Socialbots: Implicações na segurança e na credibilidade de serviços baseados no Twitter

Carlos Freitas[1], **Fabrício Benevenuto**[1], **Adriano Veloso**[1]

[1] Departamento de Ciência da Computação
Universidade Federal de Minas Gerais (UFMG) – Belo Horizonte, MG – Brazil

{alessandro, fabricio,adrianov}@dcc.ufmg.br

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.*

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.*

1. Introdução

O Twitter é inegavelmente 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 and 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, dessa forma 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 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]

A quantidade exata de bots no Twitter é desconhecida. [Chu et al. 2012] estimam que 50% das contas sejam associadas a bots. 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%, a necessidade de estratégias para a detecção de bots no Twitter é crucial para garantir a credibilidade e segurança dos serviços que usam o Twitter como fonte de dados.

Neste artigo, abordamos o problema de detectar bots no Twitter. Nosso foco está na identificação de comportamentos de bots que extrapolam as estratégias de identificação de atividade automática. Como principais contribuições podemos mencionar: (i) a caracterização do comportamento de bots em uma grande base de dados, (ii) identificação de atributos linguísticos na postagem de bots, que até onde tenhamos conhecimento nunca foram utilizados para a detecção de bots, e finalmente, (iii) a criação de um método de detecção automática de bots que explora os atributos identificados.

Nossa abordagem foi capaz de detectar mais de 92% dos bots da nossa base de dados, classificando erroneamente menos de 1% dos usuários. Para isso, nós construímos uma coleção contendo 19.115 bots, identificados através de uma abordagem de identificação de padrões automáticos de postagem. A partir desses bots, nós investigamos diversos outros aspectos capazes de diferenciá-los de usuários comuns. De posse dessas características, nós investigamos a viabilidade de uma estratégia supervisionada de classificação para a identificação de bots.

O restante do artigo está organizado da seguinte forma: Na próxima seção apresentamos trabalhos relacionados. Na Seção 3 descrevemos a construção da base de dados de bots utilizada em nossos experimentos. Na Seção 4 apresentamos um estudo dos atributos usados por nosso método. Na Seção 5 apresentamos os resultados obtidos por nosso

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

método. Finalmente, na Seção 6 apresentamos conclusões e direções para possíveis trabalhos futuros.

2. Trabalhos Relacionados

Existem vários estudos com foco na criação de bots. O projeto Realboy visa a criação de bots que imitam usuários reais de forma verossímil [Coburn and Marra 2008]. O Web Ecology Project[2] 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[3] e Twitalyzer.[4] Finalmente, [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%. De maneira geral, esses esforços demostram a vulnerabilidade do Twitter à infiltração de bots.

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 preditores se um usuário irá interagir com um bot.

Apesar de bots não serem utilizados necessariamente para postar algum tipo de spam, existem diversos esforços nessa tarefa que são complementares ao nosso esforço. Em particular, [Grier et al. 2010, Pitsillidis et al. 2010, Stringhini et al. 2010, Benevenuto et al. 2010] desenvolveram técnicas para a detecção automáticas de spambots baseadas no seu comportamento anormal. [Lee et al. 2011] realizaram um estudo de longo prazo sobre poluidores de conteúdo no Twitter usando "honeypots", cujo modelo conseguiu detectar spammers com 98% de acurácia. Contudo, não é claro o desempenho destes métodos para a detecção de bots que não estejam envolvidos em atividades relacionadas a spam. 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.

[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. [Zhang and Paxson 2011] desenvolveram um método para detecção de contas com atividade automatizada usando apenas o "timestamp" das mensagens utilizando um teste χ^2. Apesar desses métodos apresentarem bons resultados, eles podem ser facilmente burlados por bots que: (i) postem com intervalos aleatórios, ou seguem uma distribuição

[2]http://www.webecologyproject.org/category/competition/

[3]http://klout.com/

[4]http://twitalyzer.com/

similar a comportamentos típicos de humanos, (ii) diminuam a fração de tweets com URLs, e (iii) usando ferramentas para automação web que imitem um navegador, (e.g., phantomjs [5] e o fake[6]). 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.

