Caracterização e Classificação de *Bots* Utilizando a Rede de Comentários do Reddit

Rafael G. Damasceno[1], **Daniel R. Figueiredo**[1]

[1]Programa de Engenharia de Sistemas e Computação (PESC)
Instituto Alberto Luiz Coimbra de Pós-Graduação e Pesquisa de Engenharia (COPPE)
Universidade Federal do Rio de Janeiro (UFRJ)
Rio de Janeiro – RJ – Brasil

{damasceno,daniel}@cos.ufrj.br

Abstract. *Online social networks play an increasingly important role in society, given its influence on the opinion formation of millions of users. However, the presence of bots (users with programmed behavior) emerges as a concern given their potential to dissiminate information and consequently influence opinion formation, evidencing the need for its characterization and identification. This work considers a directed network of users constructed from comments in Reddit to characterize and identify bots. The network characterization highlights the significant structural differences of bots, allowing them to be classified with high accuracy using only network features.*

Resumo. *Redes sociais online desempenham um papel cada vez mais importante na sociedade, tendo em vista sua influência na formação de opinião de milhões de usuários. Contudo, a presença de bots (usuários com comportamento programado) surge como uma preocupação diante do potencial para disseminação de informação e consequente influência na formação de opinião, evidenciando a necessidade de sua caracterização e identificação. Este trabalho considera uma rede direcionada de usuários construída a partir de comentários do Reddit para caracterizar e identificar bots. A caracterização da rede evidencia diferenças estruturais significativas dos bots, permitindo sua classificação com alta acurácia utilizando apenas características da rede.*

1. Introdução

Redes sociais online (RSO) tais como Facebook e Twitter desempenham um papel central na sociedade moderna, atuando também como meio para divulgação de notícias e debate de opiniões, atingindo milhões de pessoas diariamente. Não por menos, o papel destas redes vem sendo constantemente debatido, principalmente durante momentos marcantes, tais como revoluções populares e eleições democráticas. Um ponto de destaque nestas discussões está acerca do fenômeno conhecido como epidemia em redes, o qual permite disseminar uma informação - verdadeira ou falsa - através dos relacionamentos entre usuários da rede.

Neste contexto, o espaço das redes sociais acaba se comportando como um ambiente formador de opinião. Recentemente, escândalos envolvendo o uso de dados das redes para influenciar opiniões de eleitores em candidaturas presidenciais ocuparam as capas dos jornais - como o caso envolvendo o Facebook, a empresa de análise de dados

Cambridge Analytica e a eleição do presidente norte americano Donald Trump[1]. Esta nova preocupação em torno da manipulação de opiniões levanta a atenção para comportamentos voltados unicamente para este propósito.

A presença de *bots*[2] nestas redes ocupa um papel particular dentro da referida discussão. Estes usuários podem possuir objetivos bem definidos, como realizar uma determinada interação com outro usuário sempre que um dado padrão ocorrer. Em certos casos, todavia, este objetivo não é tão claro para os outros usuários, sobretudo quando o *bot* busca se comportar e interagir como um usuário humano.

O uso de *bots* em RSO, em especial nos últimos anos, provoca uma série de discussões sobre o seu papel dentro da rede e as consequências de seu comportamento, principalmente no que tange a influência de opiniões e a propagação de conteúdo. Além disso, parte destes comportamentos podem ferir a política de uso da RSO, demandando esforços para identificar, controlar e, em alguns casos, reduzir os efeitos causados. Assim sendo, a identificação de tais usuários se torna cada vez mais importante, permitindo adotar as medidas cabíveis em cada caso. De fato, diversos trabalhos recentes atacam este problema utilizando diferentes metodologias para diferentes RSOs (ver discussão na seção 5).

Um aspecto de destaque deste trabalho está na construção da rede de usuários a ser explorada. Em particular, os vértices representam os usuários e uma aresta direcionada indica que o usuário na origem da aresta fez um comentário em uma mensagem escrita pelo usuário destino da aresta. Desta forma, os comentários dos usuários do Reddit serão usados para construir a rede (direcionada e com pesos) de relacionamentos. Repare que esta rede reflete diretamente a principal atividade do Reddit que, por sua vez, são as discussões através dos comentários dos usuários.

O objetivo do presente trabalho é a caracterização e identificação de *bots* na rede social Reddit, independentemente de seus objetivos. Diferentemente de outras abordagens, este trabalho visa explorar apenas a estrutura da rede entre usuários. Sendo assim, o conteúdo das mensagens escritas por usuários ou mesmo os nomes dos usuários (i.e., login), não serão empregados. O objetivo é evitar a subjetividade inerente ao conteúdo, tendo em vista as muitas formas de expressão escrita nas RSOs, e utilizar aspectos mais robustos, como a troca de mensagens entre pessoas.

O presente trabalho possui as seguintes principais contribuições:

- A construção da rede de usuários, induzida por comentários através de um *dataset* real com mais de 91 milhões de comentários.
- Caracterização da estrutura da rede do ponto de vista de usuários reais e *bots*. Esta caracterização evidencia a grande diferença estrutural entre usuários normais e *bots*. Uma descoberta interessante é o alto número de respostas para *bots* e número reduzido de comentários de *bots* para um mesmo usuário (a ser discutido).
- Identificação automática de *bots* utilizando uma rede neural como modelo de classificação, tendo como entrada apenas características estruturais da rede de usuários. Apesar do grande desbalanceamento entre as classes (*bots* e não-*bots*),

[1] Segundo o The New York Times, os dados coletados pela Cambridge Analytica incluíam detalhes sobre as identidades dos usuários, redes de amigos e "curtidas" [Rosenberg et al. 2018].

