BOTS SOCIAIS: IMPLICAÇÕES NA SEGURANÇA E NA CREDIBILIDADE DE SERVIÇOS BASEADOS NO TWITTER

CARLOS ALESSANDRO SENA DE FREITAS

BOTS SOCIAIS: IMPLICAÇÕES NA SEGURANÇA E NA CREDIBILIDADE DE SERVIÇOS BASEADOS NO TWITTER

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação do Instituto de Ciências Exatas da Universidade Federal de Minas Gerais — Departamento de Ciência da Computação como requisito parcial para a obtenção do grau de Mestre em Ciência da Computação.

Orientador: Adriano Alonso Veloso.
Coorientador: Fabricio Benevenuto de Souza.

Belo Horizonte

Março de 2014

Freitas, Carlos Alessandro Sena de

F866b Bots sociais: implicações na segurança e na credibilidade de serviços baseados no twitter / Carlos Alessandro Sena de Freitas. — Belo Horizonte, 2014
xx, 62 f. : il. ; 29cm

Dissertação (mestrado) — Universidade Federal de Minas Gerais — Departamento de Ciência da Computação
Orientador: Adriano Alonso Veloso.
Coorientador: Fabricio Benevenuto de Souza.

1. Computação - Teses. 2. Redes de relações sociais - Teses. 3. Redes de computadores - Medidas de segurança - Teses. 4. Aprendizado do Computador - Teses. I. Orientador. II. Coorientador. III. Título.

CDU 519.6*04 (043)

UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS EXATAS
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

FOLHA DE APROVAÇÃO

Robôs sociais: implicações na segurança e na credibilidade de serviços baseados em microblogs

CARLOS ALESSANDRO SENA DE FREITAS

Dissertação defendida e aprovada pela banca examinadora constituída pelos Senhores:

PROF. ADRIANO ALONSO VELOSO - Orientador
Departamento de Ciência da Computação - UFMG

PROF. FABRÍCIO BENEVENUTO DE SOUZA - Coorientador
Departamento de Ciência da Computação - UFMG

PROF. DORGIVAL OLAVO GUEDES NETO
Departamento de Ciência da Computação - UFMG

PROF. MARCO ANTÔNIO PINHEIRO DE CRISTO
Departamento de Ciência da Computação - UFAM

Belo Horizonte, 27 de março de 2014.

Agradecimentos

Este trabalho só foi possível graças a ajuda e apoio de pessoas que admiro, cujas contribuições e conselhos foram fundamentais para tomar as decisões corretas que culminaram nesta dissertação.

Gostaria de começar agradecendo à minha mãe Eunice Sena, ao meu pai João de Souza e ao meu irmão Jean Sena, que nunca deixaram de acreditar em mim e estiveram sempre ao meu lado. Agradeço aos meus grandes amigos Samuel Sérvulo e Rodrigo Borges, que tornaram-se verdadeiros irmãos durante estes últimos dois anos e sem os quais esse período não seria sinônimo de bons momentos e coleguismo.

Gostaria de agradecer também aos meus orientadores de graduação Edleno Moura e Marco Cristo, cuja orientação foi fundamental para minha entrada no mestrado. Trabalhar com eles foi essencial na minha formação.

A graduação foi um longa jornada, porém sempre pude contar com o apoio e as dicas de colegas como Gerson Barreiros, Javier Medina, Julio Machado, Luis Menezes, Rodrigo Borges e Rodrigo Maues, além dos colegas do laboratório BDRI André Carvalho, Antonio Sobrinho, Cristian Rossi, Diego Rodrigues, Eli Cortez, Felipe Hummel, Guilherme Monteiro, Guilherme Toda, Juliana Nunes, Karane Vieira, Klessius Berlt, Kleverson Paixão, Leticia Santos, Ludimila Carvalho, Mauro Rojas, Onilton Maciel, Vivian Lô.

Agradeço aos colegas que fizeram parte do meu dia a dia no mestrado Aline Bessa, Javier Medina, Rogerio Fonteles, Sabir Ribas, Thales Costa, e aos colegas do e-SPEED, laboratório no qual fiz pesquisa, Alex de Sá, Bruno Coutinho, Camila Araújo, Denise Eb, Diogo Rennó, Elverton Fazzion, Fernando Carvalho, Filipe Arcanjo, Gabriel Poesia, Hélio Almeida, Júlio Albinati, Luam Totti, Luiz Oliveira, Natália Tereza, Osvaldo Fonseca, Paulo Bicalho, Pedro Calais, Raphael Luciano, Tatiana Schmidt, Walter Santos e, em especial, ao Silvio Soares, que tornou-se um grande colega e amigo durante os vários momentos difíceis do mestrado. Agradeço também aos colegas do grupo de pesquisa de aprendizado de máquina LAMA Adriano Pereira, Alexandre Guelman, Antônio Carlos, Bruna Neuenschwander, Gabriel Carvalho, Itamar Hata, Isabella Brito, Mariane Souza e, em especial, ao Roberto Oliveira, que foi um grande amigo durante o mestrado. Agradeço também a Ana Paula Nunes e Aline Mourão cujo apoio e amizade foram de grande ajuda nessa jornada.

Gostaria de agradecer aos meus orientadores Adriano Veloso e Fabrício Benevenuto.

Este trabalho só foi possível graças a eles. No entanto, as contribuições se estendem muito além das páginas deste trabalho. Seus conselhos, conversas e incentivos em momentos difíceis são lições que levarei pelo resto da vida.

Gostaria de agradecer à banca examinadora e às pessoas que revisaram esta dissertação, pelo tempo dedicado e pelas dicas valiosas: Adriano Veloso, Fabrício Benevenuto, Marco Cristo, Dorgival Guedes e Samuel Sérvulo.

Finalmente, gostaria de agradecer à três pessoas sem os quais esta pesquisa não teria sido possível. Saptarshi Ghosh que forneceu a base de dados utilizada em nossa pesquisa, Guido van Rossum criador da linguagem de programação Python e, finalmente, ao criador dos Mojitos uma bebida capaz de trazer alegria as noites mais sofridas deste mestrado, quem quer que você seja OBRIGADO!!

"Don't Panic."
(Douglas Adams, The Hitchhiker's Guide to the Galaxy)

Resumo

Cada vez mais, dados extraídos de redes sociais são utilizados para a construção de novas aplicações e serviços, como plataformas para monitoramento de trânsito, identificação de surtos epidêmicos, bem como várias outras aplicações associadas à criação de cidades inteligentes, por exemplo. Entretanto, tais serviços são vulneráveis a ataques de bots − contas automatizadas − que buscam adulterar estatísticas de percepção pública postando um excessivo número de mensagens geradas automaticamente. Bots podem invalidar diversos serviços existentes, o que torna crucial entender as principais formas de ataque, bem como buscar mecanismos de defesa. Este trabalho apresenta uma ampla caracterização do comportamento de bots no Twitter. A partir de uma base de dados real contendo 19.115 bots, foram identificadas diversas características dos bots, extraídas de padrões de comportamento e de escrita de texto, que possuem alto poder discriminativo. A partir dessas características, apresentamos um método de detecção automática de bots capaz de detectar 92% deles, enquanto menos de 1% dos usuários reais são classificados erroneamente. Finalmente, realizamos um estudo sobre quais características tornam os bots mais bem sucedidos em tarefas de infiltração. Para isso, foram criados 120 socialbots no Twitter. Durante 30 dias monitoramos seu comportamento e todas suas interações com usuários da rede, assim como com 600 usuários-alvo. Durante esse período nossos bots interagiram 5.966 vezes com 2.637 usuários do Twitter.

Palavras-chave: Twitter, Bots, Redes Sociais, Aprendizado de Máquina.

Abstract

More and more, data extracted from social networks is used to build new applications and services, such as traffic monitoring platforms, identification of epidemic outbreaks, as well as several other applications related to the creation of smart cities, for example. However, such services are vulnerable to attacks from bots − automatized accounts − seeking to tamper statistics of public perception posting an excessive number of messages generated automatically. Bots can invalidate many existing services, which makes it crucial to understand the main forms of attacks and to seek defense mechanisms. This work presents a wide characterization of the behavior of bots on Twitter. From a real data set containing 19,115 bots, several characteristics of bots were identified, extracted from behavior and writing patterns, that have discriminative power. From these features, we present an automatic detection method capable to detect 92% of the bots while only less than 1% of real users are misclassified. In addition, we conducted a study on which characteristics makes a bot most successful in infiltration tasks. For this study we created 120 socialbots on Twitter. During 30 days we monitored their behavior and interactions with all network users, as well as 600 target users. During this period our bots had 5,966 interactions with 2,637 Twitter users.

Keywords: Twitter, Bots, Social Networks, Machine Learning.

Lista de Figuras

Lista de Tabelas

Sumário

Capítulo 1

Introdução

O Twitter é uma das redes sociais mais populares da atualidade, na qual seus usuários geram mais de 500 milhões de mensagens diariamente (Protalinski [2013]), o que, aliado a sua API aberta, tem tornado a plataforma largamente utilizada para serviços de extração de conhecimento. Como exemplo podemos citar a predição de mudanças no mercado de ações (Zhang & Paxson [2011]), a detecção de catástrofes em tempo real (Sakaki et al. [2010]), detecção de focos de epidemias (Gomide et al. [2011]) e também análise de opinião (Tumasjan et al. [2010]). Geralmente, esses serviços usam amostras do Twitter, tornando-se vulneráveis a ataques com o objetivo de adulterar suas estatísticas. Por exemplo, um ou mais usuários podem postar mensagens sobre um tópico específico para direcionar os resultados de um algoritmo de análise de opinião. Mais importante, robôs (ou bots) podem ser utilizados para postar mensagens enviesadas sobre um tópico específico (e.g., postar mensagens favorecendo algum candidato político).

Socialbots, bots desenvolvidos de forma a se passar por humanos, já são usados com o objetivo de enganar e influenciar outros usuários na rede (Messias et al. [2013]; Harris [2013]). Esses bots se aproveitam de um conjunto de vulnerabilidades inerentes das redes sociais atuais para se infiltrar na rede de usuários legítimos (Boshmaf et al. [2012]). Dessa forma os mesmos possuem a capacidade de comprometer a estrutura da rede social (Boshmaf et al. [2011]), permitindo assim que eles ganhem influência na rede. Bots podem ser explorados para a difusão de propaganda e informações erradas na rede. Por exemplo, uma rede de socialbots pode ser usada para a propagação de ações políticas ou publicitárias que tentam criar a impressão de que são movimentos espontâneos e populares (Ratkiewicz et al. [2011]). Além disso, bots já são usados por candidatos políticos durante campanhas eleitorais com o objetivo de alterar os "trending topics" (Orcutt [2012]), ou para aumentar artificialmente seus números de

seguidores, e consequentemente seus índices de popularidade (Calzolari [2012]). Este cenário só piora quando consideramos a existência de serviços de venda de bots.[1,2,3,4]

A quantidade exata de bots no Twitter é desconhecida. Chu et al. estimam que 50% das contas sejam associadas a bots (Chu et al. [2012]). Contudo, o Twitter afirma que contas falsas ou spammers representam apenas 5% dos seus 215 milhões de usuários ativos (Gara [2013]). Seja 5% ou 50%, entender o impacto dos bots no Twitter, assim como desenvolver estratégias para sua detecção é crucial para garantir a credibilidade e segurança dos serviços que usam o Twitter como fonte de dados.

1.1 Objetivos

Neste trabalho realizamos um estudo em largura sobre bots no Twitter, dessa forma, realizamos dois estudos complementares para entender o seu impacto, assim como a criação de uma estratégia de defesa contra ataques de bots. Os objetivos deste trabalho são:

(a) Estudar o processo de infiltração de bots na rede do Twitter. Para isso, foram criados 120 socialbots no Twitter. Durante 30 dias monitoramos seu comportamento e todas suas interações com usuários da rede, assim como com 600 usuários-alvo. Ao final de nosso experimento, analisamos que fatores tornam um bot mais bem sucedido em tarefas de infiltração

(b) A criação de uma estratégia supervisionada para detecção de bots. Para isso foi construída uma coleção contendo 19.115 bots, identificados através de uma abordagem de identificação de padrões automáticos de postagem. Além disso, estudados o poder discriminativo de diversas características dos bots, extraídas de padrões de comportamento e de escrita de texto.

1.2 Contribuições

As principais contribuições deste trabalho são:

- Estudo sobre a vulnerabilidade de usuários do Twitter a ataques de bots.

[1] http://www.jetbots.com/
[2] http://youtubebots.org/
[3] http://instagress.com/
[4] http://bestsocialbots.com/

- Estudo de vários atributos dos bots e seu impacto em tarefas de infiltração no Twitter.
- Estudo de diversas características de usuários e seu impacto em tarefas de infiltração no Twitter.
- A caracterização do comportamento de bots em uma grande base de dados .
- Identificação de atributos linguísticos na postagem de bots, que até onde tenhamos conhecimento nunca foram utilizados para a detecção de bots.
- Criação de um método de detecção automática de bots que explora os atributos identificados.
- Disponibilização das bases de dados utilizadas neste trabalho.

1.3 Organização do texto

O texto está organizado da seguinte forma: no capítulo 2, serão introduzidos os conceitos fundamentais utilizados neste trabalho e será apresentada uma revisão da bibliografia relevante ao tema; no capítulo 3, será explicada nossa estratégia de detecção de bots, assim como os resultados obtidos pela mesma; no capítulo 4, detalharemos o processo de infiltração realizado por 120 bots durante o período de 30 dias, além de apresentar os resultados obtidos; por fim, no capítulo 5, são apresentadas nossas principais conclusões e os trabalhos futuros.

Capítulo 2

Referencial Teórico e Trabalhos Relacionados

2.1 Bots

Um bot é uma aplicação de software que executa tarefas automatizadas. Normalmente, bots executam tarefas que são simples e repetitivas, a uma taxa muito mais elevada do que seria possível para um ser humano. Um conjunto de bots conectados à Internet que se comunicam com a finalidade de executar uma tarefa em comum é denominado de Botnet.

Bots, ou botnets, podem ser utilizados para uma série de ciberataques, entre os principais temos:

- **Spam:** o termo spam refere-se ao envio de mensagens não solicitadas em massa, especialmente publicidade. Enquanto a forma de spam mais conhecida são os e-mails de spam, o termo também se aplica a abusos similares em vários meios: fóruns, chats, páginas web, máquinas de buscas e redes sociais online.

- **Phishing:** é uma fraude eletrônica, caracterizada pelo ato de tentar tornar-se dono de informações pessoais (e.g., senhas e dados bancários). Para isso o fraudador se faz passar por uma pessoa ou empresa confiável enviando uma mensagem eletrônica oficial. Os principais meios usados para a fraude são e-mail, mensagens instantâneas, SMSs e redes sociais.

- **Ataque Sybil:** refere-se ao uso de múltiplas contas para burlar um sistema de reputação. Nesse tipo de ataque, um usuário mal-intencionado cria várias identidades com o objetivo de tirar proveito dessas identidades para atacar o sistema.

Por exemplo, em redes sociais como o Yelp,[1] onde os lugares são avaliados com base em notas dadas pelos usuários, um fraudador pode criar várias identidades para manipular a popularidade dos mesmos.

- **Link Farm:** referia-se originalmente ao processo de troca recíproca de hiperlinks entre páginas web com o objetivo de influenciar os resultados de máquinas de buscas. Uma fazenda de links é uma forma de spam no índice de máquinas de busca. A principal consequência desta atividade é que a qualidade dos resultados das buscas diminui. Além disso, os índices de máquinas de buscas são inflados com páginas irrelevantes, dessa forma aumentando o custo de cada consulta processada. Portanto, a identificação de web spam é um dos principais desafios de máquinas de busca Henzinger et al. [2002].

A principal técnica anti-bot utilizada é o uso de CAPTCHAs, que é na realidade um teste de Turing reverso usado para distinguir entre um utilizador humano e um bot através da codificação gráfica de textos. Outra técnica largamente usada consiste no uso de algoritmos de aprendizado de másquina para detectar padrões de comportamento considerados suspeitos.

Figura 2.1: Exemplo de CAPTCHA

2.2 Tipos de ataques e seus mecanismos de defesa

Nesta seção apresentamos os principais mecanismos de defesa propostos na literatura para os vários tipos de ataques:

2.2.1 Spam

2.2.1.1 E-mail Spam

O recebimento de mensagens eletrônicas indesejadas é ainda hoje um problema sério. Estudos indicam que foram enviados mais de 94 bilhões de mensagens de spam

[1] http://www.yelp.com/

por dia em 2012 (Grandoni [2012]). Além disso, spam ocasiona vários problemas, alguns gerando perdas financeiras diretas. Mais precisamente, o desperdício de tráfego, armazenamento e poder computacional, além do desperdício de tempo e recursos humanos (Siponen & Stucke [2006]). Finalmente, estima-se que as perdas financeiras causadas por spam anualmente sejam em torno de $20 bilhões, enquanto spammers e comerciantes anunciando spam tenham uma receita bruta de $200 milhões por ano (Rao & Reiley [2012]).

Isso mostra que a filtragem de spam é, e provavelmente continuará sendo, uma importante aplicação prática da aprendizagem de máquina. Técnicas de filtragem bem sucedidas incluem filtros baseados em "Bag-of-Words", que tratam o e-mail como um conjunto não estruturado de tokens (Pantel & Lin [1998]; Sahami et al. [1998]; Drucker et al. [1999]; Androutsopoulos et al. [2000]; Metsis & Metsis [2006]), métodos baseados em características linguísticas (Bratko et al. [2006]; Medlock [2006]; O'Brien & Vogel [2003]), filtros baseados em cabeçalhos ou meta-atributos dos e-mails (Palla & Dantu [2007]), filtros que usam a rede do usuário (James & Hendler [2004]; Boykin & Roychowdhury [2005]; Chirita et al. [2005]), métodos que detectam comportamentos típicos de spammers (Yeh et al. [2005]; Hershkop [2006]) e, finalmente, métodos de filtragem colaborativa (Lazzari et al. [2005]; Zhou et al. [2003]; Damiani et al. [2004]; Mo et al. [2006]; Garg et al. [2006]).