3. 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. A base de de dados foi criada a partir de 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, proposto recentemente[Zhang and Paxson 2011]. Além disso, nós selecionamos 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 1. Teste de atividade automática

	Reprovadas	Aprovadas	Insuficientes
Não suspensas	5.755	91.118	903.127
Suspensas	19.115	25.355	334.869

Como podemos perceber pelos resultados da Tabela 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 possuem nenhum método de automatização como usuários legítimos. De forma similar,

[5]http://phantomjs.org/

[6]http://fakeapp.com/

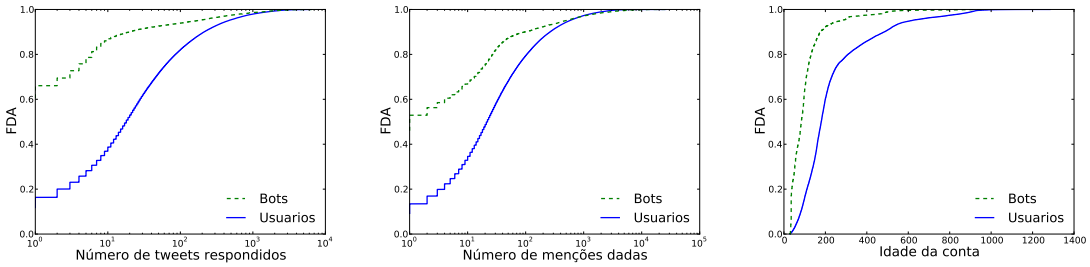
consideramos que as contas suspensas com atividade automática são bots. Dessa forma, nosso dataset contém **110.233** (91.118+19.115) contas e **42.773.272** de tweets.

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

4.1. 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 último tweet 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

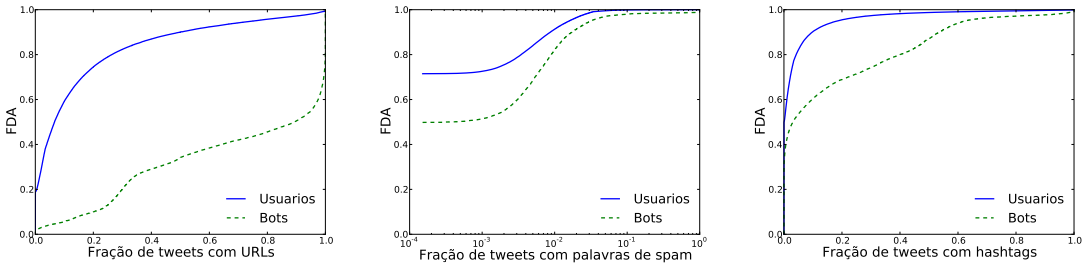
(c) Idade da Conta

Figura 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 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 1(a) e 1(b) notamos que usuários legítimos são mais sociais e ativos em conversas do que bots. Finalmente, a Figura 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.

4.2. 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, 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) Fração de tweets com URLs

(b) Fração de tweets com palavras de spam

(c) Fração de tweets com hashtags

Figura 2. Funções de distribuição acumulada de três 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 2 mostra as FDAs destes atributos. A Figura 2(a) mostra que bots postam mais tweets com URLs que usuários legítimos. Contudo, como a Figura 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 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.

4.3. 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, sílabas etc. Um n-grama de tamanho 1 é conhecido como unigrama, de tamanho 2 como brigrama 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 and 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 and 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

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

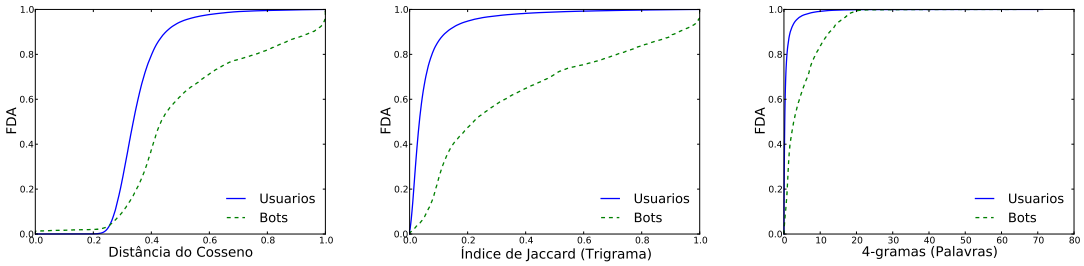
  Assumimos que a probabilidade de observar a palavra w_i é dada por apenas as últimas $n - 1$ palavras, propriedade Markoviana. Dessa forma a probabilidade condicional pode ser calculada a partir da contagem da frequência dos n-gramas nos tweets restantes do usuário.

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

  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. Funções de distribuição acumulada de três atributos linguísticos.

A seguir, realizamos uma análise de três atributos linguísticos: A distância do cosseno, o índice de Jaccard (trigrama) e o 4-gramas (palavras). A figura 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.

5. 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 5.1 as métricas usadas para avaliar os resultados da classificação. A seção 5.2 descreve o algoritmo de classificação, ou seja, o classificador, e ambiente experimental utilizado.

5.1. 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 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 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 $r_{bot} = \frac{a}{(a+b)}$.

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

Tabela 2. Exemplo de Matriz de Confusão

		Previsto	
		Bot	Usuário
Verdadeiro	Bot	a	b
	Usuário	c	d

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 contraria, 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.

5.2. 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.[7] 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.