[2] Define-se *bot* como um algoritmo que produz conteúdo automaticamente e interage com seres humanos nas mídias sociais, tentando imitar e possivelmente alterar seu comportamento [Ferrara et al. 2016].

o modelo exibiu um bom desempenho, atingindo um AUC de 0.91 na ROC, indicando a eficácia das características estruturais para a classificação.

O restante deste documento está organizado da seguinte forma. Na seção 2 a metodologia para construção da rede é apresentada, assim como detalhes do *dataset* utilizado. A seção 3 apresenta a caracterização da rede mostrando as diferenças entre usuários normais e *bots*. A identificação de *bots* utilizando apenas características estruturais como entrada de uma rede neural é apresentada na seção 4. Trabalhos relacionados são expostos na seção 5 e a seção 6 apresenta algumas considerações finais.

2. *Dataset* e Modelagem da Rede

O Reddit[3] se define, em tradução livre, como lar de milhares de comunidades, conversas intermináveis e conexões humanas autênticas. A rede social, com média de usuários ativos mensais maior que 330 milhões, possui como foco as interações através de postagem, votação (positiva ou negativa) e comentário.

As postagens são realizadas pela comunidade da rede, que pode compartilhar conteúdo postando histórias, links, imagens e vídeos. Em seguida, a comunidade pode comentar estas postagens - é importante destacar que cada comentário também pode receber comentários e assim sucessivamente. Por fim, tanto as postagens como os comentários podem ser votados, com o intuito de destacar o conteúdo mais interessante para a rede.

Mais detalhes sobre a construção da rede e o processamento do *dataset* podem ser encontrados em [Damasceno 2019].

2.1. *Dataset*

O *dataset* utilizado neste trabalho é disponibilizado através do acervo pushshift [Baumgartner 2018]. Neste acervo, é possível encontrar diferentes conjuntos de dados da plataforma Reddit coletados e separados por dia ou mês.

Para este trabalho, utilizou-se os dados de comentários da rede social, Reddit, no mês de janeiro de 2018. O arquivo compreende os comentários de todos os usuários neste período, um total de 91.558.594, onde cada linha se apresenta como um *JSON* de um total de 20 atributos. Nestes atributos estão as informações daquele comentário, desde identificadores até seu conteúdo.

Dentre as informações disponibilizadas, as relevantes para a construção da rede avaliada deste trabalho são os campos "*author*", "*id*" e "*parent_id*". O campo "*id*" identifica o comentário que foi realizado pelo usuário no campo "*author*". Enquanto o campo "*parent_id*" identifica o comentário que foi respondido pelo atual comentário (um comentário no Reddit é sempre feito em cima de outro comentário). É importante notar que todo usuário e todo comentário é identificado unicamente no sistema.

É importante ressaltar que o usuário que recebeu o comentário, ou seja, o autor do comentário identificado em "*parent_id*", é desconhecido. As consequências desta característica do *dataset* para a modelagem da rede serão abordadas na seção 2.3.

[3]https://www.reddit.com

2.2. Lista de usuários *bots*

Para utilizar a rede de usuários mostra-se necessário identificar, dentre os usuários presentes, ao menos uma parcela dos usuários *bots*. Esta caracterização permite comparar o comportamento de usuários *bots* com os outros e, posteriormente, avaliar as métricas.

Com o objetivo de compor uma lista de usuários *bots*, foram agrupados os resultados de diferentes fontes descritas a seguir. Os nomes dos usuários obtidos como resultado foram adicionados na forma minúscula em um *set* para evitar duplicatas e permitir uma comparação sem ambiguidades com os nomes dos usuários do *dataset* de comentários.

Inicialmente, foram realizadas requisições a partir da API da plataforma pushshift [Baumgartner 2018] para as postagens do usuário *BotBust*[4], este usuário *bot* apresenta outros usuários *bots* identificados e banidos dentro de tópicos da rede social. Os critérios para o banimento de um *bot* incluem realizar comentários sem serem convidados, não fornecer valor para a comunidade ou simplesmente não funcionar corretamente, porém, o banimento ocorre de forma manual a partir de usuários moderadores. Este processo coletou um total de 578 usuários identificados como *bot*.

Em seguida, foram utilizadas postagens[5,6] apresentando listagens de usuários *bots* populares - estas listas não utilizaram a estrutura da rede, somente avaliação de moderadores e do tempo de resposta dos usuários. Agrupando cada uma das listas, permitiu-se aumentar o número de usuários *bots* conhecidos para 989 e posteriormente 1055.

Dentre os 1055 usuários, é esperado que apenas uma parcela seja reconhecida dentro da rede construída para este trabalho (como apresentado na seção 3). O motivo para isto está no tamanho limitado do *dataset* de comentários, restrito ao período de um mês. Além disso, espera-se que uma parcela dos usuários presentes na rede construída sejam *bots* e não estejam identificados como tal.

2.3. Modelagem da Rede

A rede modelada para este trabalho, a partir do *dataset* apresentado na seção 2.1, apresenta-se como uma rede de usuários induzida por comentários. Os vértices da rede são os usuários e as arestas entre os pares de usuários indicam que o usuário de origem da aresta respondeu em algum comentário do usuário de destino da aresta. Ou seja, existe uma aresta direcionada $u \rightarrow v$ somente se o usuário u comentou em um comentário do usuário v. Além disso, cada aresta possui um peso associado $w_{u,v}$ que indica o número de comentários realizados por u em comentários de v.