2.2.1.2 Opinion Spam

Com a crescente popularidade de sites de *reviews* que apresentam opiniões geradas por usuários (e.g., Amazon[2] e Yelp), surge um grande potencial para o ganho monetário por meio de *Opinion spam* – *reviews* inapropriados ou fraudentos. Em contraste aos ataques de spam em serviços de e-mail, spam em *reviews* podem ser utilizados com o objetivo de influenciar o usuário na tomada de decisões (e.g., difamar um produto ao inserir vários revisões falsas de teor negativo). Jindal & Liu [2008] analisaram 5,8 milhões de *reviews* da amazon.com, identificando três principais tipos de spam: (i) opiniões falsas (comentários que promovem ou difamam os produtos), (ii) opiniões sobre marcas, porém não produtos, e (iii) *reviews* sem opinião (e.g., anúncios); além de estratégias de detecção. Posteriormente, Lim et al. [2010] desenvolveram uma técnica para detectar spammers em *reviews* com base no seus comportamentos de avaliação.

[2]http://www.amazon.com/

2.2.1.3 Spam Social

Com os serviços de e-mail melhorando significante seus métodos de detecção e filtragem de spam e a crescente popularidade das redes sociais, os spammers estão migrando para as mesmas com o objetivo de obter um maior ganho monetário. Dessa forma o spam em mídias sociais aumentou em média 355% no primeiro semestre de 2013 (Franceschi-Bicchierai [2013]). Em 2012 o Facebook informou que apenas 4% do conteúdo gerado por seus usuários apresenta algum tipo de spam, enquanto o Twitter afirma que apenas 1,5% dos tweets continham spam (Geoffrey A. Fowler [2012]). Esta prática pode comprometer a confiança dos usuários no sistema, prejudicando, assim, seu sucesso na promoção de interações sociais.

Um dos maiores desafios na detecção de spam em mídias sociais é que os spams geralmente têm forma de imagens e texto, além do contexto da rede social na qual estão inseridos. O que demanda soluções abrangentes, que possam considerar texto, imagens e os recursos da rede social, além de também serem escaláveis e capazes de realizar a detecção em tempo real. Thomas et al. [2011] descobriram que e-mails de spam diferem qualitativamente de maneira significativa de campanhas de spam no Twitter. Entre alguns estudos sobre spam em redes sociais destacamos:

Benevenuto et al. [2010b] fornecem uma visão geral da poluição em sistemas de compartilhamento de vídeo (evidência de poluição, tipos de poluição, efeito sobre o sistema e estratégias de controle). O'Callaghan et al. [2012] propuseram um método para identificar campanhas de spam no YouTube usando métodos de análise de rede. Sureka [2011] descreve um método para identificação de spammers em comentários do YouTube pela mineração do log de atividades de comentários dos usuários. Finalmente, Benevenuto et al. [2009] estudaram o comportamento de poluidores de conteúdo no YouTube e desenvolveram um método supervisionado para detectá-los.

Stringhini et al. [2010] realizaram um estudo em três principais redes sociais (Facebook, MySpace e Twitter), além de desenvolverem técnicas para identificar spam bots, assim como campanhas de spam em larga escala. Em outro trabalho, Irani et al. [2010] analisaram mais de 1,9 milhões de perfis do MySpace e criaram um método capaz de detectar perfis de spammers quase no momento de criação dos mesmos com mais de 99% de acurácia.

Grier et al. [2010] analisaram 400 milhões de tweets e detectaram que 8% continham algum tipo de spam, além disso, analisando o comportamento de spammers, verificaram que apenas 16% das contas de spam são claramente bots automatizados, enquanto que os 84% restantes parecem ser contas comprometidas sendo controladas por spammers. Benevenuto et al. [2010a] investigaram o uso de aprendizado super-

visionado para detectar spammers no Twitter, analisando atributos do usuário e seu comportamento. Lee et al. [2011] realizaram um estudo de longo prazo sobre poluidores de conteúdo no Twitter usando "honeypots", perfis criados para atrair spammers, cujo modelo conseguiu detectar spammers com 98% de acurácia. Finalmente, Thomas et al. [2013] investigaram, durante 10 meses, o mercado negro de venda de contas em serviços sociais e criaram um método para a detecção de contas fraudulentas. Esse método é capaz de detectar contas fraudulentas com 99% de precisão antes mesmo delas iniciarem qualquer atividade ilegal.

Markines et al. [2009] propuseram um método supervisionado para detecção de spam em serviços de "social bookmarking" com 98% de acurácia. Finalmente, Costa et al. [2013] desenvolveram um método de detecção de spam em dicas dentro de redes sociais baseadas em localização. De forma similar Aggarwal et al. [2013a] desenvolveram um mecanismo para detecção de spammers no Foursquare.

2.2.2 Phishing

Apesar de phishing ser um tipo de spam, caracteriza-se por possuir certas propriedades únicas, visto que mensagens de phishing são projetadas de forma a parecerem mensagens legítimas de uma empresa ou pessoa. Dessa forma, espera-se que mensagens de phishing sejam mais difíceis de detectar que mensagens gerais de spam.

Com isto em mente, vários métodos foram propostos na literatura para detectar phishing. Fette et al. [2007] desenvolveram um método para detecção de mensagens de phishing em serviços de email. Whittaker et al. [2010] descrevem um sistema em larga escala para detectar páginas que contenham phishing usando aprendizado de máquina com uma taxa de falsos positivos inferior a 0,1%. De forma complementar, foram desenvolvidos vários métodos para detecção de páginas de phishing com base em características extraídas da própria URL (Garera et al. [2007]; Blum et al. [2010]). Zhang et al. [2007] desenvolveram uma abordagem baseada em conteúdo para detectar sites de phishing baseada em TF-IDF;

Chhabra et al. [2011] identificaram ataques de phishing em redes sociais usando encurtadores de URL, além disso, detectaram que a maior parte do phishing em tweets é automatizado. Gao et al. [2010] analisaram 200.000 postagens maliciosas no Facebook e detectaram que mais de 70% das URLs direcionavam para um site de phishing, além de detectarem que 97% das mensagens eram postadas a partir de "perfis comprometidos", enquanto apenas 3% tinham como origem perfis falsos. Finalmente, Aggarwal et al. [2013b] desenvolveram um método para detecção de phishing em tempo real no Twitter.

2.2.3 Ataque Sybil

Recentemente, uma série de métodos têm sido propostos para se defender contra ataques Sybil aproveitando as redes sociais (Mislove et al. [2008]; Post et al. [2011]; Li & Subramanian [2010]; Tran et al. [2009]). Viswanath et al. [2012a] analisaram defesas Sybil baseadas em rede social e dividiram as propostas existentes em duas categorias, detecção de Sybil e tolerância Sybil. A primeira categoria, chamada de métodos de detecção de Sybil, funciona através da detecção de identidades que provavelmente são Sybils. Em contraste, os métodos de tolerância Sybil não tentam rotular identidades como Sybil ou não-Sybil. No lugar disso, seu objetivo é limitar o benefício que um atacante pode obter usando múltiplas identidades Sybil. Apesar de suas diferenças, ambas as técnicas possuem o mesmo objetivo em comum, que é o de impedir que os atacantes obtenham uma vantagem ao criar e utilizar múltiplas identidades na rede.

Os métodos de detecção de Sybils supõem que, apesar de um atacante poder criar várias identidades Sybil em redes sociais, essas identidades não podem estabelecer um número arbitrariamente grande de conexões sociais para nós não-Sybil. Dessa forma, nós Sybil tendem a ser fracamente ligados ao resto da rede, em comparação com os nós não-Sybil. Métodos de detecção analisam a rede para identificar características topológicas resultantes da limitada capacidade dos Sybils de estabelecer laços sociais (Yu et al. [2006, 2008]; Danezis & Mittal [2009]; Tran et al. [2011]). Em um estudo, Viswanath et al. [2010], descobriram que apesar das diferenças entre os métodos, todos eles consistem em identificar comunidades dentro da rede social, que é um problema largamente estudado na literatura.

Nos métodos de detecção de Sybils a presença de nós Sybils é um indício de comportamento malicioso, e dessa forma um nó não-Sybil não deveria interagir com um nó Sybil. No entanto, existem razões legítimas para que um usuário possa querer criar várias identidades. Por exemplo, os usuários podem querer dividir a sua identidade em uma que é utilizada para interagir com os colegas de trabalho e outra que é usada para interagir com amigos e familiares. Usuários postando vídeos no YouTube podem desejar publicar conteúdo sob pseudônimos a fim de evitar revelar sua identidade no mundo real, enquanto usam uma conta pessoal para classificar vídeos e postar comentários.

Uma vez que a mera presença de usuários com múltiplas contas não é necessariamente um indício de mal comportamento, os métodos de tolerância Sybil preocupam-se não com a presença de Sybils, mas sim no seu uso em atividades maliciosas. Mislove et al. [2008] propuseram um sistema que utiliza as relações de confiança existentes entre os usuários para impedir a comunicação indesejada. Tran et al. [2009] desenvolveram um sistema de votação de conteúdo que utiliza redes de confiança entre os usuários

para se defender contra ataques Sybil. Post et al. [2011] apresentaram o Bazaar, um sistema que reforça a reputação de usuários em mercados on-line. Bazaar é baseado em cálculos de fluxo máximo em uma rede de risco, uma estrutura de dados que codifica a quantidade de risco compartilhado entre os participantes recompensados. Finalmente, Viswanath et al. [2012b] apresentaram o Canal, um sistema eficiente e preciso para transferir pagamentos de crédito em grandes redes de crédito. Canal foi concebido para complementar os métodos de tolerância Sybil já existentes, como os apresentados previamente, tornando seu uso prático no mundo real.

2.2.4 Link Farm

Link Farm tem sido amplamente estudado no contexto da web. Estudos já demostraram que algoritmos de ranking podem ser influenciados por certas relações no grafo da web (Bharat & Henzinger [1998]; Lempel & Moran [2000]). Usuários maliciosos tentam tirar proveito disso para obter um alto ranking em máquinas de busca. Gyöngyi & Garcia-Molina [2005] estudaram a estrutura de link farms e como suas páginas podem se interconectar para otimizar rankings.

Várias soluções para combater link farm foram propostas. Estas soluções podem ser divididas em duas categorias principais: técnicas que usam apenas o conteúdo das páginas web, e aquelas que utilizam a estrutura dos links página, além das abordagens que usam os dois tipos de evidência.

Becchetti et al. [2006] utilizaram métricas baseadas em links para construir um classificador para detectar automaticamente Web-spam. Gyöngyi et al. [2004] propôs o algoritmo de TrustRank; este algoritmo assume que boas páginas geralmente se conectem a outras boas páginas, desta forma o algoritmo atribui altos escores para páginas confiáveis e então os propaga de forma similar ao PageRank (Page et al. [1999]). Alguns algoritmos que funcionam de forma inversa ao TrustRank, analisam a relação de uma páginas novas com páginas confirmadas de spam, também têm sido propostos para identificar páginas de spam (Krishnan [2006]; PR0-Pagerank-Penalty [2002]; Wu & Davison [2005]).

Ntoulas et al. [2006] propuseram um método de classificação que usa características baseadas no conteúdo da página para identificar páginas de spam. Mishne et al. [2005] desenvolveram método que utiliza modelos de linguagem para detectar páginas de spam.

Exemplos que utilizam o conteúdo da página em conjunto com sua estrutura de links, incluem Fetterly et al. [2004] que detectaram várias propriedades capazes de diferenciar páginas de spam, entre elas temos a distribuição de in-degress e out-degress

e a excessiva replicação de conteúdo presentes em páginas maliciosas. Castillo et al. [2007] usam a topologia da rede e o conteúdo das páginas para detectar páginas de spam com o intuito que duas páginas conectadas pertencem à mesma classe (spam ou não-spam).

Finalmente, Ghosh et al. [2012] realizaram uma análise de link farm no Twitter, descobriram que um pequeno número de contas legítimas, populares e altamente ativas são responsáveis pela maior parte de atividade de link farm no Twitter e que um grupo de spammers toma proveito desse grupo para ganhar seguidores e reputação na rede. Posteriormente, desenvolveram um método de ranking que penaliza os usuários que seguem spammers.

2.3 Socialbots

Existem vários estudos com foco na criação e análise de socialbots. O projeto Realboy visa a criação de bots que imitam usuários reais de forma verossímil (Coburn & Marra [2008]). O Web Ecology Project[3] visa a criação de socialbots para interagirem com um grupo de usuários no Twitter. Messias et al. [2013] criaram bots capazes de interagir com usuários legítimos no Twitter. Durante o período de 90 dias os mesmos conseguiram resultados significantes em sistemas medidores de influência como o Klout[4] e Twitalyzer.[5] Boshmaf et al. [2011] projetaram uma rede social de bots com o intuito de realizar uma infiltração em larga escala. O estudo demonstrou que redes sociais podem ser infiltradas com uma taxa de sucesso de até 80%. Finalmente, Elishar et al. [2012] demonstraram como adversários podem usar socialbots para coletar informações de funcionários de uma organização, a fim de reconstruir e aprender melhor rede social da mesma. Em um estudo similar Elyashar et al. [2013] usaram um sofisticado algoritmo de solicitações de amizade, a fim de se infiltrar em usuários específicos de organizações alvo com até 70% de requisições aceitas. De maneira geral, esses esforços demostram a vulnerabilidade de redes sociais à infiltração de bots.

2.3.1 Detectando Bots no Twitter

Apesar dos métodos de detecção apresentados na seção anterior poderem ser utilizados para detectar bots envolvidos em atividades maliciosas, seu desempenho não é claro na detecção de bots que não estejam envolvidos nesse tipo de atividade (e.g., bots usados

[3]http://www.webecologyproject.org/category/competition/
[4]http://klout.com/
[5]http://twitalyzer.com/

para postar a temperatura de uma região a cada minuto). Dessa forma, nosso estudo pode ser considerado ortogonal aos métodos apresentados anteriormente, visto que, foca na detecção de bots e não de padrões de ataques. Entre os principais trabalhos para detecção de bots destacamos os descritos nos próximos trabalhos:

Chu et al. [2012] usam técnicas de aprendizado de máquina para identificar três tipos de contas: usuários, bots e ciborgues (usuários assistidos por bots). Eles mostram que a regularidade de postagem, a fração de tweets com URLS e o meio de postagem (o uso de aplicativos externos) apresentam indícios de qual é o tipo da conta. Além disso, o método exige que os tweets sejam rotulados como spam e não-spam. A principal diferença desse método para o proposto neste trabalho é que o nosso método não utiliza atributos temporais, além do fato de não exigir a rotulação de tweets de spam.

Zhang & Paxson [2011] desenvolveram um método para detecção de contas com atividade automatizada usando apenas o "timestamp" das mensagens por meio de um teste χ^2. Apesar desses métodos apresentarem bons resultados, eles podem ser facilmente burlados por bots que: (i) postem com intervalos aleatórios ou sigam uma distribuição similar a comportamentos típicos de humanos, (ii) diminuam a fração de tweets com URLs, e (iii) usem ferramentas para automação web que imitem um navegador, (e.g., phantomjs[6] e o fake[7]). Dessa forma nossa abordagem visa a identificação de atributos mais difíceis de serem burlados por bots, como a estrutura dos tweets e o padrão de escrita, além das características do usuário.

2.3.2 Engenharia Reversa

De forma complementar à detecção de bots, Wagner et al. [2012] criaram um modelo de aprendizado de máquina para prever a suscetibilidade dos usuários a ataques de socialbots, utilizando três componentes diferentes de atributos (a rede do usuário, seu comportamento e características linguísticas). Seus resultados apontam que usuários mais "abertos" a interações sociais são mais suscetíveis a ataques. Posteriormente, Wald et al. [2013] realizaram um estudo similar e encontraram que o Klout score, número de seguidores e de amigos, são bons previsores se um usuário irá interagir com um bot. Neste trabalho realizamos um estudo complementar a estes trabalhos, isto é, investigamos que características tornam um bot mais popular na rede. Para isto, aplicamos engenharia reversa em algumas características detectadas no nosso estudo de detecção de bots.

[6] http://phantomjs.org/
[7] http://fakeapp.com/

Capítulo 3

Detectando bots no Twitter

Neste capítulo, abordamos o problema de detectar bots no Twitter utilizando uma abordagem supervisionada. Nosso foco está na identificação de comportamentos de bots que extrapolam as estratégias de identificação de atividade automática. O capítulo está organizado da seguinte forma: Na próxima seção descrevemos a construção de uma base de dados de bots utilizada em nossos experimentos. Na seção 3.2 apresentamos um estudo dos atributos usados por nosso método. Finalmente, na s eção 3.3 apresentamos os resultados obtidos por nosso método.

3.1 Base de dados

Para estudar o comportamento de bots no Twitter, precisamos de uma amostra ampla e representativa de bots e usuários legítimos. Até onde conhecemos, nenhuma coleção com tais características está disponível publicamente. Descrevemos a seguir como construímos a coleção para nossos experimentos. O conjunto de dados utilizado é um "snapshot" completo da rede do Twitter e todos os tweets postados por todos os usuários até agosto de 2009 (Cha et al. [2010]). Mais especificamente, o conjunto de dados contém 54.981.152 usuários ligados uns aos outros por 1.963.263.821 arestas. O conjunto de dados também contém todos os tweets postados pelos usuários coletados, que consiste em 1.755.925.520 tweets. Cerca de 8% das contas eram privadas, o que implica que apenas seus seguidores poderiam ver seus tweets. Posteriormente Ghosh et al. [2012] recoletaram os usuários desta base de dados em fevereiro de 2011, encontrando um total de 379.340 contas suspensas pelo Twitter.

