Empregando Entropia na Detecção de Comportamento Automatizado nos *Trend Topics* do Brasil

Adeilson Souza[1], **Eduardo Feitosa**[1]

[1]Instituto de Computação – Universidade Federal do Amazonas (UFAM)
Av. Gen. Rodrigo Octávio, 6200 – 69.000-077 – Manaus – AM – Brasil

{adeilson, efeitosa}@icomp.ufam.edu.br

Abstract. *Twitter enables the creation and propagation of automatized tweets for commercial purposes, but this possibility has been used by social bots - automated accounts - for malicious purposes, especially in Trend Topics (TT). Studying the characteristics and behaviors of social bots in Twitter, this article proposes six new features, based on entropy, with high discrimination power and able to distinguish automated behavior in Brazilian TT. Using a real database containing 2.853.822 accounts and 11.294.861 tweets, we identified and extracted features of tweets (texts and user behaviors) on TT. Next, applying machine learning algorithms for supervised classification, we are capable to detect 92% of automated accounts in the used database and thus get an insight into the behavior of these users.*

Resumo. *O Twitter permite a criação e propagação de postagens automatizadas para fins comerciais, mas tal característica vem sendo utilizada indevidamente por bots sociais - contas automatizadas - para fins maliciosos, especialmente nos Tópicos de Tendência (TT). Estudando as características e comportamento dos bots no Twitter, este artigo propõe seis novas características, baseadas no conceito de Entropia, com alto poder discriminativo e capazes de distiguir o comportamento automatizado nos TT no Brasil. A partir de uma base de dados real contendo 2.853.822 contas e 11.294.861 tweets, foram identificadas e extraídas características do texto das mensagens e do comportamento dos usuários nos TT. Aplicando algoritmos de aprendizagem de máquina supervisionada para classificação, foi possível detectar 92% das contas automatizadas na base de dados utilizada e, assim, obter uma visão do comportamento desses usuários.*

1. Introdução e Motivação

A presença de perfis "não humanos" em redes socais, criados com a intenção de propagar automaticamente conteúdo não desejado vem aumentando nos últimos anos. O Twitter, por exemplo, afirmou que apenas 5% dos seus 215 milhões de usuários ativos estavam ligados a atividades automatizadas [Gara 2013], mas autores como Chu et al. [Chu et al. 2012] estimam que 50% das contas no Twitter estão ou são associadas a *bots* sociais[1].

[1]*Bots* sociais são programas automatizados que utilizam contas em redes sociais, para se passar por usuários legítimos, com o intuito de enganar ou influenciar outros usuários

De acordo com as regras do Twitter [Twitter 2014], *bots* sociais são permitidos para compartilhar conteúdo útil (notícias e eventos) e propagandas autorizadas. Contudo, vários estudos mostram que os *bots* sociais no Twitter vem sendo utilizados como canal de controle para *botnets* [Nazario 2009, Kartaltepe et al. 2010], para *phishing* [Malware Bytes 2014, Symantec 2013], para propagação de spam contendo propagandas, pornografia e código malicioso [Neal Ungerleider 2015] e também para favorecer candidatos políticos [Orcutt 2012], mesmo que existam regras e práticas para criação e uso de contas automatizadas (por exemplo, postar *tweets* automatizados repetidamente para os Tópicos de Tendência é proibido).

O Twitter fornece métodos para que usuários denunciem comportamento suspeito ou conteúdos ofensivos. Tais denúncias são investigadas e caso algo seja comprovado, as contas são suspensas e/ou o acesso pelo endereço IP referente à conta é temporariamente bloqueado. Entretanto, esse mecanismo não tem se mostrado muito eficiente, uma vez que basta os usuários maliciosos criarem uma conta diferente para continuarem enviando mensagens. No caso de denúncias envolvendo *tweets* nos TT, o resultado é ainda pior, uma vez que o processo de suspensão é lento e os TT são efêmeros e, em muitos casos, duram apenas algumas horas [Martinez-Romo and Araujo 2013]. Na literatura existem diferentes trabalhos de detecção de *bots* sociais no Twitter [Benevenuto et al. 2010, Lee et al. 2010, Stringhini et al. 2010, Wang 2010, Wang 2012] que visam minimizar e combater o uso indevido de contas automatizadas através do estudo de seu comportamento, identificando características e padrões de atividades que possam ser utilizados para desenvolver contramedidas.

A fim de contribuir com a detecção de *bots* sociais no Twitter, este artigo propõe e avalia novas características baseadas no conceito de entropia, capazes de detectar comportamento automatizado nos Tópicos de Tendência do Twitter no Brasil, através da medição dos padrões de escrita dos *tweets* postados pelos usuários. Para alcançar esse objetivo, uma base com *tweets* e dados das postagens e contas de usuários que comentam sobre os TT no Brasil foi montada. Experimentos realizados mostram que o uso de entropia para medir a frequência de postagens e escrita dos *tweets* é eficaz, com taxa de acerto de cerca de 92% dos *bots* na base de dados utilizada.