As informações disponibilizadas para cada comentário no *dataset* indicam o usuário autor do comentário, o identificador do comentário realizado por este autor e o identificador do outro comentário, que sofreu o comentário do autor. Logo, o usuário que recebeu o comentário é desconhecido ao analisar somente as informações para uma única entrada do *dataset*. Esta informação está ilustrada na Figura 1a.

Como a rede deve possuir relacionamento entre usuários, constrói-se inicialmente uma série de pares de relacionamentos entre o usuário autor e o identificador do comentário que o autor está se dirigindo (que, por sua vez, pertence a algum usuário, ainda

[4]https://www.reddit.com/user/BotBust

[5]https://www.reddit.com/r/autowikibot/wiki/redditbots

[6]https://www.reddit.com/r/dataisbeautiful/comments/9mh3pn/oc_the_50_most_active_bots_on_reddit_based_on

desconhecido). Ao mesmo tempo, também é construído um dicionário (*hash table*) que indica para cada identificador de comentário quem é o usuário autor.

Figura 1. Construção da rede a partir de dados de comentários: (a) analisando somente um comentário; (b) analisando um comentário e as informações dos outros comentários do *dataset*.

Após a etapa inicial, as arestas da rede podem ser geradas percorrendo os pares de relacionamentos entre o usuário autor e o identificador do comentário que o autor está se dirigindo, substituindo o identificador do comentário pelo usuário correspondente - utilizando o dicionário construído. Caso o usuário correspondente não seja encontrado, esta aresta é descartada, visto que não poderia ser traçado o relacionamento entre dois usuários com os dados coletados.

Conclui-se que cada usuário que aparece na rede precisa necessariamente ter comentado ao menos uma vez, dado que somente desta forma o seu nome de usuário é conhecido. No entanto, isto não implica que o seu grau de saída na rede será obrigatoriamente maior que zero, na medida em que o usuário do comentário ao qual ele se dirige pode não aparecer na rede (ou seja, o autor do comentário ao qual ele se dirige é desconhecido pois o comentário não está no *dataset* e consequentemente o identificador do comentário não está presente no dicionário construído).

Seguindo esta construção, pode-se concluir que, a partir da Figura 1a, caso o comentário de identificador *dnqijyp* também esteja presente no *dataset*, será possível conhecer o seu autor e, assim, construir o relacionamento entre usuários apresentado na Figura 1b (supondo que o autor do comentário *dnqijyp* é o usuário *IndividualTwo8*).

Analisando os dados do *dataset*, foi possível perceber que parte dos autores dos comentários aparecem com o campo de nome de usuário igual à "*[deleted]*", o que indica que este usuário não existia mais na rede social no momento em que o *dataset* utilizado foi formado. Tratar estes usuários como parte da rede iria agrupar um conjunto de usuários diferentes como um único, deixando de representar a rede real e prejudicando as métricas de estudo. Logo, todos estes comentários foram ignorados para a construção da rede, restando um total de 85.006.711 comentários - aproximadamente 93% dos 91.558.594 comentários totais do *dataset* - realizados por 4.414.144 usuários diferentes.

Seguindo os procedimentos descritos acima para obter o usuário autor de cada comentário e os relacionamentos entre os pares, foi construído o conjunto de arestas (*edge list*) que compõem a rede, o resultado é a rede induzida a partir dos comentários, cujas características estão na Tabela 1. Em particular, a tabela indica o número de usuários presentes na rede (número de vértices), o número de relacionamentos entre os usuários (número de arestas), o número total de comentários na rede (soma dos pesos das arestas), e o número de componentes conexas.

Vértices	3.541.145
Arestas	38.452.098
Densidade	$3,07 \times 10^{-6}$
Soma do peso das arestas	48.754.283
Componentes Fracamente Conexas	21.872
Componentes Fortemente Conexas	969.314

Tabela 1. Avaliação preliminar da rede induzida a partir dos comentários

Analisando a Tabela 1 em comparação com os dados do *dataset* (já após a remoção dos usuários que aparecem como deletados) a rede induzida representa uma redução significativa para o número de comentários, aproximadamente 43%. Esta redução está diretamente relacionada à limitação, imposta pelo *dataset*, de não conhecer o usuário autor do comentário que recebeu resposta.

Durante a criação do conjunto de arestas para a rede, os relacionamentos em que o usuário de uma das extremidades não é conhecido são descartados. Por consequência, como são armazenados somente os relacionamentos entre os pares, os usuários que não possuem nenhuma arestas (ou seja, não realizam ou recebem comentários de um usuário conhecido) são omitidos da rede, significando uma redução de aproximadamente 20% do número de usuários. Apesar da expressividade, os usuários omitidos representam somente componentes conexas de tamanho 1 e não afetam a análise da maior componente conexa.

3. Caracterização Estrutural de Usuários e *Bots*

A rede modelada neste problema pode induzir diferentes grupos de usuários agrupados em diferentes componentes conexas da rede. Por exemplo, determinados usuários podem comentar apenas em questões de um determinado assunto específico. Por outro lado, podem existir usuários que comentem em diferentes assuntos, criando caminhos na rede entre estes diferentes grupos.

Com o objetivo de saber mais sobre a organização da rede induzida, e principalmente a presença dos *bots*, pode-se observar a distribuição das componentes fortemente conexas na Tabela 2. Cabe destacar que foi encontrado um total de 207 usuários *bots* na rede ao comparar os nomes dos usuários com os nomes dos *bots* coletados na seção 2.2.