Nossa estratégia consiste em investigar essas contas suspensas para identificar bots, através de um método de detecção de atividade automática no Twitter, que foi previamente mencionado na seção 2.3.1 (Zhang & Paxson [2011]). Além disso, nós sele-

cionamos uma amostra de um milhão de contas não suspensas que, conjuntamente com as contas suspensas, foram submetidas ao teste de atividade automática. Uma conta é reprovada no teste quando ela apresenta um comportamento altamente automatizado (e.g., postagem de tweets em intervalos regulares de tempo). Finalmente, como o método precisa de pelo menos 30 tweets para funcionar, as contas com menos de 30 tweets foram consideradas "insuficientes". Apesar do método realizar uma análise simples, o mesmo nos permitiu criar uma grande coleção rotulada e assim realizar um estudo de comportamentos mais complexos dos bots no Twitter. Nossa abordagem consiste em investigar outros aspectos relativos ao comportamento e padrões de escrita dessas contas, na tentativa de identificar mesmo bots com comportamentos mais complexos.

Tabela 3.1: Teste de atividade automática

	Com atividade automática	Sem atividade automática	< 30 tweets
Não suspensas	5.755	91.118	903.127
Suspensas	19.115	25.355	334.869

Como podemos perceber pelos resultados da tabela 3.1, cerca de 42% das contas suspensas com pelo menos 30 tweets utilizam algum método de atividade automática, enquanto menos de 6% das contas não suspensas com tweets suficientes usam um recurso similar.

Para compor nossa base de dados consideramos as contas não suspensas que não apresentaram nenhum método de automatização como usuários legítimos. De forma similar, consideramos que as contas suspensas com atividade automática são bots. Dessa forma, nossa base de dados contém **110.233** (91.118+19.115) contas e **42.773.272** de tweets.

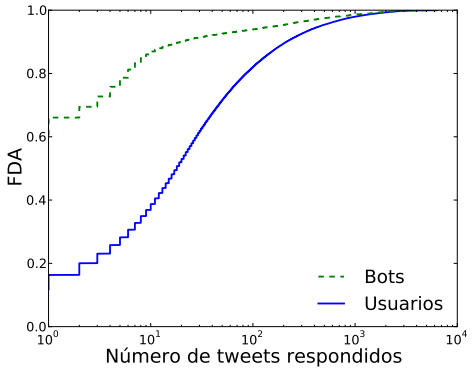
3.2 Analisando atributos de usuários

De forma diferente dos humanos, bots geralmente são criados com algum objetivo específico: invadir um grupo de usuários, espalhar spam, postar mensagens sobre um tópico em particular, etc. Além disso, bots simples não são capazes de interagir inteligentemente com outros usuários (e.g., respondendo perguntas encaminhadas aos mesmos). Dessa forma, é esperado que usuários e bots possuam comportamentos diferentes. Intuitivamente, esperamos que humanos sejam mais sociais e ativos em conversas, enquanto que os bots postam mais tweets, enviesados para algum tópico em particular ou contendo URLs. Para comprovar isto, analisamos um grande conjunto de atributos extraídos de padrões de comportamento e de escrita do texto. Consideramos três con-

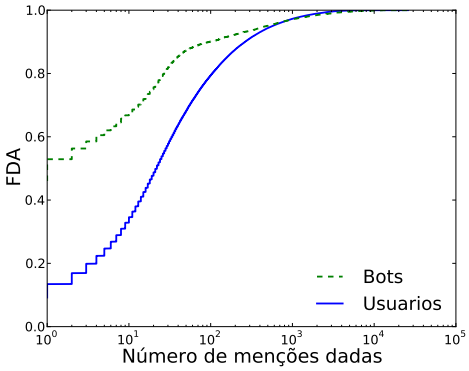
juntos de atributos: (i) atributos de conteúdo, (ii) atributos do usuário e (iii) atributos linguísticos.

3.2.1 Atributos do usuário

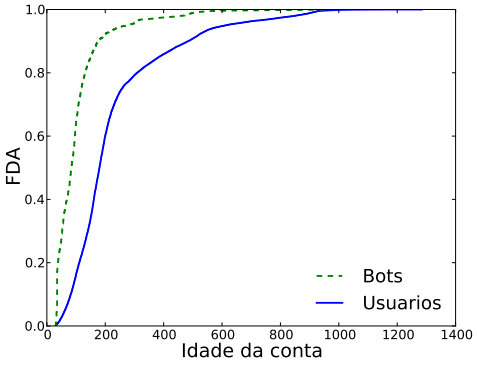
Atributos do usuário capturam características como a influência na rede do Twitter e as interações sociais do usuário. Foram consideradas as seguintes métricas como atributos de usuário: número de seguidores, número de amigos, a razão de seguidores por amigos, número de tweets, idade da conta do usuário − o número de dias entre a criação da conta e do tweet mais novo analisado por nós, número de vezes que o usuário foi mencionado, número de vezes que o usuário foi respondido, número de vezes que o usuário mencionou alguém, número de vezes que o usuário respondeu alguém, número de amigos dos seguidores do usuário, número total de tweets dos amigos do usuário e a existência de palavras associadas a spam no nome do usuário. No total, temos 12 atributos de usuário.



(a) Número de tweets respondidos



(b) Número de menções dadas



(c) Idade da Conta

Figura 3.1: Funções de distribuição acumulada de três atributos do usuário.

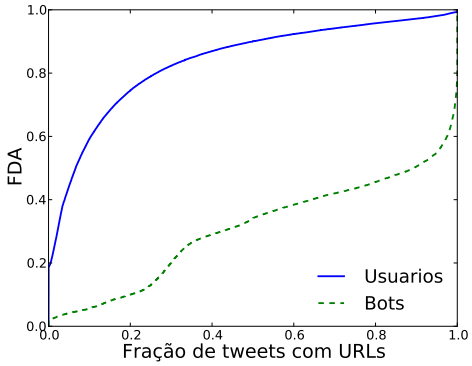
Em seguida, analisamos três características do usuário, que podem diferenciar bots de usuários legítimos. A figura 3.1 mostra a função de distribuição acumulada (FDA) dos três atributos: número de tweets respondidos, número de menções dadas e idade da conta. A partir das figuras 3.1(a) e 3.1(b) notamos que usuários legítimos são mais sociais e ativos em conversas do que bots. Finalmente, a figura 3.1(c) mostra a idade da conta do usuário. Podemos observar que bots tendem a possuir contas mais novas, provavelmente pelo fato de serem bloqueados por outros usuários ou reportados para o Twitter por realizarem atividades ilícitas, e.g., postar links de spam.

3.2.2 Atributos de conteúdo

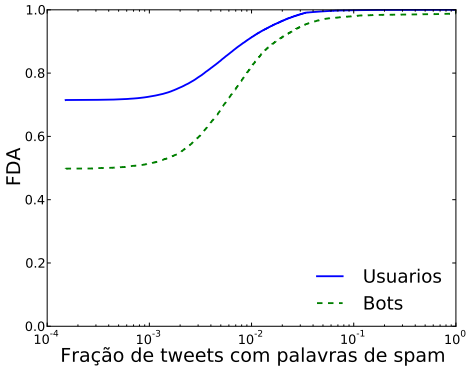
Atributos de conteúdo são baseados em propriedades dos tweets postados pelos usuários, que capturam características específicas relacionadas a forma com que os mesmos escrevem seus tweets. Devido ao fato dos usuários geralmente postarem vários tweets, utilizamos o valor máximo, mínimo, médio e a mediana das seguintes métricas: número de hashtags por palavra em cada tweet, número de URLs por palavra em cada tweet, número de palavras em cada tweet, número de caracteres em cada tweet, número de URLs em cada tweet, número de hashtags em cada tweet, número de caracteres numéricos (e.g. 1,2,3) em cada tweet, número de usuários mencionados em cada tweet, número de vezes que o tweet foi retweetado. Também utilizamos a fração de tweets contendo pelo menos uma palavra relacionada a atividades de spam[1], a fração de mensagens que eram respostas, a fração de mensagens que mencionam um outro usuário, a fração de tweets que contem hashtags, a fração de mensagens que são retweets e a fração de mensagens que contem URLs. Ao todo temos 42 atributos de conteúdo.

A seguir, apresentamos uma análise de três atributos de conteúdo: fração de URLs, fração de tweets com palavras de spam e fração de hashtags. A figura 3.2 mostra as FDAs destes atributos. A figura 3.2(a) mostra que bots postam mais tweets com URLs que usuários legítimos. Contudo, como a figura 3.2(b) indica, bots não são necessariamente spammers, o que aponta que eles possam postar URLs dos mais diversos tópicos (e.g., notícias sobre um determinado tópico). Finalmente, a figura 3.2(c) revela que bots tendem a postar mais hashtags que usuários legítimos, talvez com o intuito de aparecer mais em buscas de determinados tópicos.

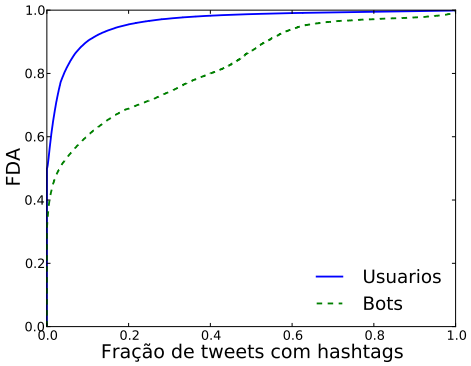
[1] http://codex.wordpress.org/pt-br:Palavras_de_Spam



(a) Fração de tweets com URLs



(b) Fração de tweets com palavras de spam



(c) Fração de tweets com hashtags

Figura 3.2: Funções de distribuição acumulada de três atributos de conteúdo.

3.2.3 Atributos linguísticos

Atributos linguísticos capturam propriedades específicas do padrão de escrita do usuário, visto que usuários que postam mensagens sobre vários tópicos geram conteúdo menos previsível do que aqueles que se restringem a um tópico em particular. Consideramos as seguintes métricas como atributos linguísticos:

- **Tamanho do Vocabulário:** Consideramos o tamanho do vocabulário do usuário, isto é, o número total de palavras diferentes usadas por ele, assim como a razão entre ele e o número de tweets do usuário.

- **N-gramas**: Dado um conjunto de tweets gerados por um usuário, para cada tweet calculamos o número de n-gramas que já foram usados pelo usuário em outros tweets, além da sua razão com o número total de n-gramas já utilizados pelo usuário. Um n-grama é uma sequência contígua de n itens de uma dada sequência de texto, os itens podem ser caracteres, palavras, silabas etc. Um n-

grama de tamanho 1 é conhecido como unigrama, de tamanho 2 como bigrama e de tamanho 3 como trigrama. Usamos a média destes valores como atributos de nosso classificador. Calculamos variações desta métrica usando n-gramas de palavras e caracteres, além de valores de n iguais a 2, 3 e 4.

- **Distância do Cosseno:** Dado um conjunto de tweets gerados por um usuário. Para cada tweet computamos a distância máxima do cosseno Baeza-Yates & Ribeiro-Neto [1999] com o resto dos tweets do usuário. A distância de dois tweets é dada por

$$dist(t_j, q) = \frac{\sum_{i=1}^{N} w_{i,j} w_{i,q}}{\sqrt{\sum_{i=1}^{N} w_{i,j}^2} \sqrt{\sum_{i=1}^{N} w_{i,q}^2}}$$

  Onde, $w_{t,d}$ é o produto da frequência do termo t no tweet d pela frequência inversa do termo nos tweets do usuário. Usamos a média destes valores como atributo no nosso classificador.

- **Índice de Jaccard:** Dado um conjunto de tweets gerados por um usuário. Para cada tweet é computado o máximo índice de Jaccard Tan et al. [2005] com o resto dos tweets postados. O índice de dois tweets é dado por

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

  Usamos a média destes valores como atributo no nosso classificador. Além disso, calculamos três variações do índice usando unigramas, bigramas e trigramas.

- **Modelo de N-gramas**: Dado um conjunto de tweets gerados por um usuário. Calculamos a probabilidade de cada tweet ser gerado pelo usuário usando um modelo de linguagem Manning & Schütze [1999], um modelo estatístico que atribui a probabilidade de uma sequência de m palavras por meio de uma distribuição de probabilidade. Para isso, usamos um modelo de n-grama, no qual a probabilidade $P(w_1, ..., w_m)$ de observar a sequência $w_1, ..., w_m$ é aproximado por

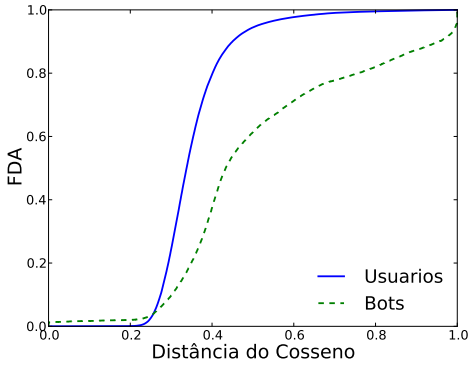
$$P(w_1, ..., w_m) \approx \prod_{i=1}^{m} P(w_i | w_{i-(n-1)}, ..., w_{i-1})$$

  Assumimos que a probabilidade de observar a palavra w_i é dada por apenas as ultimas $n - 1$ palavras, propriedade Markoviana. Dessa forma a probabilidade condicional pode ser calculada a partir da contagem da frequência dos n-gramas nos tweets restantes do usuário.

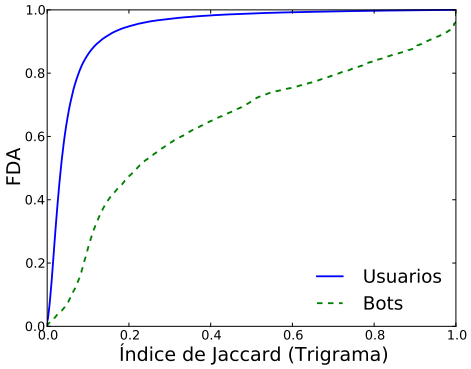
$$P(w_i|w_{i-(n-1)}, ..., w_{i-1}) = \frac{freq(w_{i-(n-1)}, ..., w_{i-1}, w_i)}{freq(w_{i-(n-1)}, ..., w_{i-1})}$$

Para cada usuário usamos a média das probabilidades de cada tweet como atributo no nosso classificador. Calculamos variações desta métrica usando bigramas e trigramas de palavras, além de n-gramas de caracteres para valores de n iguais a 2, 3 e 4.

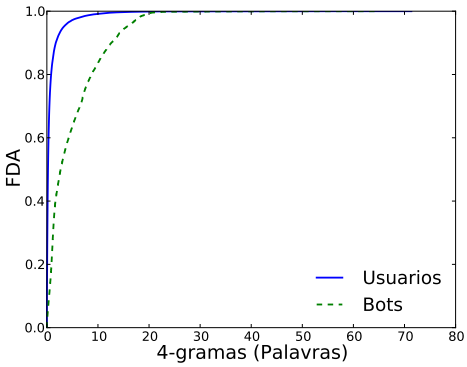
Devido ao custo computacional destas métricas foram analisados apenas os últimos 200 tweets de cada usuário. Ao todo temos 23 atributos linguísticos.



(a) Distância do Cosseno



(b) Índice de Jaccard (Trigrama)



(c) 4-gramas (Palavras)

Figura 3.3: Funções de distribuição acumulada de três atributos linguísticos.

A seguir, realizamos uma analise de três atributos linguísticos: A distância do cosseno, o índice de Jaccard (trigrama) e o 4-gramas (palavras). A figura 3.3 mostra as FDAs desses atributos. Podemos notar que o padrão de escrita dos bots é mais previsível que o dos usuários legítimos, visto que usuários legítimos usam o Twitter

para conversar sobre diversos tópicos, enquanto bots tendem a postar mensagens com foco em um tópico específico.

3.3 Detectando bots

Nesta seção, analisamos o desempenho dos atributos discutidos na seção anterior em conjunto com um algoritmo de aprendizado supervisionado para a tarefa de detectar bots no Twitter. Além disso, apresentamos na seção 3.3.1 as métricas usadas para avaliar os resultados da classificação. A seção 3.3.2 descreve o algoritmo de classificação, ou seja, o classificador, e ambiente experimental utilizado.

3.3.1 Métricas de avaliação

Para avaliar o desempenho de nossa abordagem foram utilizadas as seguintes métricas: precisão, revocação, Micro-F1, Macro-F1 e Área sob a curva ROC (AUC). A revocação(r) de uma classe X é a razão entre o número de usuários corretamente classificados e o número de usuários na classe X. A precisão(p) de uma classe X é a razão do número de usuários corretamente classificados e o número total de usuários previstos como sendo da classe X. Para explicar essas métricas, usaremos uma matriz de confusão, ilustrada na tabela 3.2. Cada uma das posições nesta matriz representa o número de elementos em cada classe original, e como eles foram previstos pelo classificador. Na tabela 3.2, os valores de precisão (p_{bot}) e revocação (r_{bot}) para a classe bot são calculados como $p_{bot} = \frac{a}{(a+c)}$ e $r_{bot} = \frac{a}{(a+b)}$.

Tabela 3.2: Exemplo de Matriz de Confusão

		Previsto	
		Bot	Usuário
Verdadeiro	Bot	a	b
	Usuário	c	d

A medida F1 é a média harmônica entre a precisão e revocação e é definida como $F1 = \frac{2pr}{(p+r)}$. Micro-F1 e Macro-F1 são duas variações da métrica geralmente utilizadas para avaliar a eficácia de um classificador. Micro-F1 é calculada computando os valores globais de precisão e revocação para todas as classes, e em seguida calculando a medida F1. Micro-F1 considera igualmente importante a classificação de cada usuário, independentemente de sua classe. Esta métrica basicamente mede a capacidade do classificador de prever corretamente a classe de um usuário. De forma contrária,

Macro-F1 é calculado computando primeiro os valores F1 para cada classe de forma isolada, e posteriormente calcular a média destes valores. Macro-F1 considera igualmente importante a eficácia do classificador em cada classe, independentemente do tamanho relativo da classe no conjunto. Desta forma, essas métricas fornecem avaliações complementares da efetividade de um classificador. Finalmente, também foi usada a Área sob a curva ROC que mede a capacidade discriminativa do classificador.

