Utilização de técnicas de Processamento de Linguagem Natural e Aprendizado de Máquina na análise do uso de *bots* no *Instagram*

Kristiano Pasini de Oliveira[1], **Cinthyan Renata Sachs Camerlengo de Barbosa**[1]

[1]Departamento de Computação – Universidade Estadual de Londrina (UEL)
Caixa Postal 10.011 – CEP 86057-970 – Londrina – PR – Brasil

kristiano.pasini@uel.br, cinthyan@uel.br

Abstract. *Social networks are platforms where millions of people communicate on a daily basis, and that number grows more each year. Financial and ideological gains that profiles and opinions with high popularity gain encourages the development of many strategies get this status. One of those are called bots, which consists of computer programs responsible for simulating common user interactions in an automated way and by doing so generates artificial metrics that manipulate the natural organization of the network. Such practice generates a lot of issues: from the maintenance of the platform by the responsible company to even negative impacts in politics and public health. Because of that, it is critical to find ways to restrict this kind of tools. In this context, this paper proposes the use of Machine Learning algorithms that may help to identify the use of bots, under the premise that human users and automated behavior can be told apart based on comments and platform data analysis. Several data will be collected, both profile data, such as follower number and followed number and comment content across posts. That information will be analyzed with Natural Language Processing techniques and to train a Machine Learning model to verify if such approaches generate good results for this type of problem.*

Resumo. *As redes sociais são plataformas nas quais milhões de usuários se comunicam diariamente e esse número vem aumentando a cada ano. Os ganhos financeiros e ideológicos garantidos por perfis e opiniões com grande popularidade fomenta o desenvolvimento de diversas técnicas para alcançar tal projeção. Uma delas é o uso de ferramentas bots, programas que automatizam interações comuns de usuários a fim de gerar métricas artificiais que manipulem a organização natural da rede. Em virtude dos malefícios causa- dos com essa prática, desde a dificuldade de manutenção dos próprios serviços até as possíveis influências negativas na política e na saúde pública, é mister a investigação de soluções que restrinjam tais ferramentas. Nesse sentido, o seguinte trabalho tem como proposta classificar um perfil da plataforma Ins- tagram como bot ou não, com base em atividades na rede e sob a hipótese de que o comportamento de um usuário pode ser diferenciado de um programa por meio da análise de comentários e dados da plataforma. Para isso serão coletados do Instagram tanto métricas da conta como número de seguidos e seguidores quanto os comentários realizados em postagens. Sendo então aplicadas técnicas de Processamento de Linguagem Natural e Aprendizado de Máquina para analisar se tais abordagens geram bons resultado para esse tipo de problema.*

1. Introdução

Nos últimos anos as redes sociais vêm cada vez mais alcançando um maior número de usuários. Dentre as diversas existentes pode-se explicitar as plataformas *Facebook*, *Twitter* e *Instagram* dada a grande popularidade dessas.

O aplicativo *Instagram*, desenvolvido e mantido pela *Meta Platforms*, segundo a própria empresa, é um aplicativo gratuito de compartilhamento de fotos e vídeos [1]. Conforme os relatórios financeiros disponibilizados pela empresa, os aplicativos do grupo família, que incluem *Facebook*, *Whatsapp* e *Instagram* ultrapassaram a margem dos três bilhões de usuários diários com um aumento de em média 5% ao ano [2].

Sendo assim, é clara a grande interação entre usuários que ocorre diariamente por meio da plataforma. Nesse contexto, devido à visibilidade tanto ideológica quanto monetária gerada por perfis que atraiam o interesse de uma grande quantidade de usuários, diversas abordagens são constantemente desenvolvidas a fim de obter maior sucesso e alcance. Entre elas podemos destacar o desenvolvimento de *bots*, programas capazes de gerar conteúdo de forma automatizada e interagir com outros usuários, simulando as atitudes de um humano [3].