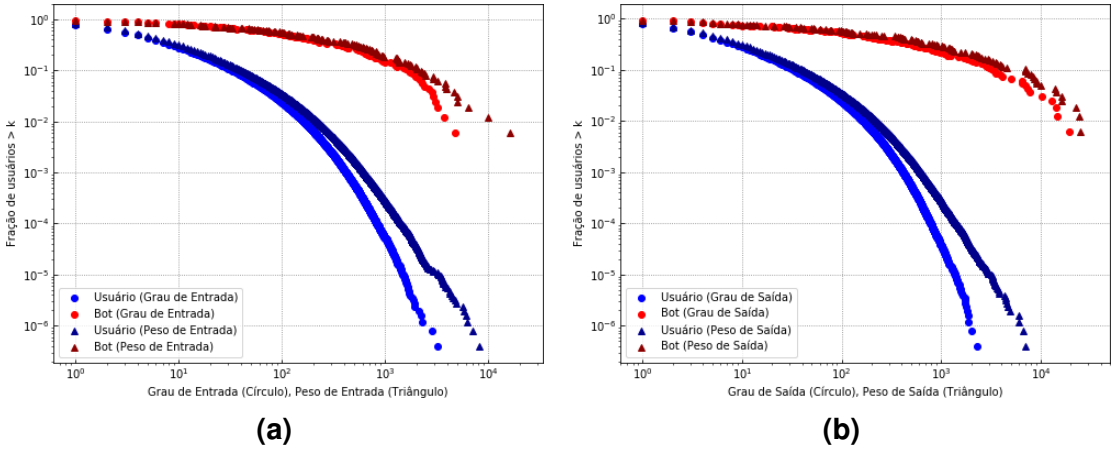
Tamanho da Componente	Número de Componentes	Número de Bots
1	955.182	43
2	13.352	1
3	661	0
4	77	0
5	25	0
6	3	0
7	7	0
9	3	0
14	1	0
28	1	0
268	1	0
2.556.439	1	163
	Total de Bots na Rede	207

Tabela 2. Avaliação das Componentes Fortemente Conexas da Rede

Analisando a Tabela 2, pode-se observar que existem 43 *bots* em componentes fortemente conexas de tamanho 1, ou seja, 43 componente formadas somente por um único *bot*. Verifica-se, também, um único *bot* em uma das componentes de tamanho 2. E, por fim, todos os *bots* restantes estão presentes na única maior componente fortemente conexa. Para as componentes fracamente conexas da rede, o mesmo comportamento pode ser observado, apresentando somente um *bot* fora da maior componente. Para este trabalho serão avaliadas as métricas com base na maior componente fortemente conexa.

3.1. Métricas estruturais

Nas Figuras 2 e 3 são apresentadas algumas das principais características extraídas da Maior Componente Fortemente Conexa. As marcações dentro dos gráficos como *bot* são referentes aos dados coletados de acordo com a seção 2.2.



(a) **(b)**

Figura 2. Função de distribuição cumulativa complementar (CCDF) do grau e peso dos vértices: (a) para o grau de entrada e peso de entrada; (b) para o grau de saída e peso de saída.

Observando o gráfico da Figura 2a, pode-se perceber que a diferença entre a curva do grau de entrada e do peso de entrada começa a aumentar para valores em ordem menor para os usuários do que para os *bots*. Este comportamento pode indicar a presença mais provável de uma discussão entre usuários comuns, dado que a curva do peso se afasta da curva do grau quando as arestas direcionadas entre os pares de usuários codificam a presença de mais comentários.

Além disso, é possível perceber que os usuários *bots* recebem proporcionalmente mais comentários direcionados a eles do que os usuários comuns. Apesar de parecer, em um primeiro momento, um comportamento estranho, esta observação pode caracterizar situações onde o usuário emite um comentário relevante e/ou controverso, causando uma série de respostas apoiando ou contestando aquela informação. Desse modo, faz sentido esperar um grau de entrada maior para *bots* que desempenham este comportamento.

Observando o gráfico da Figura 2b, pode-se perceber que a curva do grau de saída e do peso de saída começam a se distanciar para os usuários comuns após um certo valor, enquanto permanecem praticamente juntas para toda a curva de usuários bots.

Como cada valor unitário de peso indica um comentário, o comportamento normal é que a curva do peso de saída apresentasse valores próximos ou maiores que o grau de

saída. Entretanto, é notório que a curva do peso de saída para os *bots* se distancia de forma limitada da curva de grau de saída, o que indica a maioria de arestas com peso unitário. Em outras palavras, indica que há pouco envio de comentários como resposta para um mesmo usuário.

Dessa forma, os usuários *bots* parecem comentar de forma mais indiscriminada, com peso das arestas tendendo a ser mais próximo de 1, enquanto usuários comuns acabam comentando para o mesmo usuário mais de uma vez, podendo indicar, como na Figura 2a, uma discussão entre os pares de usuários.

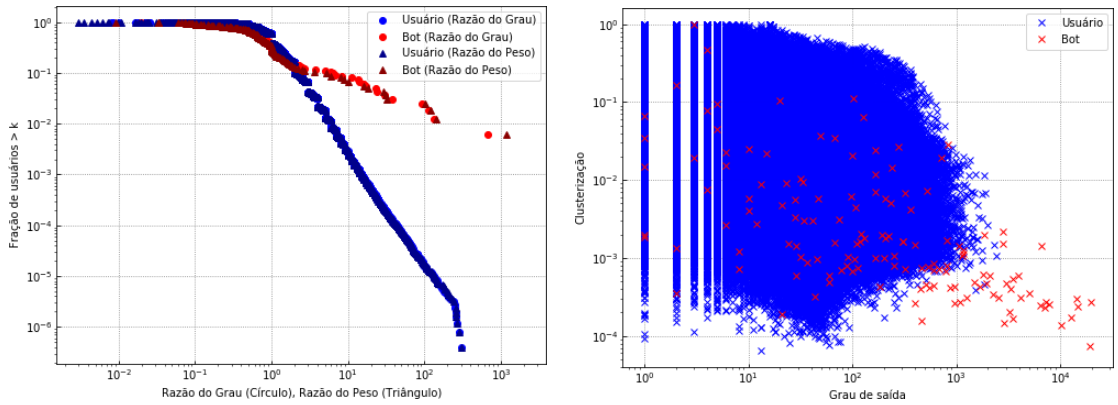


Figura 3. (a) Função de distribuição cumulativa complementar (CCDF) da razão do grau e razão do peso dos vértices; (b) Gráfico de dispersão entre o grau de saída e o coeficiente de clusterização dos vértices.