3.3.2 Classificador e ambiente experimental

Nos nossos experimentos utilizamos o classificador *Random Forest* Breiman [2001], visto que ele foi o que apresentou o melhor desempenho dentre os classificadores testados, dessa forma reportamos apenas seus resultados. A implementação utilizada em nossos experimentos é encontrada na biblioteca Scikit da linguagem de programação Python.[2] Todos os experimentos de classificação são realizados usando validação cruzada com 20 partições. Em cada teste, separamos nosso conjunto de dados em 20 amostras disjuntas, das quais uma é usada como teste e o restante como treino para nosso classificador. O processo é repetido 20 vezes, de forma que cada amostra é usada exatamente uma vez como teste. Isso gera 20 resultados diferentes, finalmente, reportamos os valores médios.

3.3.3 Resultados da classificação

A tabela 3.3 mostra a matriz de confusão obtida em nossos experimentos. Os números apresentados são as porcentagens relativas ao total de contas em cada classe. Aproximadamente 92% dos bots e 99% dos usuários foram classificados corretamente. Desta forma, apenas uma pequena fração - menos de 1% - de usuários foi erroneamente classificado.

Tabela 3.3: Matriz de Confusão

		Previsto	
		Bot	Usuário
Verdadeiro	Bot	**92.67**%	7.33%
	Usuário	0.94%	**99.06**%

Uma pequena fração (mais de 7%) dos bots foram classificados erroneamente como usuários legítimos. Após uma inspeção manual, percebemos que esses bots tendem a

[2] http://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

postar poucas URLs e hashtags, além de postarem tweets contendo citações. Este comportamento engana alguns aspectos importantes usados pelo classificador para diferenciar bots de usuários legítimos. Além disso, analisamos uma amostra dos usuários que foram classificados como bots. Notamos que esses usuários geralmente são bots cujo padrão temporal de postagem não foi detectado pelo algoritmo de detecção de atividade automática (e.g., contas que postam as noticias de um blog). Dessa forma mostrando que nossa abordagem consegue detectar bots com base em padrões mais complexos.

3.3.4 Importância dos atributos

Para medir a importância dos atributos calculamos o ganho de informação, isto é a redução esperada na entropia, de cada um dos mesmos. A tabela 3.4 apresenta o ranking com os 20 atributos mais importantes segundo esta métrica.

Tabela 3.4: Ranking dos 20 melhores atributos

Posição	Atributo
1	Idade da conta
2	Fração de tweets com URLs
3	Número de URLs por tweet (média)
4	Índice de Jaccard (Trigramas)
5	Índice de Jaccard (Bigramas)
6	Índice de Jaccard (Unigramas)
7	4-gramas (Palavras)
8	URLs por palavra (média)
9	Trigramas (Palavras)
10	Fração de respostas
11	Número de Amigos
12	Fração de mensagens que mencionam um usuário
13	URLs por palavra (media)
14	Número de menções por tweet (média)
15	Número de URLs por tweet (mediana)
16	Trigramas relativo (Palavras)
17	Número de dígitos por tweet (mediana)
18	Número de tweets dos amigos do usuário
19	Bigramas (Palavras)
20	número de mensagens respondidas

Entre os primeiros atributos do ranking temos a fração de tweets contendo URLs e o número médio de URLs por tweet, o que indica que bots postam links com maior frequência que os usuários legítimos (e.g., bots que postam links de notícias ou spam).

Além disso, podemos notar que os atributos linguísticos apresentam um grande poder discriminativo, apesar de serem redundantes, isso revela que apesar de todas as limitações do Twitter os padrões linguísticos de seus usuários são bons atributos para detecção de bots. Finalmente, podemos notar que bots são geralmente associados a contas mais novas.

Tabela 3.5: Número de atributos nas posições do topo do ranking

	Usuário	Conteúdo	Linguísticos
Top 10	1	4	5
Top 20	3	9	8
Top 30	8	12	10
Top 40	8	19	13
Top 50	9	24	17
Top 60	9	30	21
Top 70	10	37	23
Top 77	12	42	23

A tabela 3.5 apresenta um resumo dos resultados, mostrando número de atributos de cada conjunto (usuário, conteúdo e linguísticos) no top 10, 20, 30, 40, 50, 60, 70 e 77 atributos mais discriminativos de acordo com o ranking de ganho de informação. Como podemos notar os atributos de conteúdo são os mais significativos no topo do ranking, seguidos pelos atributos linguísticos o que confirma que a estrutura dos tweets e o padrão de escrita do usuário são atributos fortemente discriminativos na detecção de bots.

3.3.5 Redução do conjunto de atributos

De forma similar a detecção de spammers no Twitter, a detecção de bots é uma constante luta entre os mecanismos de detecção de bots e seus criadores. Dessa forma, esperamos que novos bots sejam mais difíceis de ser detectados por estratégias atuais de detecção. Portando, a importância dos atributos pode variar com o tempo, isto é, atributos importantes hoje podem se tornar pouco discriminativos. De modo que é importante que diferentes conjuntos de atributos possam ser usados para obter resultados de classificação precisos.

Com essa finalidade, computamos os resultados utilizando os diferentes conjuntos de atributos: do usuário (U), de conteúdo (C) e linguísticos (L), assim como a combinação dos mesmos. A tabela 3.6 apresenta o desempenho do classificador usando diferentes conjuntos de atributos.

Tabela 3.6: Resultados de nosso classificador

Atributos	Micro F1	Macro F1	AUC
L	0.954	0.916	0.976
U	0.971	0.948	0.985
C	0.964	0.936	0.982
L+U	0.977	0.960	0.991
U+C	0.978	0.962	0.991
L+C	0.973	0.951	0.987
L+U+C	0.980	0.969	0.992

Apesar dos atributos do usuário não serem individualmente os mais discriminativos, em conjunto foram os que apresentaram os melhores resultados nos nossos testes, o que pode ser explicado pelo fato que estes atributos são pouco redundantes entre si. De forma similar, os atributos linguísticos e de conteúdo por apresentarem grande redundância entre si apresentam desempenho inferior. Finalmente, a combinação de qualquer conjunto de atributos melhora os resultados de nosso classificador, atingindo o seu melhor desempenho quando todos os conjuntos são utilizados.

Capítulo 4

Infiltração na rede de usuários do Twitter

Neste capítulo, realizamos um estudo sobre a vulnerabilidade do Twitter a ataques de socialbots, além de investigarmos quais características tornam socialbots mais bem sucedidos em tarefas de infiltração no Twitter.

Enquanto outros estudos demostraram que socialbots podem se infiltrar com uma taxa de sucesso de até 80% em outras redes sociais, poucos estudos analisam o desempenho dos mesmos em tarefas de infiltração no Twitter.

Finalmente, este trabalho realiza um estudo complementar a estudos anteriores que visam detectar quais características dos usuários tornam-os suscetíveis a ataques de socialbots. Para isto, investigamos quais atributos e comportamentos tornam socialbots mais populares na rede, essas características foram obtidas a partir da aplicação de engenharia reversa nos atributos do classificador proposto no capítulo anterior.

4.1 Metodologia

Uma tarefa de infiltração possui como objetivo promover a interação de usuários-alvo na rede do Twitter com um ou mais socialbots. Um usuário-alvo pode interagir com um socialbot por meio das seguintes ações: (i) seguir o socialbot, (ii) retuitar um tweet postado pelo mesmo, (iii) mencionar o socialbot em algum tweet e, finalmente, (iv) responder a um tweet postado pelo mesmo. Para isso, foram criados 120 socialbots no Twitter. Durante 30 dias monitoramos seu comportamento e todas suas interações com usuários da rede.

Devido ao alto custo de analisar todas as possíveis variações de comportamento, este estudo é um passo inicial a fim de compreender se alguma característica pode tornar

um bot mais bem sucedido em tarefas de infiltração. Dessa forma, este trabalho visa (a) medir a vulnerabilidade de usuários do Twitter a socialbot, e (b) se o desempenho dos socialbot em tarefas de infiltração pode ser influenciado por fatores como:

- Delimitar o grupo de usuários-alvo, e.g. usuários que falem de um tema.
- Ter conhecimento sobre a rede dos usuários-alvo.
- O gênero do bot.
- O nível de atividade do bot.
- O método utilizado pelo bot para gerar os tweets.

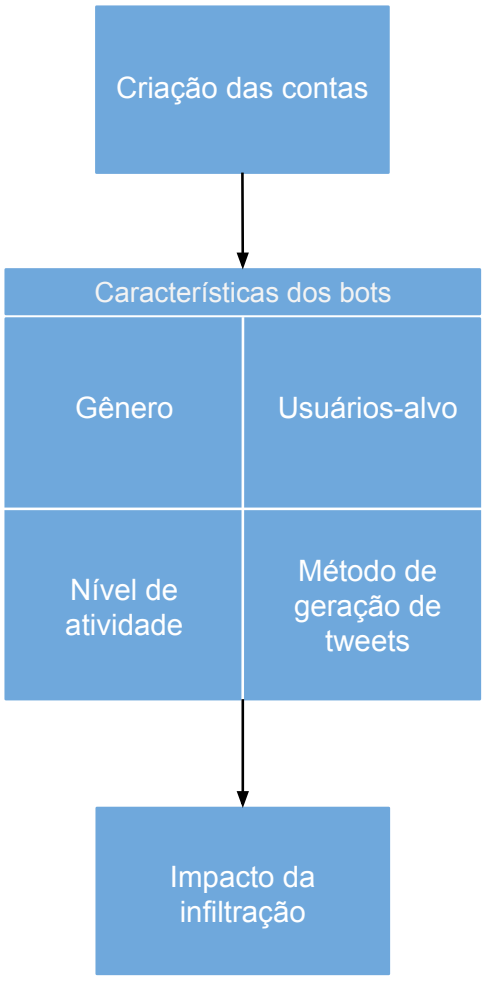


Figura 4.1: Passos do experimento de infiltração.

A figura 4.1 apresenta a metodologia utilizada no experimento de infiltração. Inicialmente detalhamos o processo de criação das contas utilizadas, o que envolve a configuração dos perfis no Twitter e a implementação dos bots posteriormente detalhamos as características e comportamentos adotados por nossos bots e, então apresentamos os resultados do nosso experimento de infiltração. Finalmente, apresentamos uma discussão sobre os resultados obtidos. A seguir descrevemos de forma detalhada cada um desses passos:

4.1.1 Criação das Contas

A seguir, detalhamos o processo de configuração dos perfis utilizados por nossos bots, a criação dos mesmos e, finalmente, as dificuldades encontradas durante nossos experimentos.

4.1.1.1 Configuração dos perfis

Para aumentar o apelo de nossos bots para usuários do Twitter, realizamos alguns passos no seu processo de criação:

- Personalização do perfil dos bots, isto é, cada bot possui um nome, biografia, foto de perfil e um plano de fundo. Dessa forma o perfil de nossos bots torna-se similar ao de usuários legítimos da rede.
- Todos os bots designados ao mesmo grupo de usuários-alvo seguem uns aos outros, dessa forma evitando que nossas contas não possuam seguidores.
- Extraímos as contas mais seguidas pelo grupo de usuários-alvo designado aos bots, e cada um dos bots segue entre uma e sete dessas contas selecionadas aleatoriamente.
- Finalmente, antes de nossos bots realizarem qualquer interação com usuários da rede eles devem postar pelo menos 10 tweets. Dessa forma, quando um usuário-alvo analisar o perfil de um de nossos bots não encontrará um perfil totalmente "vazio"

4.1.1.2 Criação dos Bots

Para que bots possam se passar por usuários legítimos é necessário que os mesmos interajam com o resto da rede. Dessa forma, nossos bots podem executar um conjunto de ações para essa finalidade: (i) postar tweets, (ii) retuitar tweets de usuários que

eles sigam e (iii) seguir usuários no Twitter. Nossos bots só seguem usuários de seus respectivos grupos-alvo e usuários que os tenham seguido.

De forma mais especifica, em intervalos aleatórios nossos bots possuem igual probabilidade de postar um novo tweet ou de retuitar um tweet existente. Além disso, toda vez que uma das ações anteriores é realizada, o bot também segue um número aleatório, entre 1 e 5, de usuários-alvo e todos os novos usuários que os tenham seguido desde a ultima ação. Para evitar que nossos bots participassem de atividades de link farm, eles só seguem usuários não-alvos se eles possuem uma quantidade de seguidores maior que a metade do número de amigos, desta forma evitando seguir usuários que possam estar envolvidos em atividades ilegais (e.g., spam, phishing e link farm).

Finalmente, os bots foram implementados utilizando como base o projeto open-source Realboy (Coburn & Marra [2008]) com algumas modificações.

4.1.2 Configuração dos Bots

Para responder nossas questões de pequisa previamente apresentadas, criamos bots cujo comportamento é definido por quatro características a fim de medir o impacto das mesmas na tarefa de infiltração proposta. A seguir, apresentamos essas características e sua distribuição nos 120 bots criados.

4.1.2.1 Gênero

Para medir o impacto no gênero de nossos bots criamos várias contas de cada tipo. Para isso, utilizamos o nome da conta e sua foto de perfil e criamos 60 bots de cada gênero.

4.1.2.2 Nível de atividade

Esta característica visa responder se bots mais ativos são mais bem sucedidos em tarefas de infiltração. Enquanto bots que postem pouco conteúdo são mais difíceis de serem detectados, também tem menos probabilidade de postarem conteúdo relevante que possa atrair novos seguidores. Para tornar nossa analise mais simples criamos bots com apenas dois níveis de atividade:

- **Muito Ativos:** Estes bots possuem intervalo de até 60 minutos entre suas ações, o intervalo é escolhido de forma aleatória e vária entre 1 minuto e 60 minutos. Ao todo metade dos bots possuem este nível de atividade.

- **Pouco Ativos:** Estes bots possuem intervalo de até 120 minutos entre suas ações, o intervalo é escolhido de forma aleatória e vária entre 1 minuto e 120 minutos. Dessa forma, metade dos bots criados possuem este nível de atividade.

Além disso, nossos bots "dormem" entre 22h e 9h, fuso horário do pacífico, dessa forma simulando os períodos de inatividade esperados de usuários humanos.

4.1.2.3 Método de geração de Tweets

Para tornar um bot bem sucedido em tarefas de infiltração é necessário que o mesmo seja capaz de postar conteúdo considerado relevante pelos seus usuários-alvo. Desta forma, o desafio é criar tweets com conteúdo relevante e bem escritos. A seguir apresentamos dois tipos de abordagens para a geração de tweets:

- **Repostagem:** Como o nome indica este método consiste em postar um tweet criado por outro usuário como se fosse de autoria própria. Para aumentar as chances de que o tweet possua conteúdo relevante extraímos as 20 palavras mais usadas pelos usuários-alvo do bot e procuramos um tweet que contenha pelo menos um desses termos. Apesar de simples e eficiente este método pode gerar tweets muito genéricos, visto que os termos mais usados por um grupo contém termos pouco discriminativos (e.g., "people", "day", "happy").

- **Gerar tweets sintéticos:** Esta abordagem gera tweets a partir de um conjunto de exemplo. A abordagem proposta neste trabalho utiliza um gerador markoviano. Para isso, inicialmente extraímos a probabilidade empírica de cada trigrama presente no conjunto de exemplo, posteriormente geramos uma cadeia de markov a partir do conjunto de trigramas obtidos e, finalmente, geramos um tweet aleatório usando esta cadeia. Foram utilizados trigramas porque apresentaram os melhores resultados quando comparados a n-gramas de outra ordem. Para aumentar as chances de que o tweet gerado seja considerado relevante pelos usuários-alvo usamos os seus tweets como conjunto de exemplo.

  A seguir, a figura 4.2 apresenta um exemplo de uma cadeia de markov usando bigramas, extraída a partir do conjunto de exemplos "I like turtles", "I like rabbits" e "I don't like snails". Um possível tweet gerado por esta cadeia é "I don't like rabbits".

  A principal vantagem deste método é que ele não exige nenhum tipo de esforço humano, além de conseguir gerar tweets que contenham os termos representativos da coleção de exemplo, dessa forma gerarando tweets sobre temas do interesse

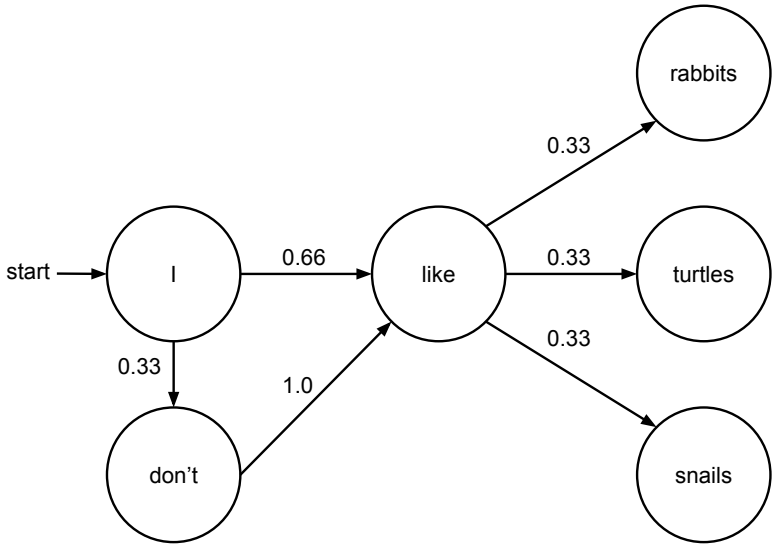


Figura 4.2: Exemplo de cadeia de markov usando bigramas.

do grupo-alvo. Contudo, a qualidade textual dos tweets pode ser baixa (e.g., alguns tweets podem ser sentenças inacabadas). Além disso, devido a forma que o método foi implementado ele é incapaz de gerar tweets contendo mentions e URLs.