O restante do artigo é organizado como segue: na Seção 2 são apresentados trabalhos relacionados. A Seção 3.1 descreve os atributos passíveis de serem empregados na detecção de contas automatizadas, bem como os seis novos atributos propostos baseados em entropia. A Seção 4 apresenta a base de dados coletada e o processo de refinamento aplicado para se obter contas de humanos e automatizadas para a realização de experimentos. Na Seção 5 são mostrados os resultados dos classificadores testados com o conjunto de atributos utilizados. Por fim, na Seção 6 são expostas as conclusões obtidas e as possibilidades de trabalhos futuros.

2. Trabalhos Relacionados

Os trabalhos que buscam identificar atividade automatizada preocupam-se em determinar se os usuários são *bots* ou humanos, a fim de dar maior credibilidade aos dados que são utilizados para a construção de novas aplicações e serviços baseados no Twitter.

Para determinar se uma conta no Twitter possui comportamento automatizado, Zhang e Paxson [Zhang and Paxson 2011] analisaram as atualizações de contas de

usuários, utilizando apenas a informação do horário de publicação (disponível publicamente) e a fonte de onde o *tweet* foi publicado. Foram avaliadas 106.573 contas distintas, coletadas durante 3 semanas em abril de 2010. Os autores aplicaram um teste chi quadrado de Pearson[2], e descobriram que contas automatizadas exibem padrões de intervalos de atualizações em intervalos de tempo uniformes, que dificilmente podem ser observados para usuários humanos. Também perceberam que uma parcela dessas contas publicam seus *tweets* diretamente da Web (via página) e que palavras relacionadas a spam estão associadas com contas que possuem um alto grau de automação.

Freitas et al. [Freitas et al. 2014] abordaram o problema de detectar *bots* no Twitter focando na identificação de comportamento que foge às estratégias de identificação de atividade automática. Para tanto, utilizaram como base o teste de atividade automática aplicado por Zhang e Paxson [Zhang and Paxson 2011] em um conjunto de contas suspensas do Twitter, identificadas por [Ghosh et al. 2012], e em um conjunto de contas não-suspensas, totalizando 110.233 usuários. Identificando atributos linguísticos nas postagens dos usuários e padrões de comportamento capazes de diferenciar usuários *bots* e humanos, os autores foram capazes de detectar cerca de 92% dos *bots* na base de dados.

John et al. [Dickerson et al. 2014] propõem uma nova abordagem utilizando análise de sentimento nos *tweets*. Para tanto, um conjunto de atributos identificados por meio da ferramenta SentiMetrix's, juntamente com características já utilizadas em outras abordagens (conteúdo do *tweet* e comportamento do usuário, por exemplo), foram utilizadas para identificar *bots* no Twitter. Os autores demonstraram que o número de fatores relacionados a sentimento são a chave para a identificação de *bots*.

Neste artigo, como a intenção restringe-se a detecção de atividade automatizada nos tópicos de tendência do Twitter no Brasil, são utilizados atributos das contas e do conteúdo dos *tweets*, bem como também seis novos atributos, baseados no conceito de entropia, para medir padrões de escrita dos *tweets* e seu poder discriminativo para a tarefa de identificar atividade automatizada.

3. Atributos Extraíveis do Twitter

A forma como os usuários escrevem seus *tweets* e interagem com outros usuários pode ajudar a diferenciar comportamentos automatizados e não automatizados. Esta seção descreve o conjunto de atributos passíveis de serem empregados na detecção de contas automatizadas. Primeiro, são descritos os atributos de conteúdo e de comportamento. Em seguida, os seis novos atributos baseados em entropia são detalhados e uma discussão sobre sua capacidade de detecção de contas automatizadas em conjunto aos outros atributos é feita.

3.1. Atributos de Conteúdo e de Comportamento

Atributos de conteúdo são aqueles baseados em propriedades do texto dos tweets postados pelos usuários. Eles identificam aspectos relacionados com a forma como os usuários escrevem seus *tweets*, tipicamente através da análise da ocorrência (quantidade) de diversos atributos no texto. É importante ressaltar que outros atributos podem ser avaliados em termos da sua existência ou não. O conjunto de

[2]É um teste não paramétrico de hipóteses que se destina a encontrar um valor da dispersão para duas variáveis nominais, avaliando a associação existente entre variáveis qualitativas.

atributos de conteúdo dos *tweets* mais comuns são: *source_tweet*, *retweet_count*, *diversidade_lexica*, *media_tweets_dia*, *media_tweets_hora*, *num_palavras*, *media_palavras_tweet*, *media_urls_tweet*, *media_urls_topico*, *media_tweets_topico*, *media_palavras_topico*, *media_hashtags_topico*, *media_hashtags_tweet*, *media_mencao_tweet*, *media_mencao_topico*, *favorite_count_tweet*, *favorited_tweet* e *retweeted*.

Já os atributos de comportamento são aqueles baseados em propriedades do comportamento do usuário, em termos de frequência de postagens, interações sociais e influência, que procuram identificar características que tenham relação com a forma como os usuários interagem e sua popularidade no Twitter. Os atributos mais comuns de comportamento do usuário são: *id_user*, *verified*, *favourites_count_user*, *listed_count*, *protected*, *default_profile*, *num_followers*, *num_followed*, *rate_followers_followed*, *num_tweets*, *count_age* e *dif_tweets*.