Observando o gráfico da Figura 3a, pode-se perceber que para valores abaixo de 1 ou pouco maiores (o grau ou peso de entrada é menor ou próximo ao de saída) existe pouca distinção entre usuários comuns e *bots*. Contudo, a partir de um certo valor é possível perceber uma proporção maior de usuários *bots* com o grau ou peso de entrada maiores que os de saída, quando comparado com usuários comuns.

Este comportamento indica que um mesmo usuário *bot* acaba, por vezes, recebendo mais comentários do que realizando. Como elucidado anteriormente, a propagação de certas informações pode acarretar no recebimento de múltiplas respostas concordando ou contestando aquela informação. Ademais, usuários *bots* parecem não se envolver em discussões diretas sobre seus comentários, como considerado a partir da Figura 2b. Desta forma, o comportamento da curva se mostra justificável pelo comportamento dos usuários.

Diante das métricas apresentadas anteriormente nas Figuras 2 e 3a, pode-se conferir os valores mínimo, máximo, médio e a mediana para cada uma delas na Tabela 3.

	Mínimo	Máximo	Média	Mediana
Grau de Entrada	1	13.485	14,45	4
Soma dos Pesos de Entrada	1	20.660	18,45	5
Grau de Saída	1	19.740	14,45	4
Soma dos Pesos de Saída	1	29.930	18,45	5
Razão de Grau (entrada / saída)	$8,16 \times 10^{-3}$	1.124,00	1,28	1,00
Razão de Peso (entrada / saída)	$3,00 \times 10^{-3}$	2.005,00	1,29	1,00

Tabela 3. Métricas para a Maior Componente Fortemente Conexa da Rede

A Figura 3b apresenta um gráfico de dispersão, buscando visualizar a distribuição da rede e características mais comuns para usuários *bots*. Avaliando este gráfico, nota-se que para valores de grau de saída inferiores à 10^3, existe um grande embaralhamento entre usuários comuns e usuários *bots*. Apesar disso, para valores de grau de saída acima deste ponto, pode-se observar que para coeficientes de clusterização muito baixos existe uma concentração quase exclusiva de usuários *bots*. Dentre os poucos usuários que aparecem como comuns nesta região, existe uma probabilidade considerável de serem, na verdade, *bots* não classificados dentre os dados deste trabalho.

Esse padrão apresentado parece comum para um usuário *bot*, visto que este deve comentar de forma mais indiscriminada, como apontado na avaliação do gráfico na Figura 2b, diminuindo a probabilidade de comentários ocorrerem entre os usuários respondidos pelo *bot* (menor clusterização). Além disso, como o funcionamento do usuário *bot* acontece de forma autônoma, faz-se possível um alto grau de saída, ou seja, comentários para muitos usuários diferentes num dado período.

Outras métricas de interesse para a componente estão apresentadas na Tabela 4

	Mínimo	Máximo	Média	Mediana
Coeficiente de Clusterização	0	1	$4,04 \times 10^{-2}$	$9,07 \times 10^{-4}$
Reciprocidade Local	0	1	$3,48 \times 10^{-1}$	$2,92 \times 10^{-1}$
Reciprocidade Local (usuários)	0	1	$3,48 \times 10^{-1}$	$2,92 \times 10^{-1}$
Reciprocidade Local (bots)	0	1	$2,03 \times 10^{-1}$	$1,66 \times 10^{-1}$
Reciprocidade da Rede	-	-	$2,78 \times 10^{-1}$	-

Tabela 4. Outras métricas para a Maior Componente Fortemente Conexa da Rede

Analisando a Tabela 4, é interessante apontar os valores relativamente altos para a reciprocidade local (definida como a fração de arestas que ocorrem nos dois sentidos [Newman 2010]). A média deste valor indica que $\approx 35\%$ dos comentários realizados acabam sendo respondidos de volta, como uma discussão ou debate entre usuários. Obviamente, a métrica não garante que os usuários possuem um relacionamento recíproco diante dos mesmos comentários, mas devido ao tamanho da rede é mais esperado que isto ocorra em um mesmo tópico de discussão. Esta métrica é consideravelmente menor quando observada somente para usuários *bots*, $\approx 20\%$, reforçando a ideia de que estes usuários não se correspondem muito com outros usuários e mantém uma comunicação mais indiscriminada.

Os valores encontrados para o coeficiente de clusterização (definida como a média da fração de arestas entre os vizinhos [Barabási 2013]) indicam que a rede não possui uma alta propensão para triângulos se formarem entre um vértice e os seus vizinhos. Em outras palavras, ocorrência de três comentários entre usuários que forme um ciclo (triângulo).

Por fim, vale ressaltar que, existem outras métricas estruturais da rede que poderiam ser exploradas nesta seção, como, por exemplo, o PageRank para determinar a importância dos vértice na rede.