A seguir, apresentamos alguns tweets gerados pelo nosso gerador:

- *I don't have an error in it :)*
- *The amount of content being published this week :: the number of people who ' ve finished this website but it makes it easier to argue that*
- *Why isn't go in the morning! night y ' all*
- *Night y ' all ???!*
- *"take me to fernandos and you'll see*
- *"end aids now, the marilyn chambers memorial film festival I ' d fix health care continues to outpace much of nation's issues move to the*

Finalmente, metade de nossos bots usam apenas o método de repostagem, enquanto que a outra metade utiliza ambos os métodos, onde cada método tem a mesma probabilidade de gerar o próximo tweet.

4.1.2.4 Usuários-alvo

Para medir o desempenho de bots em tarefas de infiltração é necessário que eles possuam um conjunto de usuários-alvo, isto é, usuários com os quais os bots pretendam interagir

de alguma forma. Definimos um usuário-alvo como sendo um usuário do Twitter que possua as seguintes características: (i) seja controlado por um humano, (ii) que poste tweets em inglês, para garantir que entendessem o idioma usado por nossos bots e, finalmente, (iii) que tenha postado pelo menos um tweet no mês de Dezembro de 2013, desta forma evitamos usuários inativos. Para garantir essas propriedades, todas as contas foram manualmente verificadas. Além disso, para responder nossas duas primeiras questões de pesquisa criamos três diferentes grupos de usuários-alvo. A seguir, detalhamos cada grupo de usuário-alvo usados em nossos experimentos:

- **Grupo 1:** composto por 200 usuários obtidos de forma aleatória no Twitter. Dessa forma poderemos medir o desempenho de nossos bots em grupos heterogêneos.

- **Grupo 2**: composto por 200 usuários que postam tweets sobre um tópico específico, nosso foco foi em um grupo de desenvolvedores. Para isto, selecionamos usuários que tenham postado pelo menos um tweet contendo algum dos termos "jQuery", "javascript" ou "nodejs". Posteriormente, selecionamos manualmente 200 contas que atendessem o critério previamente descrito.

- **Grupo 3**: composto por 200 usuários que postam tweets sobre um tópico específico − novamente focamos em um grupo de desenvolvedores, e, que além disso, possuam relações de amizade entre si. Para isso, usamos um usuário semente e coletamos sua rede de amigos, a partir da qual extraímos 200 usuários cujos perfis atendessem as restrições previamente apresentadas. Para isso selecionamos manualmente um grupo de desenvolvedores que formem uma comunidade, isto é, cujas relações de amizades formem um grafo denso. Utilizamos como semente o usuário @jeresig, visto que é um desenvolvedor muito influente no Twitter.

Esta característica visa medir o desempenho de nossos bots ao invadir cada grupo de usuários-alvo previamente descritos, isto é, se as características do grupo de usuários-alvo possui alguma influência. Dessa forma, 40 bots foram designados a cada grupo de usuários-alvo.

A seguir, realizamos uma breve caracterização de cada grupo de usuários-alvo. A figura 4.3 mostra as nuvens de tags com os 30 termos mais usados por cada grupo. Como esperado a nuvem dos dois últimos grupos apresentam termos como "code", "data", "app", e "web" que são tipicamente usados por desenvolvedor. Enquanto isso, o primeiro grupo tende a usar termos do Twitter como "via", "unfollowers" e "followed", além de termos pouco específicos.

(a) Grupo 1

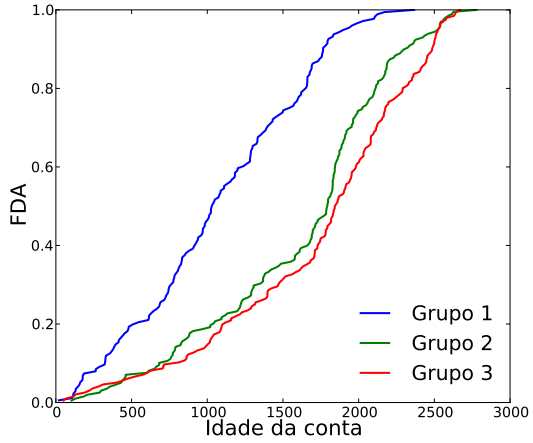
(b) Grupo 2

(c) Grupo 3

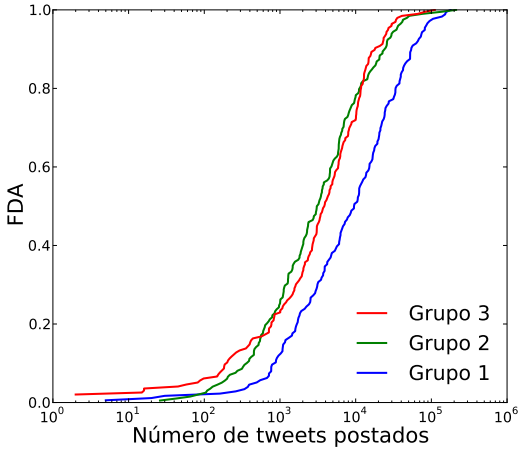
Figura 4.3: Nuvem de tags com os 30 termos mais usados por cada grupo.

Em seguida, analisamos quatro características dos grupos de usuários-alvo. A figura 4.4 mostra a função de distribuição acumulada (FDA) dos quatro atributos: idade da conta, número de tweets postados, número de seguidores e *Klout Score*. A figura 4.4(a) apresenta a idade das contas de cada grupo, notamos que usuários do grupo 1 possuem contas mais novas que dos outros grupos, apesar disso, como a figura 4.4(b) demostra estes usuários têm um maior número de tweets postados, o que é um indicio que o grupo possui um maior nível de atividade no Twitter. Finalmente, as figuras 4.4(c) e 4.4(d) mostram o número de seguidores e o *Klout Score*, respectivamente. O

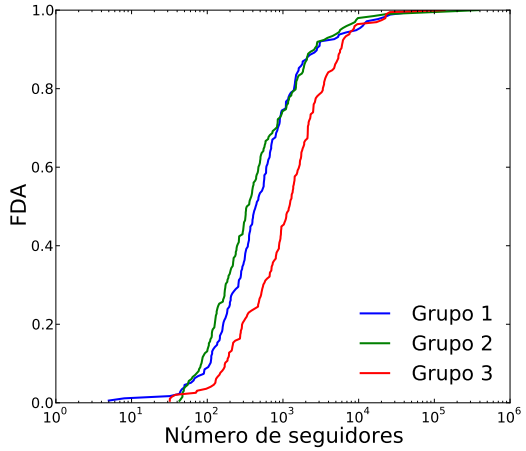
Klout[1] é um dos principais sistemas de medição de influência utilizados atualmente, para isso o sistema utiliza abordagens de medições de influência e cujos detalhes não são revelados ao público. O sistema atribui uma nota entre 0 e 100, de forma que usuários com altos valores de *Klout Score* são considerados influentes. Como podemos notar usuários do grupo 3 são mais influentes na rede que os outros grupos, isto é, possuem um maior número de seguidores e altos valores de *Klout Score*.



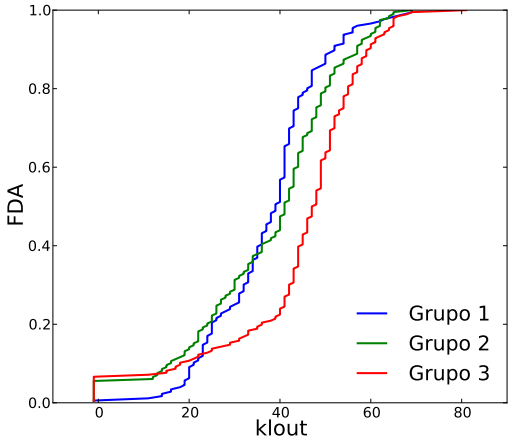
(a) Idade da Conta



(b) Número de tweets postados



(c) Número de seguidores



(d) Klout score

Figura 4.4: Funções de distribuição acumulada de quatro atributos de cada grupo.

[1]http://klout.com/

4.2 Medindo o desempenho de Infiltração

O objetivo deste estudo é investigar se, e em que medida, várias estratégias tornam socialbots capazes de se infiltrar na rede social do Twitter. Naturalmente, é necessário utilizar métricas para quantificar o desempenho de infiltração de socialbots, de modo que o desempenho de diferentes estratégias (utilizadas pelos socialbots) possam ser comparadas. Para quantificar o desempenho de infiltração usamos as três seguintes métricas, medidas no final do período do experimento:

(1) Seguidores adquiridos pelo socialbot: Contamos o número de seguidores adquiridos pelo socialbot, que é uma métrica padrão para estimar a popularidade/influência dos usuários na rede social do Twitter (Cha et al. [2010]).
(2) *Klout Score* adquirido pelo socialbot: *Klout Score*[2] é uma métrica popular para medir a influência social online de um usuário. Embora o algoritmo exato para a métrica não é conhecido publicamente, o *Klout Score* para um determinado usuário é conhecido por considerar vários dados do Twitter (e outras redes sociais on-line, se disponível), tais como o número de seguidores e seguidores do usuário, retweets, quantos spammers/contas mortas estão seguindo o usuário, quão influentes são as pessoas que retweetam/mencionar o usuário, e assim por diante[3]. Valores de *Klout Score* variam de 1 a 100, onde uma maior pontuação implica que o usuário possui uma influência social online mais elevada.
(3) Interações baseadas em mensagens com outros usuários: Medimos o número de vezes que outros usuários interagiram com um socialbot através das mensagens (tweets) postadas na rede social. Consideramos os diferentes tipos de interações baseadas em mensagens permitidas no Twitter, especificamente contamos o número total de vezes que algum usuário @menciona o bot, ou responde algum tweet do bot, ou retuita/favorita um tweet postado pelo bot. Essa métrica estima o engajamento social do bot, que é definida como a medida em que o usuário participa de uma ampla gama de papéis e relações sociais (William R. Avison & [Eds.]).

Se um bot pontua bem em relação as métricas acima, isso implica que os tweets postados por este bot são mais propensos a serem visíveis, *e.g.*, mais susceptíveis de serem incluídos nos resultados de busca do Twitter, e portanto mais susceptíveis de afetar a opinião de outros usuários (que são objetivos comuns de socialbots em redes sociais).

As seções subsequentes medem o sucesso de várias estratégias de socialbots em

[2]http://klout.com/
[3]http://en.wikipedia.org/wiki/Klout

se infiltrar na rede social de acordo com as métricas especificadas acima.

4.3 Socialbots podem infiltrar a rede do Twitter?

Nós primeiro investigamos se, e em que medida, socialbots podem se infiltrar na rede do Twitter. Para uma socialbot poder se infiltrar com sucesso na rede, ele precisa alcançar os seguintes dois objetivos: (i) evitar a detecção por mecanismos de defesa do Twitter que verificam regularmente e suspendem contas que apresentam atividade automatizada (twitter-shut-spammers [2012]), e (ii) adquirir um nível substancial de popularidade e influência na rede social, além de interagir com um grande número de usuários, ou seja, atingir altas pontuações nas métricas descritas na seção 4.2. Nesta seção, investigamos o desempenho dos socialbots com respeito aos objetivos acima.

4.3.1 Socialbots podem evadir os mecanismos de defesa?

Começamos verificando quantos dos 120 socialbots foram detectados pelo mecanismo de segurança do Twitter. Notamos que ao longo dos 30 dias em que o experimento foi realizado, 38 dos 120 socialbots foram suspensos. Isto implica que, apesar de todos os nossos socialbots ativamente postarem tweets e seguirem outros usuários durante este período, apenas 31% dos socialbots foram detectados pelos mecanismos de defesa do Twitter.

A seguir, analisamos qual dos 120 socialbots foram detectados pelo Twitter. a figura 4.5 mostra a distribuição dos quatro atributos – sexo, nível de atividade, método de postagem e grupo de usuários-alvo seguidos – entre os 120 socialbots criados. Os socialbots são indicados por identificadores numéricos na mesma ordem em que eles foram criados, ou seja, o Bot 1 foi criado primeiro e Bot 120 foi o último socialbot criado. Os socialbots que foram suspensos pelo Twitter durante o experimento, são indicados na cor vermelha, enquanto que os socialbots que não foram detectados pelo Twitter são mostrados na cor azul.

Notamos que a grande maioria dos socialbots que foram suspensos foram os que foram criados no final do processo de criação de contas (com IDs de entre 90 e 120). Isto é provavelmente porque no momento em que essas contas foram criadas, o mecanismo de defesa do Twitter tornou-se suspeito de que várias contas foram criadas a partir do mesmo bloco de endereços de IP[4]. Notamos também que os socialbots que usaram o método de postagem baseado em Markov foram mais propensos a serem suspensos.

[4]Usamos 12 endereços de IP diferentes para criar os 120 socialbots, ou seja, 10 contas foram operados a partir de cada endereço IP.

Grupo 1		Grupo 2		Grupo 3	
Masculino	Feminino	Masculino	Feminino	Masculino	Feminino
Bot 1	Bot 2	Bot 3	Bot 4	Bot 5	Bot 6
Bot 7	Bot 8	Bot 9	Bot 10	Bot 11	Bot 12
Bot 13	Bot 14	Bot 15	Bot 16	Bot 17	Bot 18
Bot 19	Bot 20	Bot 21	Bot 22	Bot 23	Bot 24
Bot 25	Bot 26	Bot 27	Bot 28	Bot 29	Bot 30
Bot 31	Bot 32	Bot 33	Bot 34	Bot 35	Bot 36
Bot 37	Bot 38	Bot 39	Bot 40	Bot 41	Bot 42
Bot 43	Bot 44	Bot 45	Bot 46	Bot 47	Bot 48
Bot 49	Bot 50	Bot 51	Bot 52	Bot 53	Bot 54
Bot 55	Bot 56	Bot 57	Bot 58	Bot 59	Bot 60
Bot 61	Bot 62	Bot 63	Bot 64	Bot 65	Bot 66
Bot 67	Bot 68	Bot 69	Bot 70	Bot 71	Bot 72
Bot 73	Bot 74	Bot 75	Bot 76	Bot 77	Bot 78
Bot 79	Bot 80	Bot 81	Bot 82	Bot 83	Bot 84
Bot 85	Bot 86	Bot 87	Bot 88	Bot 89	Bot 90
Bot 91	Bot 92	Bot 93	Bot 94	Bot 95	Bot 96
Bot 97	Bot 98	Bot 99	Bot 100	Bot 101	Bot 102
Bot 103	Bot 104	Bot 105	Bot 106	Bot 107	Bot 108
Bot 109	Bot 110	Bot 111	Bot 112	Bot 113	Bot 114
Bot 115	Bot 116	Bot 117	Bot 118	Bot 119	Bot 120

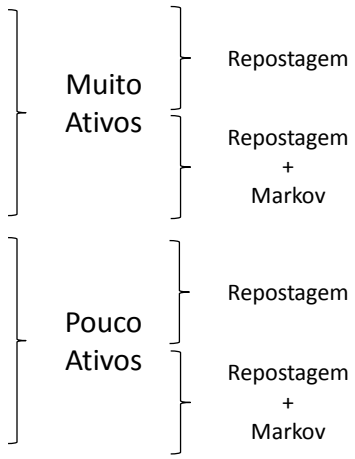


Figura 4.5: Distribuição de atributos dos 120 socialbots criados para o experimento de infiltração, mostrando aqueles socialbots, que foram detectados e suspensos pelo Twitter durante o experimento (mostrados na cor vermelha). Note-se que 69% dos socialbots (mostrados na cor azul) não foram detectados pelo Twitter.

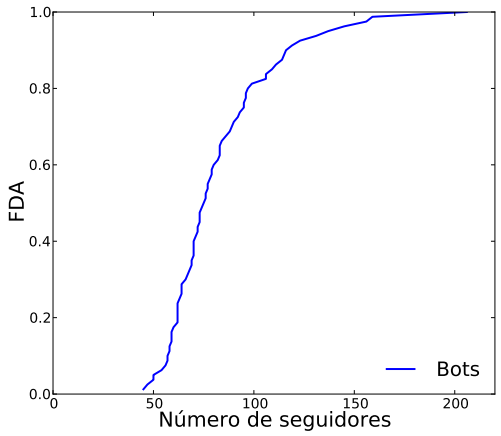
Isto é esperado, uma vez que cerca de metade dos tweets postados por essas contas foram sinteticamente gerados e, portanto, é provável que possuam uma baixa qualidade textual.

No entanto, os mecanismos de defesa do Twitter detectaram apenas uma pequena fração dos socialbots que foram criadas no início, e que adotaram a estratégia de re-postagem, ou seja, re-postaram tweets dos outros usuários. Estes números alertam que os mecanismos de defesa existentes possuem um desempenho limitado na detecção de socialbots que empregam estratégias simples, porém inteligentes para postarem tweets e links no Twitter.

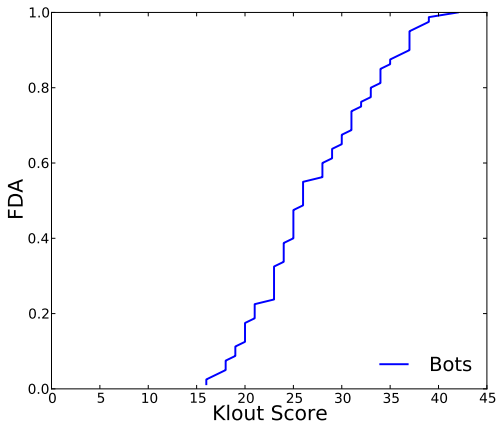
4.3.2 Bots podem se infiltrar no Twitter com sucesso?

A seguir, verificamos em que medida socialbots podem se infiltrar na rede social do Twitter, e se eles podem ganhar valores relativamente altos de popularidade/influência de acordo com as métricas estabelecidas na seção 4.2.