Tanto os atributos de conteúdo quanto os de comportamento apresentados nesta seção podem ser encontrados nos trabalhos [Benevenuto et al. 2010, Lee et al. 2010, Stringhini et al. 2010, Wang 2010, Wang 2012] e todos já demonstraram serem eficientes para medir e diferenciar a intenção dos usuários, bem como caracterizar mensagem suspeitas ou envolvidas em atividades automatizadas (com objetivos maliciosos ou não).

3.2. Atributos baseados em Entropia

Neste artigo, a entropia de um conjunto de *tweets* postados por um usuário é calculada para medir o quanto de informação os *tweets* representam, levando em consideração todos os *tweets* que ele postou na base utilizada. Para tanto, seis novos atributos, extraídos das mensagens, foram propostos e são empregados para caracterizar comportamento automatizado (Tabela 1).

Tabela 1. Atributos Propostos Baseados em Entropia

Atributos	Descrição
entropia_total	Entropia total que representa o vocabulário de um usuário
media_entropia_tweets	Entropia média por *tweet* publicado
media_entropia_topico	Entropia média por tópico
entropia_usuario_topicos_diferentes	Entropia dos *tweets* de um usuário em tópicos diferentes, levando em consideração apenas as *hashtags* utilizadas
entropia_usuarios_mesmo_topico	Entropia dos *tweets* de usuários diferentes nos 10 tópicos com maior número de *tweets*
entropia_usuarios_topicos_diferentes	Entropia apenas para um *tweet* por tópico

O primeiro atributo, $entropia\_total$, representa a medida geral de entropia para o vocabulário de cada usuário, onde o vocabulário é entendido como todos os símbolos utilizados por ele nos *tweets* coletados e os símbolos são todas as palavras e caracteres obtidos de um dado conjunto de *tweets* postados pelo usuário. É importante ressaltar que, no que tange à implementação dessa característica, para cada *tweet* ou conjunto de *tweets*, uma lista com todos símbolos é gerada. Já os atributos $media\_entropia\_tweets$ e $media\_entropia\_topico$ representam, respectivamente, o valor do atributo $entropia\_total$ dividido pelo total de *tweets* que o usuário possui no conjunto de dados e o valor do atributo $entropia\_total$ dividido pelo número de tópicos que o usuário participou no conjunto de dados.

O atributo $entropia\_usuario\_topicos\_diferentes$ calcula, para cada usuário, a entropia do conjunto de *tweets* por tópico, levando em consideração apenas as *hashtags* utilizadas. Assim, é formado o conjunto de palavras chaves para cada usuário. Devido ao tamanho limitado dos *tweets*, muitos *bots* constroem mensagens utilizando apenas *hashtags* em seus *tweets* [Freitas et al. 2014]. Essa prática mostra uma estratégia para alcançar diversos tópicos de uma só vez.

O atributo $entropia\_usuarios\_mesmo\_topico$ calcula a entropia dos *tweets* de cada usuário para os 10 tópicos com maior número de *tweets*. Essa medida representa o quanto cada usuário pode variar o vocabulário participando do mesmo tópico. A intenção é medir o conteúdo dos *tweets* quando os usuários comentam sobre os mesmos assuntos nos tópicos com o maior número de *tweets*, já que são os mais representativos em número de mensagens.

O atributo $entropia\_usuarios\_topicos\_diferentes$ calcula a entropia apenas para um *tweet* por tópico. Essa medida mostra se os usuários escrevem *tweets* únicos ou com textos muito parecidos, independentemente do tópico. Ao escolher aleatoriamente um *tweet* por tópico, a intenção é medir se o usuário na base de dados possui um vocabulário realmente diversificado. A ideia é que para comentar sobre um mesmo tópico, normalmente um usuário poste mensagens não tão diversificadas, já que fazem referência a um mesmo assunto, enquanto que um *tweet* para cada tópico deve ser mais diversificado em termos de vocabulário, já que faz referência a assuntos diversos.

Todas as seis características propostas visam quantificar o quanto cada usuário diversificou as palavras e caracteres utilizadas em seus *tweets*. Assim, se o alfabeto de um usuário é bastante diversificado, então a entropia será elevada. Uma vez que usuários humanos postam mensagens com conteúdo variado, espera-se que possuam uma entropia maior em relação aos *bots* sociais que postam mensagens automáticas com conteúdo menos diversificado. Desta forma, em conjunto com os atributos de conteúdo e de comportamento, os novos atributos propostos devem ser capazes de mostrar características relevantes para diferenciar usuários humanos de usuários automatizados.

4. Base de Dados

A **base de dados inicial** empregada no desenvolvimento deste artigo foi formada por um total de 2.853.822 usuários e 11.294.861 *tweets* únicos, coletados no período entre dezembro de 2013 e junho de 2014. Durante esse período de coleta, foram encontrados 2.712 tópicos de tendência distintos (únicos) com cerca de 2.000.000 de URLs.