5.3. Resultados da Classificação

A Tabela 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. Matriz de Confusão

		Previsto	
		Bot	Usuário
Verdadeiro	Bot	92.67%	7.33%
	Usuário	0.94%	99.16%

Uma pequena fração (mais de 7%) dos bots foram classificados erroneamente como usuários legítimos. Após uma inspeção manual, percebemos que esses bots tendem a postar poucas URLs e hashtags, além de postarem tweets contendo citações. Este

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

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 que não foram suspensos pelo Twitter e cujo padrão temporal de postagem não foi detectado pelo algoritmo de detecção de atividade automática. Dessa forma comprovamos que a abordagem proposta é mais robusta para a detecção de bots.

5.4. Importância dos Atributos

Para medir a importância dos atributos calculamos o ganho de informação, isto é redução esperada na entropia, de cada um dos mesmos. A Tabela 4 apresenta o ranking com os 20 atributos mais importantes segundo esta métrica.

Tabela 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 (Trigrama)
5	Índice de Jaccard (Bigrama)
6	Índice de Jaccard (Unigrama)
7	4-gramas (Palavra)
8	URLs por palavra (média)
9	Trigramas (Palavra)
10	Fração de respostas
11	Número de Amigos
12	Fração de mensagens que mencionam um usuário
13	Fração de respostas
14	URLs por palavra (mediana)
15	Número de menções por tweet (média)
16	Número de URLs (mediana)
17	Trigramas relativo (Palavras)
18	Número de dígitos por tweet (mediana)
19	Número de tweets dos amigos do usuário
20	Bigramas (Palavras)

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.

A Tabela 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,

Tabela 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

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.

5.5. 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. Portanto, 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 6 apresenta o desempenho do classificador usando diferentes conjuntos de atributos.

Tabela 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 (Modelo proposto)	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, notamos que nosso classificador possui um alto poder discriminativo independentemente do conjunto de atributos utilizado.

6. Conclusão

Neste trabalho abordamos o problema de detecção de bots no Twitter. 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 em nossos experimentos. Finalmente, testamos o desempenho de nosso classificador ao utilizar apenas subconjuntos de atributos, nós observamos que nossa abordagem consegue ter um bom desempenho ainda quando apenas um grupo de nossos atributos é utilizado.

Nós acreditamos que esses resultados representam um importante passo na detecção de bots com estratégias complexas e que não podem ser detectados por algoritmos de detecção de atividade automática. No futuro pretendemos implementar um sistema Web de alerta de contas suspeitas de serem bots.

7. Agradecimentos

Este trabalho teve apoio do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), da Fundação de Amparo à Pesquisa do estado de Minas Gerais (FAPEMIG), da da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) e também do Instituto Nacional de Ciência e Tecnologia para a Web (InWeb). Agradecemos também à Saptarshi Ghosh por fornecer a base de dados usada na nossa pesquisa sem a qual o presente trabalho não teria sido possível.

Referências

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

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

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

Breiman, L. (2001). Random forests. *Mach. Learn.*, 45(1):5 – 32.

Calzolari, M. C. (2012). Analysis of twitter followers of the us presidential election candidates: Barack obama and mitt romney.

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

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

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

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

Ghosh, S., Viswanath, B., Kooti, F., Sharma, N. K., Gautam, K., Benevenuto, F., Ganguly, N., and Gummadi, K. (2012). Understanding and Combating Link Farming in the Twitter Social Network. In *Proceedings of the 21st International World Wide Web Conference (WWW'12)*, Lyon, France.

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

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

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-with-the-worlds-best-spam/.

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

Manning, C. D. and Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA, USA.

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

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

Pitsillidis, A., Levchenko, K., Kreibich, C., Kanich, C., Voelker, G. M., Paxson, V., Weaver, N., and Savage, S. (2010). Botnet judo: Fighting spam with itself. In *NDSS*. The Internet Society.

Protalinski, E. (2013). Twitter sees 218m monthly active users, 163.5m monthly mobile users, 100m daily users, and 500m tweets per day.

Ratkiewicz, J., Conover, M., Meiss, M., Gonçalves, B., Patil, S., Flammini, A., and Menczer, F. (2011). Truthy: Mapping the spread of astroturf in microblog streams.

In *Proceedings of the 20th International Conference Companion on World Wide Web*, WWW '11, pages 249–252, New York, NY, USA. ACM.

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

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

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

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

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

Wagner, C., Mitter, S., Körner, C., and Strohmaier, M. (2012). When social bots attack: Modeling susceptibility of users in online social networks. In *2nd workshop on Making Sense of Microposts at WWW2012*.

Wald, R., Khoshgoftaar, T., Napolitano, A., and Sumner, C. (2013). Predicting susceptibility to social bots on twitter. In *Information Reuse and Integration (IRI), 2013 IEEE 14th International Conference on*, pages 6–13.

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