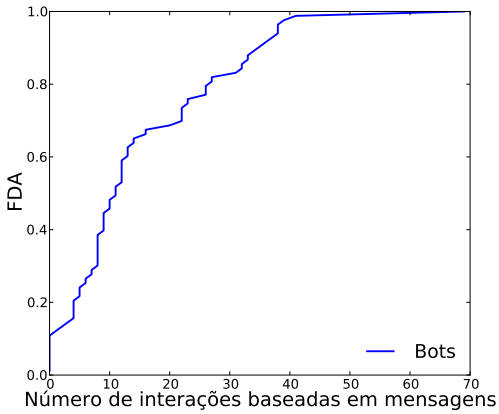
Durante o período do experimento, os 120 socialbots criados foram seguidos no total 4.999 vezes por 1.952 usuários distintos, além de terem recebido 2.128 interações baseadas em mensagens de 1.187 usuários distintos. A figura 4.6 mostra a distribuição do número de seguidores, os valores de *Klout Score* e o número de interações baseadas em mensagem adquiridas pelos socialbots no final do experimento. É evidente que



(a) Número de seguidores



(b) *Klout Score*



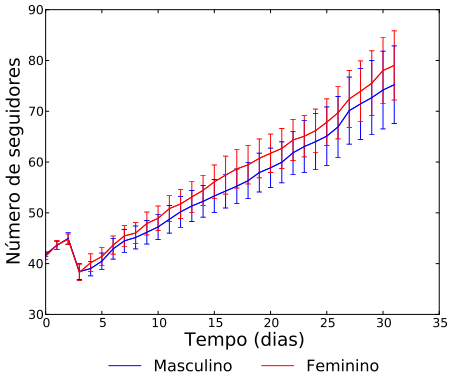
(c) Interações baseadas em mensagens

Figura 4.6: Desempenho de infiltração dos nossos socialbots: FDAs para (i) número de seguidores, (ii) *Klout Score*, e (iii) número de interações baseadas em mensagens com outros usuários.

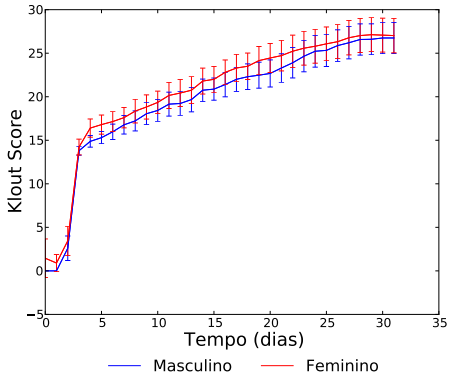
uma fração significativa dos socialbots adquiriram pontuações relativamente altas de popularidade e influência. Dentro de apenas um mês (a duração do experimento), mais de 20% dos socialbots adquiriram mais de 100 seguidores (figura 4.6(a)); apesar que 46% dos usuários do Twitter possuem menos de 100 seguidores (twitter-46pc-lt100followers [2013]). Finalmente, a figura 4.6(b) mostra que 20% dos socialbots adquiriram valores de *Klout Score* superiores a 35 no período de apenas um mês.

4.4 Impacto da Infiltração

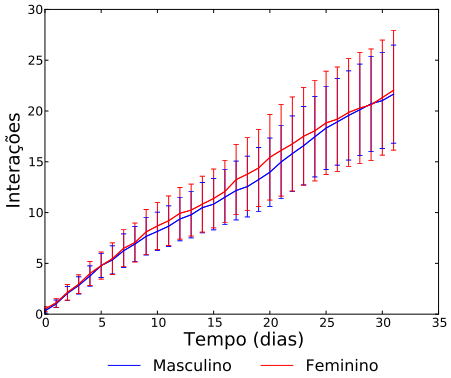
A seção anterior mostrou que uma fração significativa dos socialbots foram realmente capazes de se infiltrar e ganhar popularidade no Twitter. Esta seção analisa quais



(a) Número de seguidores



(b) *Klout Score*



(c) Interações baseadas em mensagens

Figura 4.7: Desempenho de infiltração de socialbots de diferentes gêneros durante a duração do experimento: (i) número médio de seguidores adquiridos, (ii) valor médio de *Klout Score* adquirido, e (iii) número médio de interações baseadas em mensagens com outros usuários. As curvas representam os valores médios e as barras de erro indicam os intervalos de confiança de 95%.

as estratégias usadas pelos socialbot levam a um melhor desempenho de infiltração. Lembre-se que os socialbots foram configurados com várias estratégias para cada um dos quatro atributos – de gênero, nível de atividade, método de postagem e tipo de usuários-alvo (seção 4.1). Agora investigamos quais estratégias para cada um dos quatro atributos produz o melhor desempenho de infiltração. Observe que os resultados estabelecidos nesta seção (e no próxima) consideram apenas socialbots que não foram suspensos pelo Twitter durante o experimento.

4.4.1 Gênero

Começamos analisando o impacto do gênero dos socialbots em nossos experimentos. As figuras 4.7(a) e (b) mostram, respectivamente, a média do número de seguidores e do valor de *Klout Score* adquiridos por socialbots de cada gênero ao longo do experimento. Nestas figuras, as curvas representam os valores médios, considerando todos os socialbots de um gênero particular (num dado dia durante a experiência), e as barras de erro indicam os intervalos de confiança de 95% dos valores médios. Notamos que não há diferença significativa na popularidade adquirida pelos socialbots de diferentes gêneros.

A seguir, analisamos as interações baseadas em mensagens dos socialbots de cada gênero com outros usuários. A figura 4.7(c) mostra o número médio de interações dos socialbots em cada dia durante o experimento. Novamente, observamos que os usuários interagiram quase igualmente com socialbots de ambos os sexos.[5].

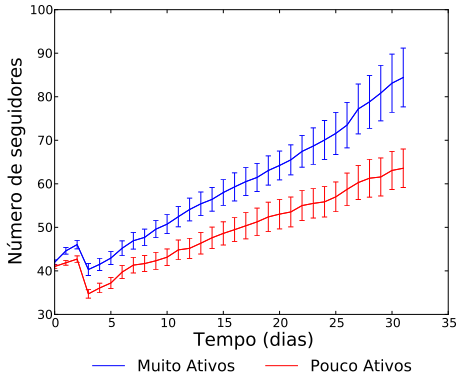
Os resultados acima indicam que o gênero especificado no perfil da conta não influencia significativamente o desempenho dos socialbots em tarefas de infiltração. Note-se que, nesta seção, estamos considerando todos os socialbots e suas interações com todos os grupos usuários-alvo. Posteriormente, na seção 4.5, quando analisarmos separadamente o desempenho de socialbots na infiltração de cada grupo de usuários-alvo, veremos que o gênero do socialbot é de fato significativo para alguns grupos-alvo específicos.

4.4.2 Nível de atividade

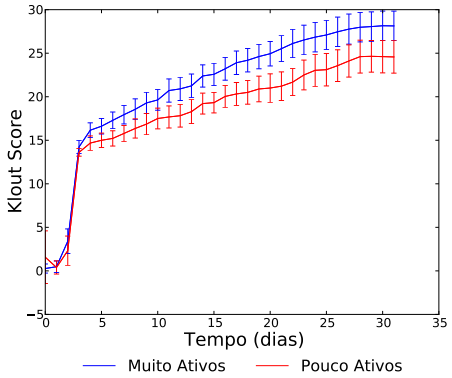
A seguir, estudamos o impacto do nível de atividade dos socialbots, que definimos como muito ou pouco ativos com base no intervalo de tempo entre as atividades realizadas pelos socialbots.

A figura 4.8(a) e (b) mostram, respectivamente, a média do número de seguidores e do valor de *Klout Score* adquiridos por socialbots (com diferentes níveis de atividade) em cada dia durante o experimento. Podemos ver que socialbots mais ativos atingiram significativamente mais popularidade e valores de *Klout Score* do que os socialbots menos ativos. A figura 4.8(c) mostra o número médio de interações baseadas em mensagens de socialbots com outros usuários no Twitter. Novamente, os socialbots mais ativos conseguiram um número muito maior de interações.

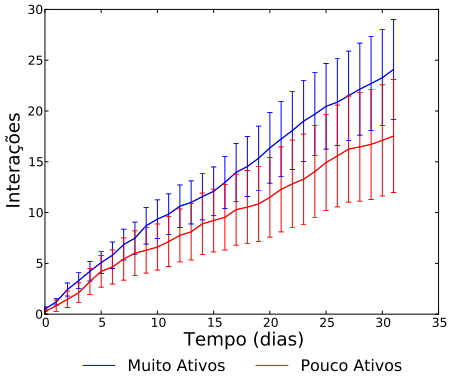
[5]O número de usuários distintos que interagiram com as socialbots femininos (1.697), foi, na verdade, um pouco maior do que o número que interagiu com os socialbots masculinos (1.528). Contudo, como é evidente a partir da figura 4.7(c), esta diferença não é significativa



(a) Número de seguidores



(b) *Klout Score*



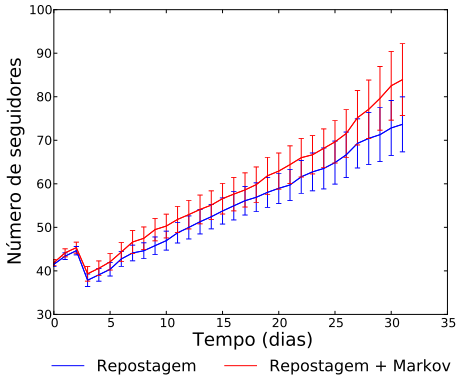
(c) Interações baseadas em mensagens

Figura 4.8: Desempenho de infiltração de socialbots com diferentes niveis de atividade ao longo do experimento: (i) número médio de seguidores adquiridos, (ii) valor médio de *Klout Score* adquirido, e (iii) número médio de interações baseadas em mensagens com outros usuários.

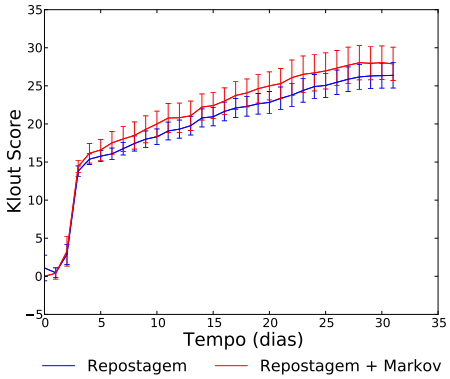
Assim, percebemos que entre mais ativos são os bots, é mais provável que eles se tornem bem sucedidos em tarefas de infiltração, bem como na obtenção de popularidade na rede social. Isto é esperado, uma vez que entre mais ativo um bot é, maior é a probabilidade de que seus tweets sejam vistos por outros usuários. No entanto, também deve notar-se que bots mais ativos, são mais propensos a serem detectados pelos mecanismos de defesa do Twitter.

4.4.3 Método de geração de tweets

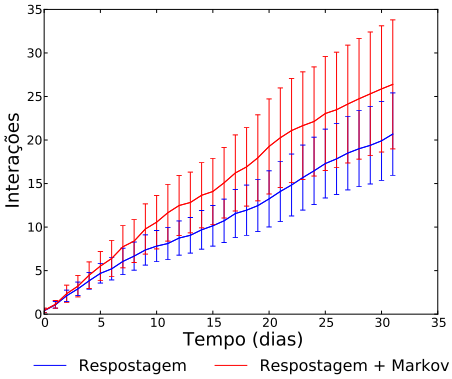
A seguir, analisamos o impacto do método de geração do tweet usado pelos socialbots. Como dito na seção 4.1 metade dos nossos socialbots apenas re-postam os tweets escritos por outros usuários (estratégia indicada como 'repostagem '), enquanto que a



(a) Número de seguidores



(b) *Klout Score*



(c) Interações baseadas em mensagens

Figura 4.9: Desempenho de infiltração de socialbots que utilizam diferentes métodos de postagem ao longo do experimento: (i) número médio de seguidores adquiridos, (ii) valor médio de *Klout Score* adquirido, e (iii) número médio de interações baseadas em mensagens com outros usuários.

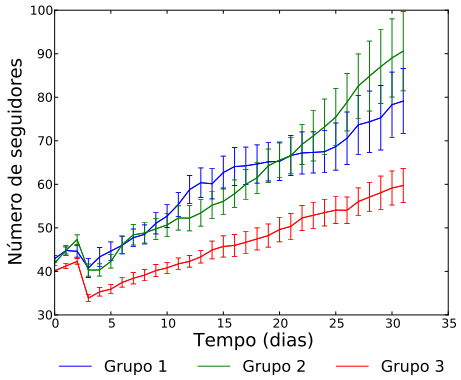
outra metade utiliza o método de repostagem, além de postar tweets sinteticamente gerados usando um gerador de Markov, com igual probabilidade (estratégia denotada como 'repostagem + Markov').

As figuras 4.9(a), (b) e (c) mostram, respectivamente, a média do número de seguidores, a média dos valores de *Klout Score*, e o número médio de interações baseadas em mensagens adquiridas pelos socialbots empregando as duas estratégias de postagem (em cada dia durante o experimento). Vê-se que os socialbots empregando o método 'repostagem + estratégia Markov' adquiriram níveis ligeiramente mais elevados de popularidade (número de seguidores e pontuação Klout), e uma maior quantidade de interações (engajamento social) com outros usuários.

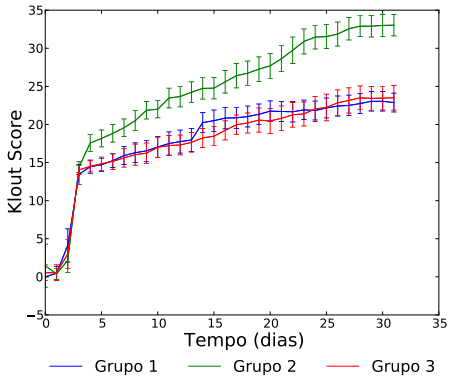
O fato que os socialbots que geraram automaticamente cerca de metade dos seus

tweets terem alcançado um maior engajamento social é surpreendente, uma vez que indica que os usuários de Twitter não são capazes de distinguir entre (contas que postam) Tweets gerados por humanos e tweets gerados automaticamente utilizando modelos estatísticos simples. Isto é possivelmente porque uma grande fração dos tweets no Twitter são escritos em um estilo gramaticalmente incoerente e informal (Kouloumpis et al. [2011]), de modo que até mesmo modelos estatísticos simples podem produzir tweets com qualidade semelhante aos postado por seres humanos no Twitter.

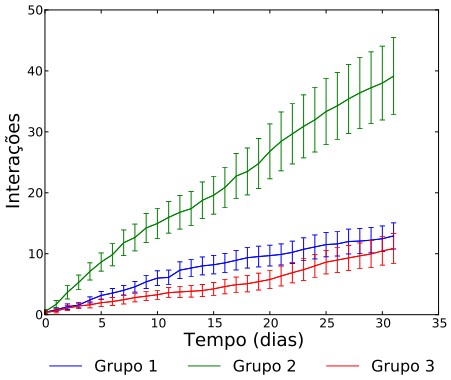
4.4.4 Usuários-alvo



(a) Número de seguidores



(b) *Klout Score*



(c) Interações baseadas em mensagens

Figura 4.10: Desempenho de infiltração de socialbots que seguem diferentes grupos de usuários-alvo ao longo do experimento: (i) número médio de seguidores adquiridos, (ii) valor médio de *Klout Score* adquirido, e (iii) número médio de interações baseadas em mensagens com outros usuários.

Finalmente, analisamos o desempenho de infiltração dos socialbots que foram designados à seguir diferentes conjuntos de usuários-alvo. Na seção 4.1 reportamos

que os socialbots foram divididos em três grupos-alvo diferentes – O Grupo 1 seguiu usuários selecionados aleatoriamente, o Grupo 2 seguiu os usuários-alvo que postam os tweets sobre um tema específico (no caso desenvolvimento de software) e o Grupo 3 de socialbots seguiu usuários-alvo que além de postar tweets sobre o tema específico (desenvolvimento de software), também são socialmente bem relacionados entre si.

A figura 4.10(a) mostra o número médio de seguidores adquiridos por cada grupo de socialbots durante todo o experimento. Notamos que os socialbots no Grupo 3 tiveram o menor número de seguidores, enquanto que os do Grupo 2 tiveram um número significativamente maior de seguidores. A figura 4.10(b) mostra os valores médios de *Klout Score* alcançados pelos nossos socialbots ao longo do tempo. Novamente, os socialbots do Grupo 2 obtiveram os maiores valores de *Klout Score*, enquanto que os outros grupos apresentaram um desempenho similar. A figura 4.10(c) mostra o número médio de interações baseadas em mensagens de cada grupo de socialbots (com outros usuários do Twitter) ao longo do tempo. Mais uma vez, vemos que socialbots no Grupo 2 tem um número significativamente maior de interações com outros usuários, e os do Grupo 3 apresentaram o menor número de interações.

Estas resultados levam a algumas observações interessantes. Seguir um conjunto de usuários que postem tweets sobre um tema específico em comum (por exemplo, desenvolvimento de software) é uma abordagem mais promissora do que seguir usuários aleatórios (como feito pelos bots do Grupo 1). No entanto, embora tanto os usuários-alvo do Grupo 2 e do Grupo 3 postem tweets sobre um tema comum, os socialbots no Grupo 2 alcançaram significativamente maior popularidade e engajamento social – isto implica que se infiltrar em grupos de usuários-alvos interconectados (Grupo 3) é muito mais difícil do que se envolver com os usuários sem qualquer relação entre si (Grupo 2). Note-se que esta observação difere daquelas feitas por uma pesquisa semelhante no Facebook (Elyashar et al. [2013]), onde constatou-se que socialbots podem efetivamente se infiltrar nas redes sociais entre membros de organizações específicas.

4.5 Avaliando a Importância dos Atributos

Nesta seção, nosso objetivo é avaliar a importância relativa dos diferentes atributos e estratégias de infiltração de socialbots. Nosso objetivo é quantificar qual a estratégia (ou combinação de estratégias) que possui o maior impacto em decidir como socialbots podem infiltrar-se em grupos específicos de usuários-alvo. Note-se que, diferentemente da Seção 4.4, aqui nós consideramos o desempenho dos socialbots em infiltrar grupos específicos de usuários-alvo.