Dado o grande tamanho da base de dados, foi necessário utilizar algum mecanismo para reduzir a quantidade de usuários. Ao analisar a base de dados coletados, percebeu-se que embora muitos usuários ativos possuam grande número de mensagens publicadas no Twitter esses mesmos usuários publicam poucos *tweets* nos tópicos de tendência. Assim, após análises na literatura, optou-se por empregar o mecanismo implementado em [Zhang and Paxson 2011] para reduzir a quantidade de usuários. A ideia é que contas com menos de 30 *tweets* publicados durante o período de coleta foram consideradas insuficientes e foram descartadas por falharem no teste de atividade automática.

Por exemplo, o usuário X possui um total de 500 *tweets* (obtidos através do atributo $statuses\_count$) publicados no Twitter até o último dia de coleta, mas desse total

apenas 28 *tweets* foram publicados nos tópicos de tendência. Este usuário é considerado inválido por conter dados insuficientes para investigação de comportamento automatizado nos tópicos de tendência.

Após verificar o número de *tweets* de cada usuário no conjunto de dados, uma nova base de dados foi montada, contendo 50.012 usuários com mais de 30 *tweets* nos tópicos de tendência do Brasil. Em seguida, aplicando o teste de atividade automática aos horários e frequência de postagens por minuto, hora e dia, proposto por [Zhang and Paxson 2011], foram rotuladas 37.919 contas (aprovadas no teste de atividade automática) como usuários humanos e 12.093 contas foram reprovadas (pertencentes a contas de *bots*) rotuladas como bots. Para confirmar e assegurar a validade do teste, os perfis das contas classificadas como automatizadas e seus *tweets* foram consultados utilizando scripts que acessam os perfis dos usuários.

No final, a **base de dados final**, para treino e teste, é formada por 50.012 usuários (37.919 humanos e 12.093 automatizados), com 4.352.107 *tweets* e 829.082 URLs, sendo 322.918 URLs únicas. Assim, foi montada uma base de treinamento com 25.026 amostras rotuladas como Humano e 7.981 amostras rotuladas como Automatizado e uma outra base para teste contendo 12.893 amostras rotuladas como Humano e 4.112 amostras rotuladas como Automatizado. Essa distribuição de valores se refere a 64% das contas para treinamento e 36% para teste, como descrito na literatura de aprendizagem de máquina [Alpaydin 2010].

5. Avaliação e Resultados

Nesta Seção, o desempenho dos atributos discutidos na Seção 3.1, especialmente os seis novos propostos, são avaliados através de algoritmos de aprendizado de máquina supervisionada a fim de validar sua efetividade na tarefa de detectar *bots* sociais no Twitter. Na Seção 5.1, o processo de treinamento dos classificadores é descrito, incluindo as métricas de avaliação e os ajustes nos classificadores avaliados. Já na Seção 5.2, o teste de validação das novas características é apresentado e seus resultados discutidos.

5.1. Treinamento

O processo de treinamento é necessário para ajustes de parâmetros individuais ou em conjunto para cada classificador, a fim de obter o melhor desempenho de cada classificador sobre o conjunto de dados utilizado nesta fase. O resultado esperado é um modelo capaz de identificar comportamento automatizado utilizando os atributos propostos e, posteriormente, testar em uma nova base de dados para avaliar o desempenho dos classificadores.

No processo de treinamento de classificadores foram utilizados oito (08) classificadores. Contudo, serão apresentados somente os resultados dos quatro (04) melhores: Decorate, SVM, RandomForest e J48. Todos os classificadores foram testados usando a ferramenta Weka e utilizando o processo de validação cruzada de 10 partições para o treinamento.

Antes de demonstrar o melhor desempenho de cada classificador no processo de treinamento, é preciso informar as métricas de avaliação utilizadas. São elas: Precisão, Revocação, Medida F e Área sob a curva ROC (AUC). A Precisão é a porcentagem de amostras positivas classificadas corretamente sobre o total de amostras classificadas como positivas. A Revocação é a porcentagem de amostras positivas classificadas corretamente

sobre o total de amostras positivas. A Medida F é um indicador que combina a Precisão e a Revocação. Seu resultado está no intervalo [0,1], sendo o melhor resultado 1 e o pior resultado 0. AUC é um indicador usualmente adotado como medida de qualidade do classificador em aprendizagem de máquina [Sokolova et al. 2006].

5.1.1. Ajustes de Parâmetros e Escolha do Melhor Classificador

Para obter o melhor resultado sobre o conjunto de dados para cada classificador, foram realizados treinamentos (20.092 instâncias) onde os valores dos principais parâmetros de cada classificador foram ajustados até a obtenção do valor mais adequado. A Tabela 2 demonstra a comparação entre os resultados dos quatro classificadores, a fim de determinar o que melhor se ajusta ao conjunto de treinamento. Explicações sobre os parâmetros de ajuste, bem como sobre classificadores podem ser encontrados em [Alpaydin 2010].