Essa prática é contra as diretrizes da comunidade do *Instagram*, visto que esse conteúdo artificial cria uma experiência negativa e prejudica a capacidade das pessoas de interagirem de forma autêntica, além de também ameaçar a segurança, a estabilidade e a usabilidade dos serviços [4]. Nota-se também que tais atividades podem ser utilizadas de forma maliciosa em assuntos políticos e de saúde pública.

Ruediger [5] afirma que os *bots* sociais podem produzir opiniões artificiais, além de poder gerar dimensões irreais para essas ou determinadas fíguras públicas, destacando também a existência de certa preocupação a respeito da propagação de notícias falsas ou de campanhas de poluição da rede, ofuscando debates com informações irrelevantes às discussões levantadas. Esse problema é ainda mais evidenciado pelo fato de tais robôs estarem envolvidos em um número notável de interações pela internet, visto que um estudo feito em 2018 pelo *Ghost Data* apontou que aproximadamente 95 milhoes de contas do *Instagram* eram automatizadas e que em 2016, *bots* geraram mais tráfego na internet do que os próprios usuários comuns [6].

Tendo em vista esse cenário, o presente trabalho busca, por meio do uso de técnicas de Processamento de Linguagem Natural e Aprendizado de Máquina, desenvolver um algoritmo capaz de análisar comentários e dados da plataforma *Instagram* e classificar um perfil como *bot* ou não. Tal objetivo se baseia na hipótese de que o compartamento de um humano difere do de um *bot* e tais diferenças podem ser verificadas por meio da análise de métricas de perfis e atividade na plataforma.

Este projeto está organizado da seguinte forma: A Seção 2 apresenta trabalhos encontrados acerca da verificações de *bots* em redes sociais, principalmente na plataforma do *Twitter* como partida para o estudo no *Instagram*, explicitando alguns métodos de Processamento de Linguagem Natural (PLN), Aprendizado de Máquina (AM) e coleta de dados de tais pesquisas. A Seção 3 descreve os objetivos que esperam ser alcançado com este trabalho. A Seção 4 detalha quais foram os procedimentos metodológicos aplicados e é separada em quatro diferentes tópicos sequênciais, contendo a revisão bibliográfica, a coleta e a análise exploratória dos dados, a aplicação dos algoritmos de AM e a avaliação

dos resultados obtidos por meio do cálculo de diferentes métricas, respectivamente. A Seção 5 apresenta o cronograma esperado para a execução das atividades propostas na Seção 4.

2. Fundamentação Teórico-Metodológica e Estado da Arte

Existem diversos estudos sobre a identificação de *bots* em redes sociais. O *Twitter*, por exemplo, por disponibilizar até o início de 2023 o acesso a sua API de forma gratuita, permitindo a coleta de um grande volume de dados para análise, garantiu a criação de múltiplos *datasets* para pesquisas na área.

Ferrara e Kudugunta [7] demonstram a utilização de uma rede neural profunda na detecção de perfis automatizados. Nesse trabalho foram coletados tanto os dados da plataforma quantos os conteúdo dos *tweets* na classificação dos perfis. Como resultado foi demonstrado uma acurácia de 96% ao análisar apenas mensagens avulsas e de 99% ao incluir análises em nível de perfil. Tais resultados indicam a eficiência de abordagens de Aprendizado de Máquina para a resolução desse problema.

Rodríguez-Ruiz *et al.* [8] discute as diferentes abordagens utilizadas na detecção de *bots* no *Twitter*, destacando o uso tanto de algoritmos supervisionados quanto os não supervisionados, sendo o primeiro o mais utilizado. O autor também pondera que, apesar da maior parte dos mecanismos propostos terem obtidos resultados satisfatórios, esses requerem dados rotulados de perfis automatizados para se obter padrões caractéristicos. Além disso, destaca a heterogeneidade desse tipo de ferramenta e a necessidade de métodos de detecção que não se baseiem unicamente em *bots* já existentes.