Utilizamos um experimento fatorial para avaliar o impacto relativo das diferentes estratégias de infiltração. Começamos por descrever brevemente como nós projetamos nossos experimentos e, em seguida, discutimos os resultados obtidos.

4.5.1 Experimento 2^k fatorial

A seguir incluímos uma breve descrição da teoria de um experimento 2^k fatorial (Jain [1991]). Este tipo de experimento é geralmente necessário em cenários com um grande número de fatores, como uma tentativa para reduzir o número de fatores que farão parte do experimento. Particularmente, experimentos 2^k fatorial referem-se a projetos experimentais com k fatores em que cada fator tem o número mínimo de níveis, apenas dois. Como exemplo ilustrativo, suponha um cenário experimental que possui três fatores – memória, disco e CPU de uma máquina – que podem afetar o desempenho de um algoritmo. Suponha agora que cada experimento leva cerca de um dia para ser executado e existem 10 possíveis tipos de memória, 10 tipos de discos, e 10 tipos de CPUs a ser testados. Para a execução de um experimento com todas as possíveis combinações seriam necessários $10 \times 10 \times 10 = 1.000$ dias. Em vez de utilizar todas as combinações possíveis, um projeto 2^k iria considerar dois tipos (geralmente extremos) de memória, dois tipos de disco, e dois tipos de CPUs para comparar, o que resultaria em apenas $2^3 = 8$ dias de experimentos. A teoria dos experimentos fatoriais (Jain [1991]) então, permite estimar o quanto cada fator impacta sobre o resultado final, uma informação importante para ajudar a decidir sobre quais os fatores um experimento deve se concentrar .

Note-se que, de forma diferente do exemplo acima, o nosso objetivo aqui não é reduzir o número de cenários experimentais. Em vez disso, usamos um experimento 2^k fatorial para inferir o quanto um fator – os quais, no nosso caso, correspondem a atributos como gênero, nível de atividade, e método postagem – afetam as diferentes métricas de infiltração.

4.5.2 Experimento fatorial na infiltração de socialbots

O objetivo dos socialbots poderia ser o de se infiltrar em um grupo específico de usuários-alvo. Por isso, consideramos aqui individualmente o sucesso de nossos socialbots na infiltrando de cada um dos três grupos-alvo (que foram descritos na Seção 4.1). Para cada grupo de usuários-alvo, consideramos as três métricas de infiltração detalhadas anteriormente – o número de seguidores adquiridos, o número de interações baseadas em mensagens e os valores de *Klout Score*. Então, para cada métrica e cada

Fator	-1	$+1$
Gênero (**G**)	Feminino	Masculino
Nível de atividade (**A**)	Pouco ativos	Muito ativos
Método de postagem (**P**)	Repostagem	Repostagem+Markov

Tabela 4.1: Fatores utilizados no experimento fatorial para o estudo de infiltração de socialbots.

grupo-alvo, executamos um experimento 2^3 fatorial considerando os atributos e seus valores, conforme descritos na Tabela 4.1, resultando em $3 \times 3 \times 2^3 = 216$ experimentos. Realizamos experimentos que associam 1 ou -1 para as estratégias empregadas por cada atributo. Para todas as configurações experimentais e para cada conjunto de dados foi usada a média de até 5 resultados, que é o número de socialbots criados em cada configuração.

A ideia básica de um modelo fatorial consiste em formular y, no nosso caso o impacto de infiltração, como uma função de um número de fatores e as suas possíveis combinações, tal como definido pela equação 4.1. Aqui, GP, AP, AG, e conta GAP representam todas as combinações possíveis entre os fatores. Por exemplo, os experimentos para 'GP' tenta medir o impacto de uma determinada combinação dos atributos gênero (G) e método de postagem (P) (*e.g.*, 'Feminino e Repostagem', ou 'Masculino e Repostagem + Markov').

$$y \;=\; Q_0 \;+\; \sum_{i \in F} Q_i \cdot x_i \tag{4.1}$$

onde $F = \{G, A, P, GA, GP, AP, GAP\}$ e x_i é definido da seguinte forma.

$$x_G = \begin{cases} \text{-1} & \text{se Feminino} \\ +1 & \text{se Masculino} \end{cases}$$

$$x_A = \begin{cases} \text{-1} & \text{se Pouco ativo} \\ +1 & \text{se Muito ativo} \end{cases}$$

$$x_P = \begin{cases} \text{-1} & \text{se Repostagem} \\ +1 & \text{se Repostagem + Markov} \end{cases}$$

e os valores x_i's para as combinações dos atributos (*e.g.*, AG, GP) são definidas a partir dos valores de x_G, x_A, e x_P seguindo o padrão descrito em Jain [1991].

Na equação acima, Q_i é o desempenho de infiltração (de acordo com uma determinada métrica, como número de seguidores, ou valor de *Klout Score*) quando a estratégia $i \in F$ é aplicada, e Q_0 representa o desempenho médio de infiltração, calculado sobre todos os atributos e suas possíveis combinações. Ao medir empiricamente y de acordo com diferentes combinações de atributos (que, no nosso caso, referem-se

às várias estratégias dos socialbots), podemos estimar os diferentes valores de Q_i e Q_0. Isso nos permite entender quanto cada atributo afeta o desempenho final de infiltração para uma métrica especifica.

Em vez de apresentar resultados para todos os valores possíveis de Q_i, nos concentramos nas variações de Q_i devido a alterações nos atributos (ou suas combinações), o que ajuda a estimar a importância de um determinado fator no resultado final. Como exemplo, se descobrirmos que um fator é responsável por apenas 1% da variação total nos resultados, podemos inferir que este atributo não é importante para a infiltração de socialbots no Twitter.

Como proposto em Jain [1991], a importância dos vários fatores podem ser quantitativamente estimada através da medição da proporção da variação total no resultado final que é explicada por cada fator. Para calcular esta variação, em primeiro lugar consideramos a variação de y (conforme definido pela Equação 4.1) em todas as execuções, e depois calculamos SS_T como a soma da diferença de quadrados entre cada valor medido de y e o valor médio de y. Em seguida, calculamos SS_i como a variação devido apenas às mudanças no fator i, que pode ser calculado de forma semelhante ao SS_T, mas considerando apenas as execuções em que os valores do fator i foram alteradas. Finalmente, calcula-se a fração da variação devido ao fator i como $\frac{SS_i}{SS_T}$. Agora usamos essa métrica para calcular o impacto de cada atributo para as diferentes métricas de infiltração e grupos de usuários-alvo.

4.5.3 Importância dos Atributos

Começamos analisando até que ponto cada um dos atributos impacta o número de seguidores adquiridos pelos socialbots. A tabela 4.2 mostra a variação explicada por cada atributo no número de seguidores adquiridos pelos socialbots de cada um dos grupos-alvo. Notamos que o nível de atividade de um socialbot é o atributo mais importante para o Grupo 1 (usuários aleatórios) de usuários-alvo, sendo responsável por decidir 53,75% do número de seguidores adquiridos por um socialbot. O segundo atributo mais importante é o método de postagem (*i.e.*, técnica usada para gerar os tweets), que responde por 12,44% da variação do número de seguidores. A combinação destes dois atributos (coluna PA na tabela 4.2) leva também a uma variação elevada (cerca de 20%) no número de seguidores.

Observações semelhantes podem ser feitos a partir da tabela 4.3 e da tabela 4.4, que mostra a variação percentual explicada por cada atributo no número de interações baseadas em mensagens (isto é, número de tweets retuitados ou favoritados, número de menções e o número de respostas) e nos valores de *Klout Score*, respectivamente.

	Gênero (G)	Nível de atividade (A)	Método de postagem (P)	GA	GP	AP	GAP
Grupo 1	7,43	**53,75**	12,44	5,20	0,85	**20,10**	0,23
Grupo 2	3,99	**72,65**	2,77	4,38	3,53	2,81	9,87
Grupo 3	**20,52**	**49,27**	2,02	2,40	5,42	12,71	7,66

Tabela 4.2: A variação percentual no número de seguidores explicada por cada tipo de atributo

	Gênero (G)	Nível de atividade (A)	Método de postagem (P)	GA	GP	AP	GAP
Grupo 1	0,03	**36,58**	13,87	0,31	2,83	**44,74**	1,64
Grupo 2	0,00	**40,56**	7,26	20,67	19,39	6,34	5,77
Grupo 3	12,71	**43,23**	4,51	19,60	8,18	1,19	10,58

Tabela 4.3: A variação percentual do número de interações baseadas em mensagens explicada por cada tipo de atributo

	Gênero (G)	Nível de atividade (A)	Método de postagem (P)	GA	GP	AP	GAP
Grupo 1	0,46	**41,32**	21,69	0,00	0,61	**35,90**	0,02
Grupo 2	7,58	**31,98**	12,62	15,93	15,93	10,19	5,78
Grupo 3	12,58	**31,42**	17,92	12,94	12,37	2,13	10,65

Tabela 4.4: A variação percentual nos valores de *Klout Score* explicada por cada tipo de atributo

Observamos, também, que a importância de alguns dos atributos varia significativamente com o grupo de usuários-alvo dos socialbots. Por exemplo, o gênero do socialbot apresentou uma grande importância com usuários-alvo do Grupo 3, sendo responsável por 20,52% da variação do número de seguidores (tabela 4.2) e 12,71% das interações baseadas em mensagens (Tabela 4.3) quando os usuários-alvo são deste grupo.[6] No entanto, o gênero não parece ter muita influência sobre os outros grupos-alvo. Isso sugere que o gênero dos socialbots pode fazer a diferença se os usuários-alvo são suscetíveis a seguir e interagir com os usuários de um determinado sexo.

4.6 Discussão dos resultados

A seguir discutimos os resultados apresentados previamente. Na seção 4.4 analisamos o impacto de vários atributos dos socialbots – como o sexo mencionado no perfil – no seu desempenho de infiltração, enquanto certos atributos não afetam significativamente o desempenho de infiltração, outros atributos, como o nível de atividade e a escolha dos usuários-alvo apresentaram grande impacto sobre o desempenho de infiltração.

[6]Descobrimos que os usuários do Grupo 3 eram mais propensos a seguir e interagir com socialbots com perfis femininos.

Posteriormente na seção 4.5 analisamos a importância relativa dos diferentes atributos utilizando um experimento fatorial. Observamos que o atributo com maior impacto na infiltração é o nível de atividade chegando a ser responsável por 70% do total de seguidores de um grupo de socialbots. Além disso, notamos também, que a importância de alguns dos atributos varia significativamente com o grupo de usuários-alvo dos socialbots.

Capítulo 5

Conclusão e Trabalhos Futuros

Neste trabalho realizamos um estudo sobre bots no Twitter, inicialmente abordamos o problema de detecção de bots. Apresentamos uma ampla caracterização do comportamento de bots no Twitter usando três conjuntos de atributos: do usuário, de conteúdo e linguísticos. Nossa análise aponta que os bots tendem a postar mais tweets contendo URLs e hashtags que usuários, além de possuírem um padrão de escrita mais detectável que o de usuários. Além disso, usuários tendem a ser mais "sociais" e participativos em conversas do que os bots.

Com base em nossas medições e caracterização, criamos um método de detecção automática de bots usando um algoritmo de classificação supervisionado. Nosso método foi capaz de detectar 92% dos bots enquanto apenas menos de 1% dos usuários são classificados erroneamente. Posteriormente, estudamos o desempenho de cada atributo proposto e notamos que a idade da conta, a fração de URLs e o padrão de escrita possuem alto poder discriminativo. Finalmente, testamos o desempenho de nosso classificador ao utilizar apenas subconjuntos de atributos. Observamos que nossa abordagem consegue ter um bom desempenho ainda quando apenas um grupo de nossos atributos é utilizado.

Posteriormente, realizamos um estudo sobre quais características tornam socialbots mais bem sucedidos em tarefas de infiltração. Para isso, foram criados 120 socialbots no Twitter. Durante 30 dias monitoramos seu comportamento e todas suas interações com usuários da rede, incluindo 600 usuários-alvo. Durante esse período 2.637 usuários, sendo 103 usuários-alvo, interagiram 5.966 vezes com nossos bots.

Detectamos que características dos bots, como o seu nível de atividade, influenciam significativamente na sua popularidade no Twitter. Além disso, notamos que infiltrar grupos de amigos não foi mais fácil do que infiltrar um grupo de usuários não conectados. Esse resultado mostra que tarefas de infiltração no Twitter diferem das de

outras redes sociais como o Facebook. Finalmente, notamos que bots mais populares não apresentam necessariamente um melhor desempenho em tarefas de infiltração.

Acreditamos que esses resultados representam um importante passo no entendimento do impacto de socialbots, além do desenvolvimento de métodos de detecção de bots com estratégias complexas, que não podem ser detectados por algoritmos de detecção de atividade automática. Como trabalhos futuros pretendemos investigar que outros atributos e estratégias podem elevar a popularidade de bots no Twitter. Além disso, pretendemos implementar um sistema Web de alerta de contas suspeitas de serem bots.

Referências Bibliográficas

Aggarwal, A.; Almeida, J. & Kumaraguru, P. (2013a). Detection of spam tipping behaviour on foursquare. Em *Proceedings of the 22nd International Conference on World Wide Web Companion*, WWW '13 Companion, pp. 641--648, Republic and Canton of Geneva, Switzerland. International World Wide Web Conferences Steering Committee.

Aggarwal, A.; Rajadesingan, A. & Kumaraguru, P. (2013b). Phishari: Automatic realtime phishing detection on twitter. *CoRR*, abs/1301.6899.

Androutsopoulos, I.; Paliouras, G.; Karkaletsis, V.; Sakkis, G.; Spyropoulos, C. D. & Stamatopoulos, P. (2000). Learning to filter spam e-mail: A comparison of a naive bayesian and a memory-based approach. pp. 1--13.

Baeza-Yates, R. A. & Ribeiro-Neto, B. (1999). *Modern Information Retrieval*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA. ISBN 020139829X.

Becchetti, L.; Castillo, C.; Donato, D.; Leonardi, S. & Baeza-Yates, R. (2006). Link-based characterization and detection of web spam. Em *In AIRWeb*.

Benevenuto, F.; Magno, G.; Rodrigues, T. & Almeida, V. (2010a). Detecting spammers on Twitter. Em *Proceedings of the Seventh Annual Collaboration, Electronic messaging, Anti-Abuse and Spam Conference (CEAS)*.

Benevenuto, F.; Rodrigues, T.; Almeida, V.; Almeida, J. & Gonçalves, M. (2009). Detecting spammers and content promoters in online video social networks. Em *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '09, pp. 620--627, New York, NY, USA. ACM.

Benevenuto, F.; Rodrigues, T.; Almeida, V.; Almeida, J.; Gonçalves, M. & Ross, K. (2010b). Video pollution on the web. *First Monday*, 15(4).

Bharat, K. & Henzinger, M. R. (1998). Improved algorithms for topic distillation in a hyperlinked environment. Em *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '98, pp. 104--111, New York, NY, USA. ACM.

Blum, A.; Wardman, B.; Solorio, T. & Warner, G. (2010). Lexical feature based phishing url detection using online learning. Em *Proceedings of the 3rd ACM Workshop on Artificial Intelligence and Security*, AISec '10, pp. 54--60, New York, NY, USA. ACM.

Boshmaf, Y.; Muslukhov, I.; Beznosov, K. & Ripeanu, M. (2011). The socialbot network: when bots socialize for fame and money. Em *Proceedings of the 27th Annual Computer Security Applications Conference*, ACSAC '11, pp. 93--102, New York, NY, USA. ACM.

Boshmaf, Y.; Muslukhov, I.; Beznosov, K. & Ripeanu, M. (2012). Key challenges in defending against malicious socialbots. Em *Proceedings of the 5th USENIX Conference on Large-Scale Exploits and Emergent Threats*, LEET'12, pp. 12--12, Berkeley, CA, USA. USENIX Association.

Boykin, P. & Roychowdhury, V. (2005). Leveraging social networks to fight spam. *Computer*, 38(4):61–68. ISSN 0018-9162.

Bratko, A.; Cormack, G. V.; R, D.; Filipič, B.; Chan, P.; Lynam, T. R. & Lynam, T. R. (2006). Spam filtering using statistical data compression models. *Journal of Machine Learning Research*, 7:2673--2698.

Breiman, L. (2001). Random forests. *Mach. Learn.*, 45(1):5--32. ISSN 0885-6125.

Calzolari, M. C. (2012). Analysis of twitter followers of the us presidential election candidates: Barack obama and mitt romney. `http://digitalevaluations.com/DigitalEvaluations-Obama_Romney.pdf`.

Castillo, C.; Donato, D.; Gionis, A.; Murdock, V. & Silvestri, F. (2007). Know your neighbors: Web spam detection using the web topology. Em *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '07, pp. 423--430, New York, NY, USA. ACM.

Cha, M.; Haddadi, H.; Benevenuto, F. & Gummadi, K. P. (2010). Measuring User Influence in Twitter: The Million Follower Fallacy. Em *Proceedings of the 4th International AAAI Conference on Weblogs and Social Media (ICWSM)*, Washington DC, USA.

Chhabra, S.; Aggarwal, A.; Benevenuto, F. & Kumaraguru, P. (2011). Phi.sh/$ocial: The phishing landscape through short urls. Em *Proceedings of the 8th Annual Collaboration, Electronic messaging, Anti-Abuse and Spam Conference (CEAS)*.

Chirita, P.-A.; Diederich, J. & Nejdl, W. (2005). Mailrank: Using ranking for spam detection. Em *Proceedings of the 14th ACM International Conference on Information and Knowledge Management*, CIKM '05, pp. 373--380, New York, NY, USA. ACM.