Tabela 2. Comparação entre os Classificadores

Classificador	SVM	J48	Decorate	RandomForest
Parâmetros Ajustados	C=10, Kernel Polinomial Grau 1.0	FC=0,50	R=3.0 I=90	I=220 K=40
Precisão	76,90%	89,80%	89,70%	97,50%
Revocação	59,60,%	90,30%	92,80%	94,70%
Medida F	0.672	0.684	0.705	0.884
Área ROC	0.517	0.782	0.906	0.937

Como observado na Tabela 2, o classificador RandomForest foi o que obteve o melhor desempenho em relação aos demais classificadores.

A matriz de confusão (Tabela 3) ilustra o desempenho do RandomForest. Para a classe Automatizado, o modelo gerado acertou em torno de 92,40% das amostras no teste realizado para o melhor conjunto de parâmetros escolhidos no processo de treinamento. De igual modo, para a classe Humano, cerca de 94,73% das amostras foram classificadas corretamente, elevando assim o poder discriminativo do classificador testado.

Tabela 3. Matriz de confusão do RandomForest

RandomForest		Previsto	
		Humano	Automatizado
Correto	Humano	94,73%	5,27%
	Automatizado	7,60%	92,40%

Além disso, o RandomForest apresentou taxas de falso positivo de 5,30% para classe Humano e 7,60% para classe Automatizado. Assim, após o treinamento realizado, o RandomForest demonstrou ser o mais adequado para o problema de detectar e distinguir usuários humanos e automatizados na base de dados utilizada neste artigo, juntamente com o conjunto de atributos sugeridos.

5.2. Testes

A fim de avaliar o poder discriminativo para detectar comportamento automatizado, o classificador RandomForest que obteve os melhores resultados, foi testado em uma nova base de dados rotulada com 12.893 usuários humanos e 4.112 usuários automatizados.

5.2.1. Relevância dos atributos

A fim de medir a relevância dos atributos utilizados foi calculado o ganho de informação, isto é, o quanto cada atributo é representativo para o problema de distinguir usuários humanos e usuários automatizados. A Tabela 4 apresenta o ranking do ganho de informação dos 36 atributos empregados neste trabalho. Vale ressaltar que esses atributos foram apresentados na Seção 3.1.

Tabela 4. Ranking Ganho de Informação

Num	Atributo	Num	Atributo
1	media_tweets_hora	2	media_tweets_dia
3	media_tweets_topico	4	media_entropia_tweets
5	media_hash_topico	6	entropia_usuario_topicos_diferentes
7	diversidade_lexica	8	media_palavras_tweet
9	media_palavras_topico	10	media_mencao_topico
11	num_palavras	12	media_hash_tweet
13	entropia_usuarios_mesmo_topico	14	media_entropia_topico
15	media_urls_tweet	16	entropia_usuarios_topicos_diferentes
17	media_urls_topico	18	media_menção_tweet
19	dif_tweets	20	statuses_count
21	idade_conta	22	entropia_total
23	listed_count	24	razao_foll_fri
25	followers_count	26	friends_count
27	favourites_count_user	28	retweet_count
29	source_tweet	30	num_fontes
31	favorite_count_tweet	32	verified
33	default_profile	34	protected
35	favorited_tweet	36	retweeted

Os três (3) atributos de maior relevância estão relacionados a média de *tweets* postados por hora ($media\_tweets\_hora$), por dia ($media\_tweets\_dia$) e por tópicos ($media\_tweets\_topico$). Os dois primeiros ajudam a provar que usuários humanos compartilham mensagens de maneira não uniforme tanto por hora quanto no decorrer de um único dia, uma vez que as pessoas comentam sobre um tópico específico. Por outro lado, as contas automatizadas comentam em diversos tópicos de forma aleatória e mantém uma quantidade uniforme de mensagens postadas por faixa horária. No caso de postagens por tópico ($media\_tweets\_topico$), os *bots* normalmente participam ativamente em inúmeros tópicos enquanto os usuários humanos comentam de maneira esporádica, as vezes concentrando-se em poucos tópicos.

Os atributos relacionados à forma como os usuários escrevem seus *tweets* também demonstram ter uma grande relevância. Atributos como a média de *hashtags* por tópico ($media\_hash\_topico$) e por *tweet* ($media\_hash\_tweet$) demonstram que os *bots* abusam do uso de palavras chaves para marcar seus *tweets* e assim têm maior chances de serem encontrados, por exemplo, em uma busca por assuntos. A ideia é fazer com que um mesmo *tweet* esteja ligado a vários tópicos de uma só vez, alcançando assim um maior número de usuários, ganhando maior visibilidade e, por consequência, aumentando a possibilidade de alcançar mais seguidores. Do mesmo modo, usuários *bots* compartilham muitas URLs em suas mensagens e também fazem menção a outros usuários da rede. O uso extensivo de URLs pode ser uma forma de tentar atrair os usuários mencionados e os seguidores desses usuários a clicar nos links, que em muitos casos direcionam para conteúdo externo como vídeos, imagem ou outros sites.

Os atributos baseados no cálculo da entropia das mensagens demonstram ser eficazes e importantes para distinguir os usuários. Observando que os atributos relativos à média de entropia por *tweet* e por tópico estão entre os 20 principais atributos, é possível assumir que existe uma alto grau de similaridade entre mensagens de usuários *bots* para tópicos diferentes. Isso fica muito evidente em um busca na base de dados, onde o padrão de escrita destas mensagens normalmente abusa do uso de múltiplas hashtags muitas vezes só alternando a ordem destas na mensagem fazendo com que a mesma mensagem esteja presente em diversos tópicos.