A detecção de *bots* também pode ser realizada sob uma perspectiva de classificação de classe única [8]. Tal metodologia busca analisar apenas dados de usuários legítimos com a finalidade de se encontrar comportamentos deviantes. Os re- sultados obtidos apontaram uma acurácia acima de 90%. Com tais resultados, Rodríguez- Ruiz *et al.* [8] conclui que classificadores desse tipo podem servir como uma forma inicial de detecção, apontando padrões de possíveis novos *bots*.

Miller *et al.* [9] explora dois algoritmos de aprendizado não supervisionado para a detecção de mensagens de spam. São eles o *Density-based clustering* e o *K-means based clustering*, apresentando métricas obtidas para cada um. Espera-se que tal trabalho permita uma melhor definição dos algoritmos a serem utilizados e da definição da utilização de uma abordagem supervisionada ou não supervisionada.

Alothali *et al.* [3] cita diferentes métodos de detecção de *bots* em redes sociais. Entre elas, pode-se destacar o *crowdsourcing* que consiste na utilização da inteligência humana para diferenciar contas falsas de usuários comuns. Tal abordagem envolve diferentes critérios, de acordo com as métricas escolhidas para justificar a classificação dada. O *crowdsourcing* será utilizado para a criação do *dataset* deste trabalho.

Um exemplo de ferramenta disponível *online* para a verificação de perfis em tempo real é a PEGABOT[1], feita pelo Instituto do Tecnologia e Sociedade do Rio de Janeiro (ITS Rio) em parceria com o Instituto Equidade e Tecnologia com o objetivo de se classificar perfis do *Twitter* com base na probabilidade de serem *bots*. No entanto, com a mudança

[1] https://pegabot.com.br/

dos planos gratuitos da API do *Twitter* permitindo apenas a postagem e não mais a leitura de conteúdo, a plataforma teve seus serviços interrompidos conforme comunicado no próprio perfil da ferramenta.

Miranda [10] realizou a aplicação de Processamento de Linguagem Natural e Aprendizado de Máquina em dados coletados da atividade de perfis no *Twitter*, desenvolvendo uma análise da relevância de variáveis morfológicas no processo de detecção de *bots*. Utilizou de diversas ferramentas como o próprio PEGABOT, o algoritmo de Aprendizado de Máquina chamado *Naive Bayes* e algumas bibliotecas para Processamento de Linguagem Natural como o pacote spaCy[2] e o LeIA[3] (*Léxico para Inferência Adaptada*). Espera-se que os resultados obtidos e as ferramentas utilizadas auxiliem na metodologia deste trabalho.

O uso de PLN como ferramenta para a solução de problemas em redes sociais pode envolver também outras temáticas. Mioni [11] aplicou PLN para detectar comporta- mento tóxico no *Twitter*, descrevendo cada uma das etapas seguidas na análise dos *tweets*. A base de dados utilizada foi um *corpus* contendo 450 tweets capturados enquanto um episódio do programa *MasterChef* Brasil era exibido. Dentre os resultados do trabalho espera-se que o estudo de processamento da língua portuguesa em conteúdos inseridos por usuários em plataformas como redes sociais e o processo de análise de dados reais da *web* contribuam com a coleta e o processamento dos comentários no *Instagram*.

No caso do *Instagram*, Akyon e Kalfaoglu [6] apresentam algumas diretrizes para a análise de perfis na perspectiva de verificação de contas falsas. Características como a ausência de foto de perfil, nomes de usuários incomuns e confusos, com número elevado de números, por exemplo, são grande indicativos para *bots*. Alto número de seguidos e baixo de seguidores também se enquadram no grupo, visto que o principal objetivo de tais programas é a manipulação de métricas da plataforma. A apresentação da importância dos dados da plataforma e a abordagem de coleta de perfis descritos pelos autores influenciou diretamente na formulação do processo de criação do *dataset* deste trabalho.

A plataforma *Instagram* foi escolhida como fonte de dados para a análise deste trabalho, apesar de impor mais barreiras ao acesso de sua API. Tal escolha se deu com base nas mudanças de acesso a API do *Twitter*, menor foco em exploração de múltiplas atividades de um único perfil e a maior familiaridade do autor com a primeira.