4. Classificação de Perfis Falsos

A partir das métricas extraídas na seção 3.1, é viável imaginar como estas características podem ser empregadas para classificação de um usuário qualquer, seguindo algum pro-

cesso dentro da computação. Desta forma, é possível conceber algum modelo classificador de aprendizado de máquina (ou vários), que lide com as características de cada usuário e os classifique em uma das possíveis classes: ”*bot*” ou ”*não bot*”.

Este trabalho explora um modelo e observa o seu resultado, verificando se consegue gerar uma predição razoável para o problema, indicando o potencial no emprego deste paradigma de rede e conceitos de Redes Complexas.

O problema deste trabalho trata de classes desbalanceadas, onde, como observado na Tabela 2, referente à componente utilizada para avaliação, apenas uma pequena parcela de usuários são conhecidos como *bot*. Caso esta característica não seja tratada, o modelo tende a favorecer a classe que contém o maior número de amostras, neste caso, apontando todos os usuários como ”*não bot*”.

4.1. Avaliação e Tratamento das Métricas para o Modelo

Dentro desta ideia de conceber um modelo classificador que lide com as características de cada usuário, vale analisar a Figura 4, com a matriz de correlação para as características dos vértices na Maior Componente Fortemente Conexa da rede.

Como esperado, é possível reparar uma forte correlação entre métricas derivadas de outras, como, por exemplo, soma dos graus de entrada e saída com as métricas individuais. Mas também é possível reparar correlações entre métricas como o grau de entrada e o grau de saída, como foi observado também em análise das métricas a partir dos gráficos da função de distribuição cumulativa complementar (CCDF) na seção 3.1.

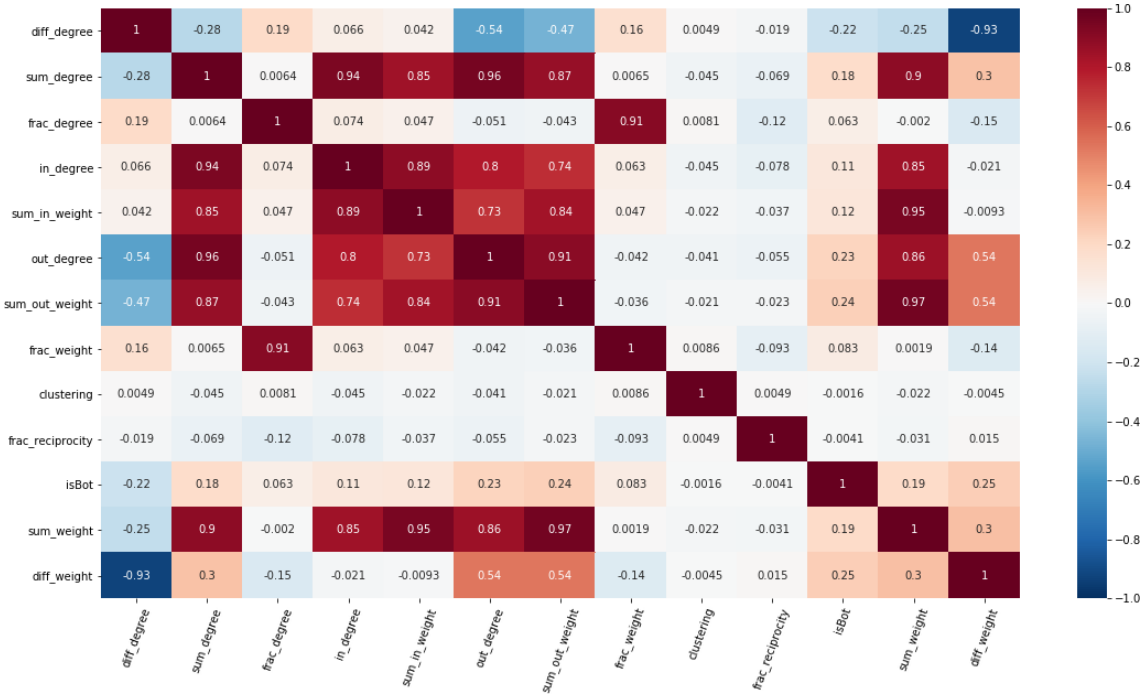


Figura 4. Matriz de Correlação para as características dos vértices na Maior Componente Fortemente Conexa

Analisando cuidadosamente a matriz de correlação é possível remover parte das características de alta correlação sem prejudicar um modelo classificador adotado. Este

processo pode simplificar o modelo, por reduzir a dimensionalidade do problema, e melhorar a predição. Ressalta-se, todavia, que este trabalho não se preocupa em verificar o melhor conjunto de características para aplicação dentro de um modelo classificador, visto que este processo requer uma avaliação dentre diferentes combinações para evitar a perda de características importantes para a classe de interesse.

Por outro lado, como as métricas de avaliação possuem diferente variações entre os valores mínimos e máximos, todas as métricas que não eram classificadas entre zero e um foram normalizadas. Este processo simples pode ser capaz de determinar se um modelo neural é capaz ou não de aprender com o passar de um número determinado de épocas. Quando as variações de valores são muito desiguais, isto pode fazer com que os gradientes oscilem muito até encontrar o caminho para o mínimo global ou local. Em contrapartida, as métricas normalizadas permitem que o gradiente possa convergir mais rapidamente.