Chu, Z.; Gianvecchio, S.; Wang, H. & Jajodia, S. (2012). Detecting automation of twitter accounts: Are you a human, bot, or cyborg? *IEEE Trans. Dependable Secur. Comput.*, 9(6):811--824. ISSN 1545-5971.

Coburn, Z. & Marra, G. (2008). Realboy: believable twitter bots. `http://ca.olin.edu/2008/realboy/index.html`.

Costa, H.; Benevenuto, F. & de Campos Merschmann, L. H. (2013). Detecting tip spam in location-based social networks. Em *Proceedings of the 28th Annual ACM Symposium on Applied Computing (SAC)*.

Damiani, E.; De Capitani di Vimercati, S.; Paraboschi, S. & Samarati, P. (2004). P2p-based collaborative spam detection and filtering. Em *Peer-to-Peer Computing, 2004. Proceedings. Proceedings. Fourth International Conference on*, pp. 176–183.

Danezis, G. & Mittal, P. (2009). Sybilinfer: Detecting sybil nodes using social networks. Em *NDSS*. The Internet Society.

Drucker, H.; Wu, S. & Vapnik, V. (1999). Support vector machines for spam categorization. *Neural Networks, IEEE Transactions on*, 10(5):1048–1054. ISSN 1045-9227.

Elishar, A.; Fire, M.; Kagan, D. & Elovici, Y. (2012). Organizational intrusion: Organization mining using socialbots. Em *Proceedings of the 2012 International Conference on Social Informatics*, SOCIALINFORMATICS '12, pp. 7--12, Washington, DC, USA. IEEE Computer Society.

Elyashar, A.; Fire, M.; Kagan, D. & Elovici, Y. (2013). Homing socialbots: Intrusion on a specific organization's employee using socialbots. Em *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, ASONAM '13, pp. 1358--1365, New York, NY, USA. ACM.

Fette, I.; Sadeh, N. & Tomasic, A. (2007). Learning to detect phishing emails. Em *Proceedings of the 16th International Conference on World Wide Web*, WWW '07, pp. 649--656, New York, NY, USA. ACM.

Fetterly, D.; Manasse, M. & Najork, M. (2004). Spam, damn spam, and statistics: Using statistical analysis to locate spam web pages. Em *Proceedings of the 7th International Workshop on the Web and Databases: Colocated with ACM SIGMOD/PODS 2004*, WebDB '04, pp. 1--6, New York, NY, USA. ACM.

Franceschi-Bicchierai, L. (2013). Social media spam increased 355half of 2013. `http://mashable.com/2013/09/30/social-media-spam-study/`.

Gao, H.; Hu, J.; Wilson, C.; Li, Z.; Chen, Y. & Zhao, B. Y. (2010). Detecting and characterizing social spam campaigns. Em *Proceedings of the 10th ACM SIGCOMM Conference on Internet Measurement*, IMC '10, pp. 35--47, New York, NY, USA. ACM.

Gara, T. (2013). One big doubt hanging over twitter's ipo: Fake accounts. `http://online.wsj.com/news/articles/SB10001424052702303492504579113754194762812`.

Garera, S.; Provos, N.; Chew, M. & Rubin, A. D. (2007). A framework for detection and measurement of phishing attacks. Em *Proceedings of the 2007 ACM Workshop on Recurring Malcode*, WORM '07, pp. 1--8, New York, NY, USA. ACM.

Garg, A.; Battiti, R. & Cascella, R. G. (2006). "may i borrow your filter?"exchanging filters to combat spam in a community. Em *Proceedings of the 20th International Conference on Advanced Information Networking and Applications - Volume 02*, AINA '06, pp. 489--493, Washington, DC, USA. IEEE Computer Society.

Geoffrey A. Fowler, Shayndi Raice, A. E. (2012). Spam finds new target. `http://online.wsj.com/news/articles/SB10001424052970203686204577112942734977800`.

Ghosh, S.; Viswanath, B.; Kooti, F.; Sharma, N. K.; Korlam, G.; Benevenuto, F.; Ganguly, N. & Gummadi, K. P. (2012). Understanding and combating link farming in the twitter social network. Em *Proceedings of the 21st International Conference on World Wide Web*, WWW '12, pp. 61--70, New York, NY, USA. ACM.

Gomide, J.; Veloso, A.; Jr., W. M.; Almeida, V.; Benevenuto, F.; Ferraz, F. & Teixeira, M. (2011). Dengue surveillance based on a computational model of spatio-temporal locality of twitter. Em *ACM Web Science Conference (WebSci)*.

Grandoni, D. (2012). Spam costs you a lot more than you'd think. `http://www.huffingtonpost.com/2012/08/08/cost-of-spam_n_1757726.html`.

Grier, C.; Thomas, K.; Paxson, V. & Zhang, M. (2010). @spam: The underground on 140 characters or less. Em *Proceedings of the 17th ACM Conference on Computer and Communications Security*, CCS '10, pp. 27--37, New York, NY, USA. ACM.

Gyöngyi, Z. & Garcia-Molina, H. (2005). Link spam alliances. Em *Proceedings of the 31st International Conference on Very Large Data Bases*, VLDB '05, pp. 517--528. VLDB Endowment.

Gyöngyi, Z.; Garcia-Molina, H. & Pedersen, J. (2004). Combating web spam with trustrank. Em *Proceedings of the Thirtieth International Conference on Very Large Data Bases - Volume 30*, VLDB '04, pp. 576--587. VLDB Endowment.

Harris, D. (2013). Can evil data scientists fool us all with the world's best spam? `http://gigaom.com/2013/02/28/can-evil-data-scientists-fool-us-all` % url-with-the-worlds-best-spam/.

Henzinger, M. R.; Motwani, R. & Silverstein, C. (2002). Challenges in web search engines. *SIGIR Forum*, 36(2):11--22. ISSN 0163-5840.

Hershkop, S. (2006). Behavior-based email analysis with application to spam detection. Relatório técnico.

Irani, D.; Webb, S. & Pu, C. (2010). Study of static classification of social spam profiles in myspace. Em Cohen, W. W. & Gosling, S., editores, *ICWSM*. The AAAI Press.

Jain, R. (1991). *The Art of Computer Systems Performance Analysis: Techniques for Experimental Design, Measurement, Simulation, and Modeling*. John Wiley and Sons, INC.

James, J. G. & Hendler, J. (2004). Reputation network analysis for email filtering. Em *In Proc. of the Conference on Email and Anti-Spam (CEAS), Mountain View*.

Jindal, N. & Liu, B. (2008). Opinion spam and analysis. Em *Proceedings of the 2008 International Conference on Web Search and Data Mining*, WSDM '08, pp. 219--230, New York, NY, USA. ACM.

Kouloumpis, E.; Wilson, T. & Moore, J. (2011). Twitter Sentiment Analysis: The Good, the Bad and the OMG! Em *Int'l Conference on Weblogs and Social Media (ICWSM)*.

Krishnan, V. (2006). Web spam detection with anti-trust rank. Em *In AIRWEB*, pp. 37--40.

Lazzari, L.; Mari, M. & Poggi, A. (2005). Cafe - collaborative agents for filtering e-mails. Em *Enabling Technologies: Infrastructure for Collaborative Enterprise, 2005. 14th IEEE International Workshops on*, pp. 356–361. ISSN 1524-4547.

Lee, K.; Eoff, B. D. & Caverlee, J. (2011). Seven months with the devils: A long-term study of content polluters on twitter. Em Adamic, L. A.; Baeza-Yates, R. A. & Counts, S., editores, *ICWSM*. The AAAI Press.

Lempel, R. & Moran, S. (2000). The stochastic approach for link-structure analysis (salsa) and the tkc effect. Em *Proceedings of the 9th International World Wide Web Conference on Computer Networks : The International Journal of Computer and Telecommunications Netowrking*, pp. 387--401, Amsterdam, The Netherlands, The Netherlands. North-Holland Publishing Co.

Li, J. & Subramanian, L. (2010). Optimal sybil-resilient node admission control. Relatório técnico.

Lim, E.-P.; Nguyen, V.-A.; Jindal, N.; Liu, B. & Lauw, H. W. (2010). Detecting product review spammers using rating behaviors. Em *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, CIKM '10, pp. 939--948, New York, NY, USA. ACM.

Manning, C. D. & Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA, USA. ISBN 0-262-13360-1.

Markines, B.; Cattuto, C. & Menczer, F. (2009). Social spam detection. Em *Proceedings of the 5th International Workshop on Adversarial Information Retrieval on the Web*, AIRWeb '09, pp. 41--48, New York, NY, USA. ACM.

Medlock, B. (2006). An adaptive approach to spam filtering on a new corpus.

Messias, J.; Schmidt, L.; Rabelo, R. & Benevenuto, F. (2013). You followed my bot! transforming robots into influential users in twitter. *First Monday*, 18(7).

Metsis, V. & Metsis, V. (2006). Spam filtering with naive bayes – which naive bayes? Em *Third Conference on Email and Anti-Spam (CEAS)*.

Mishne, G.; Carmel, D. & Lempel, R. (2005). Blocking blog spam with language model disagreement. Em *Proceedings of the First International Workshop on Adversarial Information Retrieval on the Web (AIRWeb)*, Chiba, Japan.

Mislove, A.; Post, A.; Druschel, P. & Gummadi, K. P. (2008). Ostra: Leveraging trust to thwart unwanted communication. Em *Proceedings of the 5th USENIX Symposium on Networked Systems Design and Implementation*, NSDI'08, pp. 15--30, Berkeley, CA, USA. USENIX Association.

Mo, G.; Zhao, W.; Cao, H. & Dong, J. (2006). Multi-agent interaction based collaborative p2p system for fighting spam. Em *IAT*, pp. 428–431. IEEE Computer Society.

Ntoulas, A.; Najork, M.; Manasse, M. & Fetterly, D. (2006). Detecting spam web pages through content analysis. Em *Proceedings of the 15th International Conference on World Wide Web*, WWW '06, pp. 83--92, New York, NY, USA. ACM.

O'Brien, C. & Vogel, C. (2003). Spam filters: Bayes vs. chi-squared; letters vs. words. Em *Proceedings of the 1st International Symposium on Information and Communication Technologies*, ISICT '03, pp. 291--296. Trinity College Dublin.

O'Callaghan, D.; Harrigan, M.; Carthy, J. & Cunningham, P. (2012). Network analysis of recurring youtube spam campaigns.

Orcutt, M. (2012). Twitter mischief plagues mexico's election. `http://www.technologyreview.com/news/428286/twitter-mischief-plagues-mexicos-election/`.

Page, L.; Brin, S.; Motwani, R. & Winograd, T. (1999). The pagerank citation ranking: Bringing order to the web.

Palla, S. & Dantu, R. (2007). Unwanted smtp paths and relays. Em *Communication Systems Software and Middleware, 2007. COMSWARE 2007. 2nd International Conference on*, pp. 1–8.

Pantel, P. & Lin, D. (1998). Spamcop: A spam classification & organization program. Em *In Learning for Text Categorization: Papers from the 1998 Workshop*, pp. 95--98.

Post, A.; Shah, V. & Mislove, A. (2011). Bazaar: Strengthening user reputations in online marketplaces. Em *Proceedings of the 8th USENIX Conference on Networked Systems Design and Implementation*, NSDI'11, pp. 14--14, Berkeley, CA, USA. USENIX Association.

PR0-Pagerank-Penalty (2002). Pr0 - google's pagerank 0 penalty. `http://pr.efactory.de/e-pr0.shtml`.

Protalinski, E. (2013). Twitter sees 218m monthly active users, 163.5m monthly mobile users, 100m daily users, and 500m tweets per day. `http://thenextweb.com/twitter/2013/10/03/twitter-says-it-sees-215-million-monthly-active-users-100-million-daily-users-and-500-million-tweets-per-day/`.

Rao, J. M. & Reiley, D. H. (2012). The economics of spam. *Journal of Economic Perspectives*, 26(3):87–110.

Ratkiewicz, J.; Conover, M.; Meiss, M.; Gonçalves, B.; Patil, S.; Flammini, A. & Menczer, F. (2011). Truthy: Mapping the spread of astroturf in microblog streams. Em *Proceedings of the 20th International Conference Companion on World Wide Web*, WWW '11, pp. 249--252, New York, NY, USA. ACM.

Sahami, M.; Dumais, S.; Heckerman, D. & Horvitz, E. (1998). A bayesian approach to filtering junk e-mail.

Sakaki, T.; Okazaki, M. & Matsuo, Y. (2010). Earthquake shakes twitter users: Real-time event detection by social sensors. Em *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pp. 851--860, New York, NY, USA. ACM.

Siponen, M. T. & Stucke, C. (2006). Effective anti-spam strategies in companies: An international study. Em *HICSS*. IEEE Computer Society.

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

Sureka, A. (2011). Mining user comment activity for detecting forum spammers in youtube. *CoRR*, abs/1103.5044. informal publication.

Tan, P.-N.; Steinbach, M. & Kumar, V. (2005). *Introduction to Data Mining, (First Edition)*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA. ISBN 0321321367.

Thomas, K.; Grier, C.; Ma, J.; Paxson, V. & Song, D. (2011). Design and evaluation of a real-time url spam filtering service. Em *Proceedings of the 2011 IEEE Symposium on Security and Privacy*, SP '11, pp. 447--462, Washington, DC, USA. IEEE Computer Society.

Thomas, K.; McCoy, D.; Grier, C.; Kolcz, A. & Paxson, V. (2013). Trafficking fraudulent accounts: The role of the underground market in twitter spam and abuse. Em *Proceedings of the 22nd Usenix Security Symposium*.

Tran, D. N.; Li, J.; Subramanian, L. & Chow, S. S. M. (2011). Optimal sybil-resilient node admission control. Em *INFOCOM*, pp. 3218–3226. IEEE.

Tran, N.; Min, B.; Li, J. & Subramanian, L. (2009). Sybil-resilient online content voting. Em *In Proceedings of the 6th Symposium on Networked System Design and Implementation (NSDI)*.

Tumasjan, A.; Sprenger, T. O.; Sandner, P. G. & Welpe, I. M. (2010). Predicting elections with twitter: What 140 characters reveal about political sentiment. Em *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*, pp. 178--185.

twitter-46pc-lt100followers (2013). 46% of twitter users have less than 100 followers - simplify360. `http://simplify360.com/blog/46-of-twitter-users-have-less-than-100-followers/`.

twitter-shut-spammers (2012). Shutting down spammers. `https://blog.twitter.com/2012/shutting-down-spammers`.

Viswanath, B.; Mondal, M.; Clement, A.; Druschel, P.; Gummadi, K.; Mislove, A. & Post, A. (2012a). Exploring the design space of social network-based sybil defenses. Em *Communication Systems and Networks (COMSNETS), 2012 Fourth International Conference on*, pp. 1–8.

Viswanath, B.; Mondal, M.; Gummadi, K. P.; Mislove, A. & Post, A. (2012b). Canal: Scaling social network-based sybil tolerance schemes. Em *Proceedings of the 7th ACM European Conference on Computer Systems*, EuroSys '12, pp. 309--322, New York, NY, USA. ACM.

Viswanath, B.; Post, A.; Gummadi, K. P. & Mislove, A. (2010). An analysis of social network-based sybil defenses. *SIGCOMM Comput. Commun. Rev.*, 41(4):--. ISSN 0146-4833.

Wagner, C.; Mitter, S.; Körner, C. & Strohmaier, M. (2012). When social bots attack: Modeling susceptibility of users in online social networks. Em *2nd workshop on Making Sense of Microposts at WWW '12*.

Wald, R.; Khoshgoftaar, T. M.; Napolitano, A. & Sumner, C. (2013). Which users reply to and interact with twitter social bots? Em *Tools with Artificial Intelligence (ICTAI), 2013 IEEE 25th International Conference on*, pp. 135–144. ISSN 1082-3409.

Whittaker, C.; Ryner, B. & Nazif, M. (2010). Large-scale automatic classification of phishing pages. Em *NDSS*. The Internet Society.

William R. Avison, J. D. M. & (Eds.), B. A. P. (2007). *Mental Health, Social Mirror*. Springer.

Wu, B. & Davison, B. D. (2005). Identifying link farm spam pages. Em *Special Interest Tracks and Posters of the 14th International Conference on World Wide Web*, WWW '05, pp. 820--829, New York, NY, USA. ACM.

Yeh, C.-Y.; Wu, C.-H. & Doong, S.-H. (2005). Effective spam classification based on meta-heuristics. Em *Systems, Man and Cybernetics, 2005 IEEE International Conference on*, volume 4, pp. 3872–3877 Vol. 4.

Yu, H.; Gibbons, P.; Kaminsky, M. & Xiao, F. (2008). Sybillimit: A near-optimal social network defense against sybil attacks. Em *Security and Privacy, 2008. SP 2008. IEEE Symposium on*, pp. 3–17. ISSN 1081-6011.

Yu, H.; Kaminsky, M.; Gibbons, P. B. & Flaxman, A. (2006). Sybilguard: Defending against sybil attacks via social networks. *SIGCOMM Comput. Commun. Rev.*, 36(4):267--278. ISSN 0146-4833.

Zhang, C. M. & Paxson, V. (2011). Detecting and analyzing automated activity on twitter. Em *Proceedings of the 12th International Conference on Passive and Active Measurement*, PAM'11, pp. 102--111, Berlin, Heidelberg. Springer-Verlag.

Zhang, Y.; Hong, J. I. & Cranor, L. F. (2007). Cantina: A content-based approach to detecting phishing web sites. Em *Proceedings of the 16th International Conference on World Wide Web*, WWW '07, pp. 639--648, New York, NY, USA. ACM.

Zhou, F.; Zhuang, L.; Zhao, B. Y.; Huang, L.; Joseph, A. D. & Kubiatowicz, J. (2003). Approximate object location and spam filtering on peer-to-peer systems. Em *Proceedings of the ACM/IFIP/USENIX 2003 International Conference on Middleware*, Middleware '03, pp. 1--20, New York, NY, USA. Springer-Verlag New York, Inc.