Consequentemente a medida da entropia destas mensagens é praticamente a mesma já que o texto é praticamente o mesmo. No Twitter não é permitido publicar a mesma mensagem mais de uma vez em um curto intervalo de tempo, por isso os *bots* utilizam diversas *hashtags* ou invertem a ordem em que elas aparecem na mensagem para dar a ilusão de que são mensagens diferentes. Também utilizam a lista de tópicos para inserir alguns na mensagem e alternar entre esses tópicos, criando diversas mensagens similares. Uma análise na base de dados permite observar que usuários humanos compartilham mensagens com conteúdo variado.

5.2.2. Análise da Relevância dos Atributos Baseados em Entropia

Como forma de provar a importância dos atributos baseados em entropia, os quatro atributos mais bem ranqueados na subseção 5.2.1 foram avaliados usando 2.000 amostras, sendo 1.000 amostras de cada classe (Humano e Automatizado). Vale explicar que ao analisar diversas amostras da base final, contendo 2.000, 4.000 e 8.000 contas, percebeu-se a existência de uma similaridade nos resultados obtidos. Assim, optou-se por um conjunto de 1.000 amostras para cada classe, uma vez que processar a base final por completo (todas as amostras), de uma única vez, demanda um maior poder computacional e tempo. Os valores da entropia calculada para cada atributo variam de 0 a 6, e para fins de demonstração considera-se os valores no intervalo de [0,1) como entropia muito baixa, de [1,2) como entropia baixa, de [2,3) como entropia média baixa, de [3,4) como entropia média alta, de [4,5) como entropia alta e de [5,6] como entropia muito alta.

Entropia por *Tweet* e por Tópico

A Figura 1 demonstra que a entropia dos usuários humanos por *tweet* é medida como média e alta, enquanto usuários automatizados ficam na faixa entre muito baixa até média. Embora ambos os valores para as duas classes atinjam cerca de 80% dos *tweets* avaliados, os dados reafirmam que humanos possuem um vocabulário mais diversificado do que os automatizados. Além disso, os *bots* sociais tendem a ser mais repetitivos e menos espontâneos na escrita dos *tweets*.

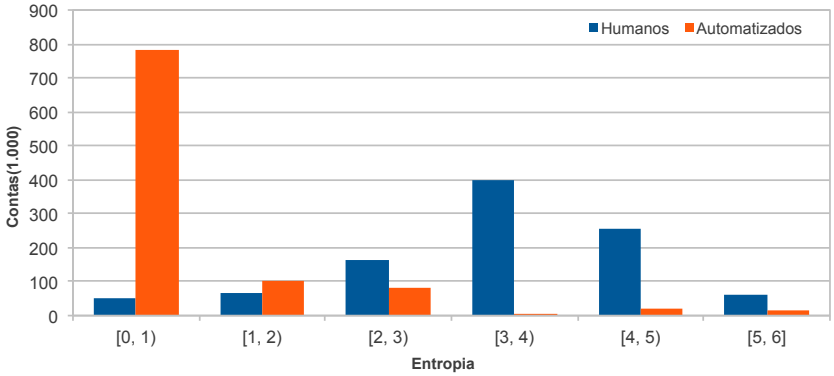

Figura 1. Entropia dos *Tweets*

Entropia dos Tópicos

De igual modo, a Figura 2 mostra que a entropia para o total de tópicos comentados é semelhante à entropia por *tweet*, uma vez que os *tweets* fazem referência aos tópicos.

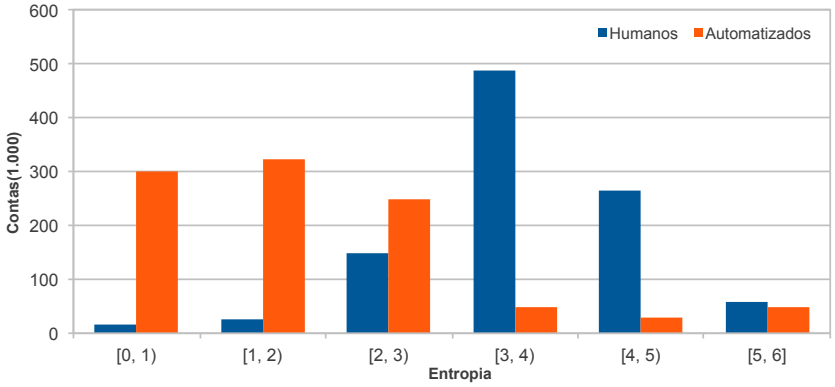

Figura 2. Entropia dos Tópicos

Entropia nos mesmos Tópicos e em Tópicos Diferentes

Na Figura 3 é possível notar que os usuários humanos, cerca de 87% que comentam em um mesmo tópico, possuem entropia considerada de baixa até média/alta, enquanto usuários automatizados, cerca de 82%, possuem entropia na faixa entre muito baixa e baixa. Usuários humanos que comentam várias vezes em um mesmo tópico não chegam a diversificar tanto o vocabulário, já que o assunto de todos seus *tweets* será específico para o tópico em questão. Por outro lado, os usuários automatizados postam

tweets massivamente para elevar a popularidade de um ou mais tópicos e acabam por repetir *hashtags* e URLs de maneira frequente nos *tweets*.