3. Objetivos

O presente trabalho tem como objetivo desenvolver um algoritmo capaz de classificar perfis da rede social *Instagram* como *bots* ou não com base em comentários de postagens e dados da plataforma.

4. Procedimentos metodológicos

Os procedimentos adotados para a execução deste trabalho consistem em quatro principais atividades: a revisão bibliográfica; a coleta dos dados e análise exploratória; a aplicação dos algoritmos; e finalmente o cálculo das métricas e análise dos resultados.

[2] https://spacy.io/

[3] https://github.com/rafjaa/

4.1. Revisão Bibliográfica

Para a revisão bibliográfica serão buscadas publicações científicas que envolvam a aplicação de diferentes técnicas de Apredizado de Máquina e de Processamento de Linguagem Natural para a detecção de *bots* em redes sociais. Tal conhecimento será importante para guiar as escolhas realizadas neste trabalho, permitindo o entendimento sobre os conceitos e algoritmos utilizados com base em resultados obtidos por pesquisas já rea- lizadas.

4.2. Coleta dos dados e Análise exploratória

A coleta dos dados será feita por meio da recuperação manual de perfis e comentários em *posts* com diferentes temáticas, por meio da ferramenta de exploração do *Instagram*. Será realizada a análise tanto de perfis pessoais quanto de negócio, com popularidade considerável (mais de 100 comentários por postagem). Hashtags e conteúdos de comentário que referenciem usuários da plataforma serão tratados a fim de manter a privacidade. Além do texto puro será coletada também a popularidade de cada comentário (quantidade de curtidas). Para o conjunto de dados da plataforma serão coletadas informações que espera-se que apresentem influências na detecção de *bots*. O número de seguidores e seguidos, por exemplo, podem indicar comportamentos automatizados visto que estão diretamente relacionados à busca de elevação da popularidade de outros usuários [6], e, portanto, podem ser boas métricas a nível de perfil.

Além disso, serão incluídos dados como a quantidade de dígitos no *username*, idade da conta, existência de foto, possíveis comportamentos que podem representar um baixo nível de detalhamento e baixa individualidade, diretamente relacionados ao processo de geração de contas falsas. Conforme Akyon e Kalfaoglu [6], a métrica de quantidade de dígitos no nome do usuário é justificada pelo fato de que mais de 50% das contas *bots* possuem tal característica, ao passo que para usuários reais esse valor cai para 11%. Como último critério será analisado se o perfil é privado e, em caso negativo, se existem postagens (campos binários).

Para o processamento dos dados textuais coletados será utilizado a biblioteca spaCy. O spaCy é uma biblioteca de Processamento de Linguagem Natural para análise sintática que pode ser escrita nas linguagens Python e Cython. Suporta a Língua Portuguesa e possuí código aberto, ou seja, disponibiliza de forma gratuita o *download* dos códigos fonte [12]. As etapas realizadas serão a tokenização dos comentários, a determinação das categorias gramáticas dos tokens e a remoção das *stopwords*. Após o tratamento do texto, serão análisados aspectos como termos populares, quantidade de palavras, uso de verbos, adjetivos, pronomes e demais classes apontadas pelo spaCy. Também serão verificados usos de *hashtags e emojis*. Ao analisar dados textuais no *Twitter*, Miranda [10] aponta o maior uso de *hashtags* e *emojis* por perfis *bots*, sendo o último o mais significativo de acordo com os resultados encontrados. Dentre os trinta *tokens* mais utilizados por perfis classificados como *bot*, oito eram emojis, ao passo que esse valor caía para três para usuários reais.