4.2. Aplicação e Avaliação de uma Rede Neural

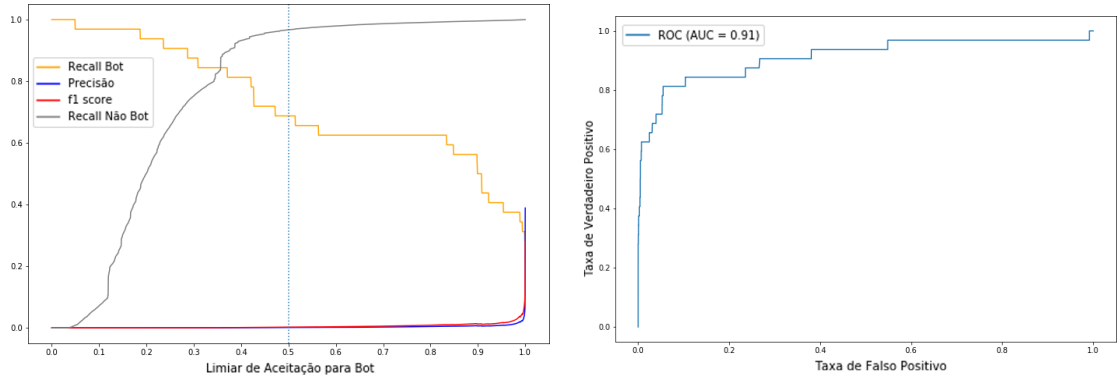
Para validar a ideia de classificação utilizando a estrutura da rede adotou-se um modelo relativamente simples de rede neural. O modelo adotado de rede neural sem realimentação (*feedforward*) contou com apenas duas camadas, sendo 62 neurônios presentes na camada de entrada e 2 neurônios de saída.

O modelo classificador foi utilizado com 12 variáveis de entrada (métricas extraídas a partir da estrutura da rede) e retorna a partir da sua camada de saída valores probabilísticos da classificação, onde os valores para as classes "*bot*" e "*não bot*" são complementares. Foram realizadas 200 épocas de treinamento, ou seja, o algoritmo estuda a amostra completa de dados um total de 200 vezes. Além disso, foram atribuídos pesos às classes de acordo com o desbalanceamento do problema.

Os dados de entrada foram divididos em 5 *folds* iguais, mantendo a proporção de cada classe "*bot*" e "*não bot*". Três desses *folds* foram agrupados e utilizados como conjunto de treinamento. Um *fold* foi utilizado para o conjunto de validação, ou seja, utilizado ao final de cada época para testar o desempenho parcial do modelo. Após o final do treinamento da rede neural, o último *fold* foi utilizado como conjunto de teste e os valores da predição deste conjunto serão utilizados para avaliar a classificação.

A Figura 5a, apresentada a seguir, contém um gráfico com as métricas de *recall* para as duas classes, "*bot*" e "*não bot*", além de precisão e *F1 score* somente para a classe "*bot*". Neste gráfico, o eixo x representa o valor a partir do qual uma probabilidade do classificador para uma dada amostra é considerada como uma classificação em "*bot*".

Avaliando as curvas das métricas, é possível perceber valores próximos de zero para as métricas de precisão e F1. Na verdade, estas métricas não são uma boa forma de avaliação para um problema desbalanceado como este. Isto pode ser verificado ao analisar a métrica de precisão, onde o valor máximo para o numerador, ou seja, verdadeiro positivo, será sempre muito pequeno devido a classe *bot* minoritária. Logo, qualquer valor de falso positivo (usuários comuns classificados como bots) irá prejudicar a métrica por estar no denominador. Fora isso, a classificação inicial do problema para *não bot* provavelmente não é totalmente correta, podendo existir usuários considerados comuns que são, na verdade, *bots*.



(a) **(b)**

Figura 5. Avaliação do modelo de Rede Neural: (a) para Recall, Precisão e F1 Score; (b) no Espaço ROC para métrica AUC.

Para avaliar melhor as taxas de *recall* foi adotando um limiar de 0,5 para a aceitação de *bots*, ou seja, amostras avaliadas pelo classificador com probabilidades iguais ou acima deste valor classificam o usuário como *bot*. Para este valor de limiar, foi possível obter um *recall* de $0,69$ para a classe *bot* e *recall* $0,97$ para a classe *não bot*.

Apesar de um acerto maior para a classe *não bot*, este comportamento é importante num contexto de uma aplicação que pode, por exemplo, suspender temporariamente usuários identificados como *bot*. Mesmo com interesse na classificação de usuários *bots* não seria interessante possuir alto erro para a classe oposta.

A Figura 5b apresenta a curva *ROC* (*Receiver Operating Characteristic*). Em problemas desbalanceados a decisão pelo mínimo erro da classificação favorece a classe majoritária. Dessa forma, a curva *ROC* busca avaliar o desempenho relacionado com o quanto o classificador acerta da classe positiva e o quanto deixa de errar da classe negativa [Zaki et al. 2014].

Para avaliar a curva da Figura 5b é utilizada a métrica *AUC* que indica a área sob a curva *ROC*. Esta métrica vale para problemas de duas classes [Zaki et al. 2014], como é o caso deste problema, e o resultado desta métrica permite obter um valor de avaliação do modelo e permite fazer comparações com outras propostas. Os valores da métrica são compreendidos entre zero e um, onde para obter valor máximo seria necessário taxa máxima de verdadeiro positivo ao mesmo tempo que taxa mínima de falso positivo.

Desta forma, o valor de $0,91$ encontrado caracteriza uma ótima classificação, visto o problema desbalanceado entre duas classes com interesse na classe minoritária *bot*.