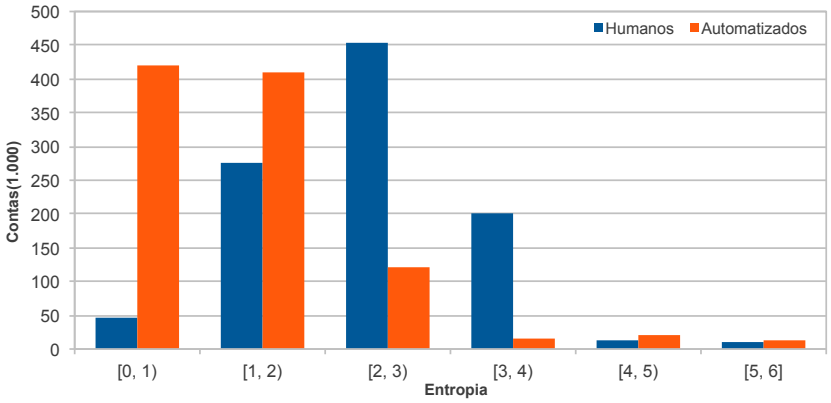

Figura 3. Entropia no mesmo Tópico

Entropia em Tópicos Diferentes

Na Figura 4, para *tweets* publicados em tópicos diferentes, nota-se que os usuários humanos, cerca de 94%, possuem entropia considerada alta ou muito alta, enquanto os usuários automatizados, cerca de 95%, possuem entropia considerada muito baixa a média. Isso reflete a maior divergência em termos de vocabulário, uma vez que, mesmo para tópicos diferentes, usuários automatizados publicam *tweets* com pouca diversidade de palavras, enquanto humanos escrevem de maneira natural e espontânea.

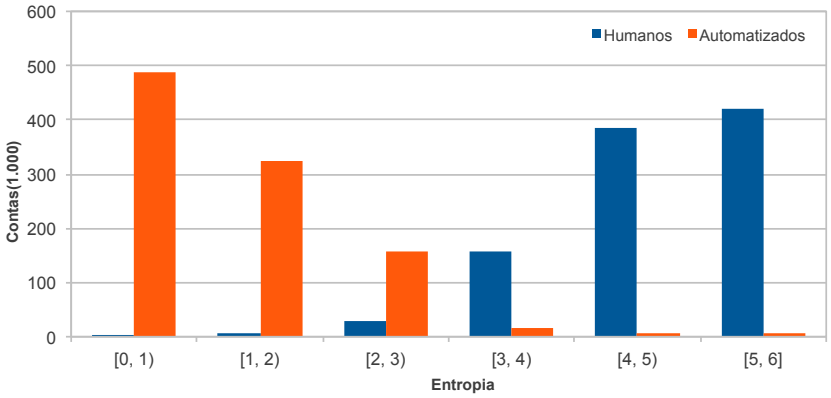

Figura 4. Entropia em Tópicos diferentes

5.3. Redução do Conjunto de Atributos

A fim de avaliar o poder discriminativo dos três (3) conjuntos de atributos separadamente, foram testados os atributos do conteúdo (C), do usuário (U) e de entropia (E) no classificador RandomForest. A Tabela 5 apresenta o desempenho obtido para cada um destes conjuntos ou ainda para mais de um conjunto.

Os resultados para o conjunto de atributos de conteúdo são os melhores, conseguindo identificar a maior parte dos *bots* sociais em relação aos outros conjuntos de

Tabela 5. Resultado do RandomForest

Atributos	Humanos	Automatizados
C	89,80%	86,95%
U	79,90%	82,90%
E	86,20%	83,00%
C+U	91,80%	91,70%
C+E	92,80%	92,10%
U+E	91,80%	90,50%
C+U+E	94,73%	92,40%

atributos. Em seguida está o conjunto de atributos baseados na entropia do texto das mensagens e por último os atributos do usuário.

Os atributos de conteúdo juntamente com os atributos de entropia representam o melhor resultado, ou seja, revelam que a forma como os usuários escrevem seus *tweets* é a principal característica para distinguir humanos de *bots*. O motivo é simples, esses dois conjuntos de atributos analisam justamente a forma como os *tweets* são escritos, quanto de similaridade possuem entre si e também o quanto a entropia do texto dos *tweets* é representativa para cada usuário.

Os atributos de usuários mostram-se, nesta escala, com menor poder discriminativo tanto isoladamente quanto em conjunto com outra categoria de atributos. Isso se deve ao fato principalmente de que muitos atributos do usuário não fazem distinção entre as classes humanos e *bots*, justamente porque na base final existem muitos *bots* ativos que participam normalmente dos tópicos de tendências postando notícias e atualizações de blogs e sites externos.

