oportunidades nesta área, que se prendem essencialmente com o maior volume de investigação realizado na rede Twitter. No caso concreto de Portugal, esta plataforma não apresenta o mesmo peso em termos de audiências que o Instagram, o Facebook ou o TikTok [6]. A par e passo,

destaca-se também o esquema reativo dos modelos de deteção atuais e os pressupostos estáticos de neutralidade, que não permitem a adaptação dinâmica dos algoritmos às constantes mutações que os *social bots* vão sofrendo.

AGRADECIMENTOS

Gostaríamos de expressar o nosso sincero obrigado ao consultor de Assuntos Públicos da LLYC, João Santos Silva, pelos conhecimentos e insights transmitidos, que foram fundamentais para a construção do artigo de divulgação.

REFERÊNCIAS

[1] Adikari, S., & Dutta, K. (2020). Identifying Fake Profiles in LinkedIn. *Proceedings - Pacific Asia Conference on Information Systems*. https://arxiv.org/abs/2006.01381v1

[2] Akamai. (2023). *O que é um botnet?* https://www.akamai.com/pt/glossary/what-is-a-botnet

[3] Aldayel, A., & Magdy, W. (2022). Characterizing the role of bots' in polarized stance on social media. *Social Network Analysis and Mining*, *12*(1). https://doi.org/10.1007/S13278-022-00858-Z

[4] Aljabri, M., Zagrouba, R., Shaahid, A., Alnasser, F., Saleh, A., & Alomari, D. M. (2023). Machine learning-based social media bot detection: a comprehensive literature review. *Social Network Analysis and Mining*, *13*(1), 1–40. https://doi.org/10.1007/S13278-022-01020-5

[5] Cresci, S. (2020). A decade of social bot detection. *Communications of the ACM*, *63*(10), 72–83. https://doi.org/10.1145/3409116

[6] DataReportal. (2023, January). *Digital 2023: Global Overview Report*. https://datareportal.com/reports/digital-2023-global-overview-report

[7] Graells-Garrido, E., & Baeza-Yates, R. (2022). Bots don't Vote, but They Surely Bother!: A Study of Anomalous Accounts in a National Referendum. *ACM International Conference Proceeding Series*, 302–306. https://doi.org/10.1145/3501247.3531576

[8] Hayawi, K., Saha, S., Masud, M. M., Mathew, S. S., & Kaosar, M. (2023). Social media bot detection with deep learning methods: a systematic review. *Neural Computing and Applications*, *35*(12), 8903–8918. https://doi.org/10.1007/S00521-023-08352-Z/FIGURES/4

[9] Himelein-Wachowiak, M., Giorgi, S., Devoto, A., Rahman, M., Ungar, L., Schwartz, H. A., Epstein, D. H., Leggio, L., & Curtis, B. (2021). Bots and Misinformation Spread on Social Media: Implications for COVID-19. *Journal of Medical Internet Research*, *23*(5). https://doi.org/10.2196/26933

[10] Hootsuite. (2022, December 1). *What is Social Listening, Why it Matters*. https://blog.hootsuite.com/social-listening-business/

[11] Khaund, T., Kirdemir, B., Agarwal, N., Liu, H., & Morstatter, F. (2022). Social Bots and Their Coordination During Online Campaigns: A Survey. *IEEE Transactions on Computational Social Systems*, *9*(2), 530–545. https://doi.org/10.1109/TCSS.2021.3103515

[12] Kolomeets, M., & Chechulin, A. (2023). *Social bot metrics*. *13*, 36. https://doi.org/10.1007/s13278-023-01038-3

[13] Kosmajac, D., & Keselj, V. (2019). Twitter Bot Detection using Diversity Measures. *Proceedings of the 3rd International Conference on Natural Language and Speech Processing*, 1–8. https://aclanthology.org/W19-7401

[14] Shen, F., Zhang, E., Zhang, H., Ren, W., Jia, Q., & He, Y. (2023). Examining the differences between human and bot social media accounts: A case study of the Russia-Ukraine War. *First Monday*. https://doi.org/10.5210/FM.V28I2.12777

[15] Suarez-Lledo, V., & Alvarez-Galvez, J. (2022). Assessing the Role of Social Bots During the COVID-19 Pandemic: Infodemic, Disagreement, and Criticism. *Journal of Medical Internet Research*, *24*(8). https://doi.org/10.2196/36085

[16] Velasco-Mata, J., González-Castro, V., Fidalgo, E., & Alegre, E. (2023). Real-time botnet detection on large network bandwidths using machine learning. *Scientific Reports*, *13*(1). https://doi.org/10.1038/S41598-023-31260-0

[17] Wu, J., Teng, E., & Cao, Z. (2022). Twitter Bot Detection Through Unsupervised Machine Learning. *Proceedings - 2022 IEEE International Conference on Big Data, Big Data 2022*, 5833–5839. https://doi.org/10.1109/BIGDATA55660.2022.10020983