5. Trabalhos Relacionados

A presença cada vez mais expressiva, em número e influência, de *bots* em redes sociais online vem chamando atenção da academia e indústria, e deu origem a um número expressivo de trabalhos diversos e recentes sobre o tema. As abordagens propostas na literatura para identificação de *bots* foi recentemente dividida em três classes: sistemas de detecção de *bots* baseados em informações de redes sociais; sistemas baseados em *crowdsourcing* (classificação humana manual); e, métodos de aprendizado baseado na identificação de características que são discriminantes entre *bots* e humanos [Ferrara et al. 2016]. Este

trabalho se enquadra na primeira categoria, pois explora informações provenientes da interação entre os usuários.

Uma das redes sociais mais estudadas é o Twitter, tendo em vista o número e o grau de atividade de seus usuários. Uma recente caracterização do comportamento de *bots* e usuários comuns a partir de 65 milhões de tweets identificou uma grande diferença na quantidades de publicação de tweets e reciprocidade de amizades, entre outros [Gilani et al. 2017]. Além da caracterização, diferentes trabalhos visam a identificação de *bots* no Twitter, em geral utilizando características extraídas do conteúdo e metadados dos tweets e usuários que servem de entrada para algoritmos de classificação baseados em aprendizado de máquina [Wang 2010, Varol et al. 2017, Kudugunta and Ferrara 2018]. Neste cenário, um *framework* que utiliza três redes diferentes para extrair características foi proposto e avaliado em larga escala [Varol et al. 2017], assim como o uso de uma rede neural profunda baseada em longa memória de curto prazo (LSTM) que explora o conteúdo e metadados dos tweets [Kudugunta and Ferrara 2018].

Uma outra direção é o estudo da influência dos *bots* em usuários comuns, e um recente trabalho buscou caracterizar e identificar usuários suscetíveis a *bots*, novamente em redes construídas a partir de metadados de usuários e tweets [Wagner et al. 2012].

Por fim, alguns dos trabalhos de caracterização e identificação de *bots* utilizam o conteúdo das mensagens, como os textos dos tweets e as *hastags*. Estas abordagens são geralmente frágeis pois o conteúdo e a grafia podem ser facilmente modificadas. Por outro lado, abordagens que utilizam a interação entre os usuários tendem a ser mais robustas, pois tais características são mais estáveis, visto que refletem a natureza da interação humana. Esta foi a abordagem adotada neste presente trabalho.

6. Conclusão

O presente trabalho buscou apresentar uma abordagem utilizando o paradigma de rede para a identificação de usuários *bots* em uma rede social. Foram debatidos alguns conceitos de Redes Complexas, destacando a relevância deste campo para o problema em estudo. Além disso, demonstrou como a estrutura da rede expõe significados sobre os participantes da mesma e como os relacionamentos podem evidenciar estes significados.

O objetivo de classificar os usuários como *bots* ou *não bots* se mostrou viável, mesmo para um modelo classificador relativamente simples e sem muitas otimizações. Apesar das limitações dos dados utilizados para este trabalho e, consequentemente, para a construção da rede modelada - onde nem todas as arestas estão presentes - a classificação de usuários, mesmo se tratando de um problema desbalanceado, atingiu taxas significativas de acerto para a classe de interesse e também da classe onde o usuário não é *bot*.

Os resultados obtidos para o classificador foram importantes para a forma como o problema se apresenta. Além disso, a métrica estudada para avaliar o classificador de forma mais designada para este problema, *AUC*, apresentou resultado extremamente satisfatório, $0,91$.

Existem outras métricas que poderiam ser extraídas da estrutura local do vértice. Algumas ideias, como calcular os vizinhos em comum ou grau dos vizinhos, poderiam trazer resultados interessantes que auxiliariam para a identificação de usuários *bot*.

Por fim, este projeto abrange a ideia do paradigma de rede e o campo de Ciência das Redes sendo utilizados como ferramentas para a classificação de usuários *bots* em redes sociais online. Os resultados obtidos deixam claro que as métricas estruturais que podem ser extraídas de uma rede conseguem exemplificar fenômenos e comportamentos dos pares de objetos codificados.

Referências

Barabási, A.-L. (2013). Network science. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 371(1987):20120375.

Baumgartner, J. (2018). Reddit comments dataset. https://files.pushshift.io/reddit/comments/. [Online; accessed 10-march-2019].

Damasceno, R. G. (2019). Classificação de perfis falsos em redes sociais online utilizando a estrutura da rede. Final course assignment, Universidade Federal do Rio de Janeiro, Rio de Janeiro, BR.

Ferrara, E., Varol, O., Davis, C., Menczer, F., and Flammini, A. (2016). The rise of social bots. *Communications of the ACM*, 59(7):96–104.

Gilani, Z., Farahbakhsh, R., Tyson, G., Wang, L., and Crowcroft, J. (2017). Of bots and humans (on twitter). In *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*, pages 349–354. ACM.

Kudugunta, S. and Ferrara, E. (2018). Deep neural networks for bot detection. *Information Sciences*, 467:312–322.

Newman, M. (2010). *Networks: an introduction*. Oxford university press.

Rosenberg, M., Confessore, N., and Cadwalladr, C. (2018). How Trump consultants exploited the facebook data of millions. *The New York Times*.

Varol, O., Ferrara, E., Davis, C., Menczer, F., and Flammini, A. (2017). Online human-bot interactions: Detection, estimation, and characterization.

Wagner, C., Mitter, S., Körner, C., and Strohmaier, M. (2012). When social bots attack: Modeling susceptibility of users in online social networks. In *# MSM*, pages 41–48.

Wang, A. H. (2010). Detecting spam bots in online social networking sites: a machine learning approach. In *IFIP Annual Conference on Data and Applications Security and Privacy*, pages 335–342. Springer.

Zaki, M. J., Meira Jr, W., and Meira, W. (2014). *Data mining and analysis: fundamental concepts and algorithms*. Cambridge University Press.