A rotulação dos perfis como *bots* será feita pela atribuição de um grau de automatização pelo processo de *crowdsourcing*, no qual os dados coletados serão apre- sentados a diferentes equipes que avaliarão perfil, a fim de obter uma pontuação final após a união das diferentes perspectivas, sendo descritos quais os critérios utilizados

em cada avaliação. Tal abordagem se baseia na possibilidade de aplicação da análise humana para esse problema, pois como é destacado por Alothali *et al.* [3] essa metodologia apresenta boa acurácia dos resultados, apesar de demandar um maior tempo. Nesse contexto, o autor cita como exemplo uma competição de verificação de *bots* no *Twitter* organizada pela DARPA em 2015 na qual times submetiam seus chutes na verificação de perfis envolvidos em discussões de suporte às vacinações. Três times dentre os seis participantes tiveram pontuação máxima em relação à habilidade de verificação de contas automatizadas.

Com o *dataset* formado serão desenvolvidos gráficos para verificar as relações entre as variáveis com o intuito de que sejam apresentadas ao modelo somente as consideradas relevantes, ou seja, que indicarem ter alguma influência sobre os resultados.

4.3. Aplicação dos algoritmos

Inicialmente espera-se garantir a utilização de ambos métodos supervisionados quanto não supervisionados, dado que o uso de vários tipos de algoritmos busca explorar diferentes aspectos do *dataset* a fim de encontrar uma boa forma de detecção de *bots* [6].

Os algoritmos supervisionados a serem utilizados inicialmente serão os baseados em árvore, mais especificamente o *Random Forest* dado sua grande popularidade e bons resultados conforme os estudos realizados por pesquisas na área [3]. O *Random Forest* é um algoritmo do tipo *ensemble* criado para oferecer uma boa resolução genérica de problemas complexos. Utiliza da aleatoriedade e da combinação de múltiplas árvores de decisão [13].

O algoritmo de aprendizado não-supervisionado utilizado será o *K-means*, um algoritmo conhecido, capaz de agrupar dados em diferentes clusters [9]. A verificação de diferentes parâmetros para o processo de *clustering* é um fator importante para a obtenção dos melhores resultados do modelo, pois como destaca Miller *et al.* [9] mesmo que o *K-means* seja apto a agrupar pontos similares por meio da distância euclidiana, o usuário precisa conhecer a quantidade de *clusters* que devem ser encontrados. Caso um valor de *k* incorreto seja selecionado o agrupamento pode ser impreciso e, portanto, múltiplas escolhas do valor *k* são recomendados para encontrar o melhor modelo. Nesse sentido, **é** importante investigar quantos diferentes tipos de *bots* existem atualmente para alcançar a maior eficiência do algoritmo, realizando testes com diferentes parâmetros.

4.4. Métricas e análise dos resultados

Para o algoritmo não supervisionado, os valores atribuídos manualmente no processo de *crowdsourcing* serão utilizados para se classificar os diferentes clusters a tipos de comportamentos que envolvam ou não o uso de *bots*.

Serão calculadas métricas estatísticas para cada algoritmo, com o intuito de verificar se os dados utilizados obtiveram resultados satisfatórios. As métricas analisadas inicialmente para ambas as abordagens serão o *recall*, o *F-Measure*, a *acurácia* e a *precisão* dado que cada métrica tem sua importância e quando combinadas, compensam pelas fraquezas das outras [9]. O cálculo das quatro métricas pode ser visualizado nas equações 1 a 4 [14].

$$Precisão = \frac{VP}{VP + FP} \tag{1}$$

$$Recall = \frac{VP}{VP + FN} \tag{2}$$

$$F\text{—}measure = 2 \times \frac{Recall \times Precisão}{Recall + Precisão} \tag{3}$$

$$Acurácia = \frac{VP + VN}{VP + FP + VN + FN} \tag{4}$$

5. Cronograma de Execução

Para a execução deste trabalho, espera-se a seguinte distribuição de tempo para cada umas das atividades, sendo elas:

1. Revisão bibliográfica;
2. Coleta de dados e análise exploratória;
3. Aplicação dos algoritmos;
4. Métricas e análise dos resultados;

Tabela 1. Cronograma de Execução

Atividade	ago	set	out	nov	dez	jan	fev	mar	abr	maio
1	x	x								
2		x	x	x	x					
3						x	x			
4								x	x	x

6. Contribuições e/ou Resultados esperados

Espera-se que a pesquisa realizada seja capaz de explicitar elementos gramaticais que possam ser relevantes na classificação de comentários como escritos por usuários reais ou *bots*, em postagens de diferentes temáticas na plataforma *Instagram*. Também espera-se encontrar possíveis novos elementos textuais que auxiliem na detecção do uso da automatização em redes sociais.