Comparando os resultados dos grupos de características avaliados neste artigo (conteúdo, usuários e entropia) com os alcançados no trabalho de [Freitas et al. 2014] (grupos usuário, conteúdo e linguístico), ambos acertando um pouco mais de 92% na identificação de *bots* sociais, percebe-se que a maneira como essas contas automatizadas escrevem seus *tweets* é fator determinante para sua detecção e caracterização. No caso da avaliação deste artigo, feita sobre os tópicos de tendência do Brasil, o uso da entropia para medir a frequência de postagens e escrita dos *tweets* demonstra ser eficaz na diferenciação de *bots* e usuários comuns.

6. Conclusão

Esta artigo apresentou seis (6) novas características, baseadas no conceito de entropia, aplicadas a detecção de comportamento automatizado no Twitter. Essas novas características permitem: (i) medir e diferenciar o padrão de escrita; (i) mensurar o vocabulário dos usuários; (iii) inferir quão diversificado o vocabulário de um usuário é; (iv) e classificá-lo com comportamento automatizado ou não.

Os resultados, utilizando aprendizagem de máquina, mostram que a junção dos atributos de usuário e de conteúdo com os atributos propostos permitem detectar comportamento automatizado nos tópicos de tendência do Twitter no Brasil. Empregando o classificador RandomForest, alcançou-se 92,40% de usuários automatizados e 94,73% de usuários humanos classificados corretamente. É necessário ressaltar que os resulta-

dos apresentados são exclusivos para a base de dados utilizada neste trabalho, mas que o resultado é similar a outros trabalhos da literatura.

Além disso, este artigo apresentou um metodologia para implantação de um processo de detecção de comportamento automatizado nos tópicos de tendência do Twitter. A importância deste trabalho se deve ao fato de que embora o próprio Twitter proíba a postagem de *tweets* automatizados nos tópicos de tendência, tal prática tem se tornado cada vez mais comum, podendo simplesmente degradar a experiência dos usuários como também expô-los a ataques e atividades maliciosas.

Como trabalho futuro é necessário criar grupos de atributos cada vez mais robustos e que representem, com maior relevância, comportamento automatizado. Outro ponto a ser considerado é criar mecanismos de detecção online de comportamento automatizado e/ou malicioso.

Referências

Alpaydin, E. (2010). *Introduction to Machine Learning*. The MIT Press, 2nd edition.

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

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

Dickerson, J. P., Kagan, V., and Subrahmanian, V. S. (2014). Using sentiment to detect bots on twitter: Are humans more opinionated than bots? In *2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2014, Beijing, China, August 17-20, 2014*, pages 620–627.

Freitas, C., Benevenuto, F., and Veloso, A. (2014). Socialbots: Implicações na segurança e na credibilidade de serviços baseados no twitter. In *Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos (SBRC)*.

Gara, T. (2013). One Big Doubt Hanging Over Twitter's IPO: Fake Accounts. `http://goo.gl/c1cSR4`.

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

Kartaltepe, E., Morales, J., Xu, S., and Sandhu, R. (2010). Social network-based botnet command-and-control: Emerging threats and countermeasures. In *Applied Cryptography and Network Security*, volume 6123 of *Lecture Notes in Computer Science*, pages 511–528. Springer Berlin Heidelberg.

Lee, K., Caverlee, J., and Webb, S. (2010). Uncovering social spammers: Social honeypots + machine learning. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '10, pages 435–442, New York, NY, USA. ACM.

Malware Bytes (2014). Twitter Phishing Spamrun: “Strange Rumors About You’. https://blog.malwarebytes.org/fraud-scam/2014/09/twitter-phishing-spamrun-strange-rumors-about-you/.

Martinez-Romo, J. and Araujo, L. (2013). Detecting malicious tweets in trending topics using a statistical analysis of language. *Expert Systems with Applications*, 40(8):2992 – 3000.

Nazario, J. (2009). Twitter-based Botnet Command Channel. https://asert.arbornetworks.com/twitter-based-botnet-command-channel/.

Neal Ungerleider (2015). Almost 10% of Twitter is Spam. http://www.fastcompany.com/3044485/almost-10-of-twitter-is-spam.

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

Sokolova, M., Japkowicz, N., and Szpakowicz, S. (2006). Beyond accuracy, f-score and roc: A family of discriminant measures for performance evaluation. In *Proceedings of the 19th Australian Joint Conference on Artificial Intelligence: Advances in Artificial Intelligence*, AI’06, pages 1015–1021, Berlin, Heidelberg. Springer-Verlag.

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.

Symantec (2013). Phishing: The Easy Way to Compromise Twitter Accounts. http://goo.gl/KJfu39.

Twitter (2014). Denunciar spam no Twitter. https://support.twitter.com/articles/263349-como-denunciar-por-spam-no-twitter.

Wang, A. (2012). Machine learning for the detection of spam in twitter networks. In *e-Business and Telecommunications*, volume 222 of *Communications in Computer and Information Science*, pages 319–333. Springer Berlin Heidelberg.

Wang, A. H. (2010). Don’t follow me: Spam detection in twitter. In *Security and Cryptography (SECRYPT), Proceedings of the 2010 International Conference on*, pages 1–10.

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.