7. Espaço para assinaturas

Londrina, 18/09/2023.

Kristiano Pasini de Oliveira

Cinthyan Renata Sachs Camerlengo de Barbosa

Referências

[1] Instagram. *Sobre o Instagram*. Disponível em https://help.instagram.com/424737657584573. Acesso em 08 de julho 2023.

[2] Meta Platforms. *Meta reports first quarter 2023 results*. Disponível em https://investor.fb.com/investor-news/press-release-details/2023/Meta-Reports-First-Quarter-2023-Results/default.aspx. Acesso em 08 de julho 2023.

[3] E. A. Mohamed, E. Alothali, N. Zaki and H. Alashwal. Detecting social bots on twitter: a literature review. *13th International Conference on Innovations in Information Technology (IIT). Al Ain, United Arab Emirates*, pages 175–180, 2018.

[4] Meta. *Spam*. Disponível em https://transparency.fb.com/pt-br/policies/community-standards/spam/?source=https%3A%2F%2Fwww.facebook.com%2Fcommunitystandards%2Fspam/. Acesso em 08 de julho 2023.

[5] M. A. Ruediger. *Robôs, redes sociais e política no Brasil: estudo sobre interferência ilegítimas no debate público na web, riscos à democracia e processo eleitoral de 2018*. FGV DAPP, Rio de Janeiro, 2017.

[6] Fatih Cagatay Akyon and M. Esat Kalfaoglu. Instagram fake and automated account detection. *10th Innovations in Intelligent Systems and Applications Conference*, pages 1–7, 2019.

[7] Emilio Ferrara and Sneha Kudugunta. Deep neural networks for bot detection. *Information Sciences*, 467:312–322, 2018.

[8] Jorge Rodríguez-Ruiz, Javier Israel Mata-Sánchez, Raúl Monroy, Octavio Loyola-González and Armando López-Cuevas. A one-class classification approach for bot detection on twitter. *Computers Security*, 91:101715, 2020.

[9] Zachary Miller, Brian Dickinson, William Deitrick, Wei Hu, and Alex Hai Wang. Twitter spammer detection using data stream clustering. *Information Sciences*, 260:64-73, 2014.

[10] Gabriel Alves Miranda. *Detecção de bots no twitter através de técnicas de Processamento de Linguagem Natural*. Trabalho de Conclusão de Curso, Departamento de Matemática e Estatística da Universidade Federal Fluminense, 2022.

[11] José Luiz Villela Marcondes Mioni. *Processamento da Língua Portuguesa na Detecção de Toxicidade na rede social Twitter*. Dissertação de Mestrado. Departamento de Computação da Universidade Estadual de Londrina, 2023.

[12] Carolinne Roque e Faria. *Ferramenta Carolina para Identificação de Pragas e Doenças na Cultura da Soja utilizando Processamento de Linguagem Natural*. Dissertação de Mestrado. Departamento de Computação da Universidade Estadual de Londrina, 2021.

[13] Arthur Alexandre Artoni. *Aplicação de Aprendizado de Máquina no auxílio ao diagnóstico do Transtorno do Espectro Autista*. Dissertacão de Mestrado. Departamento de Computação da Universidade Estadual de Londrina, 2020.

[14] Felipe Kitamura and Bradley J. Erickson. Magician's corner: 9. performance metrics for machine learning models. *Radiology: Artificial Intelligence*, 3(3): e200126, 2021.