Utilizando Análise de Sentimentos para Definição da Homofilia Política dos Usuários do Twitter durante a Eleição Presidencial Americana de 2016

Josemar Alves Caetano, Hélder Seixas Lima, Mateus Freira dos Santos, Humberto Torres Marques-Neto

Programa de Pós-Graduação em Informática
Pontifícia Universidade Católica de Minas Gerais (PUC Minas)
Belo Horizonte – MG – Brasil – 31980-110

{josemar.caetano, helder.seixas, mateus.freira}@sga.pucminas.br

humberto@pucminas.br

Abstract. *This paper proposes a political homophily analysis among Twitter users during the 2016 US presidential election. We collected tweets, user profiles, and their network for this study over 122 days (08/01/2016 to 11/30/2016). Based on our dataset we considered the sentiment related to the candidates Donald Trump and Hillary Clinton for homophily analysis. The results showed that exist greater homophily among users that share negative sentiment associated with both candidates, higher in the case of candidate Donald Trump. The results also revealed that exists higher heterophily among users that did not express sentiment related to both candidates.*

Resumo. *Este trabalho propõe uma análise da homofilia política entre usuários do Twitter durante a eleição presidencial americana de 2016. Os tweets, perfis de usuários e suas redes de contatos foram coletados para este estudo ao longo de 122 dias (01/08/2016 até 30/11/2016). Considerando a base de dados coletada, o sentimento em relação aos candidatos Donald Trump e Hillary Clinton foi utilizado para a análise da homofilia. Os resultados mostraram que existe maior homofilia entre usuários que compartilham sentimento negativo em relação a ambos candidatos, sendo mais evidente no caso do candidato Donald Trump. Os resultados também mostraram que existe maior heterofilia entre usuários que não manifestaram sentimento em relação aos dois candidatos.*

1. Introdução

As redes sociais *online* são hoje um dos principais ambientes de debate, discussão e troca de informações entre as pessoas. A homofilia, tendência dos indivíduos possuírem características e comportamento similares aos dos seus pares, é um fenômeno já percebido nas redes sociais há algum tempo [Easley and Kleinberg 2010]. As características que os pares, por exemplo amigos, têm em comum vão desde questões não mutáveis como a etnia, até características mutáveis como crenças, profissão [McPherson et al. 2001] e sentimentos a respeito de um assunto [Yuan et al. 2014].

Percebe-se que a política é um assunto recorrente em debates nas redes sociais, o que pode transformar essas redes em um ambiente de forte confronto ideológico e de opiniões [Wang et al. 2012]. A eleição presidencial americana do ano de 2016 caracterizou-se por uma

disputa acirrada, especialmente após as primárias partidárias que resultaram na disputa entre Donald Trump, representante do Partido Republicano, e Hillary Clinton, representante do Partido Democrata[1]. Os embates políticos e ideológicos entre os dois candidatos refletiram nas discussões entre seus apoiadores nas redes sociais[2].

O objetivo deste trabalho é identificar e analisar a homofilia política entre os usuários do Twitter ao longo da campanha presidencial americana de 2016. O discurso político dos usuários foi caracterizado através da análise de sentimentos de *tweets* publicados ao longo do período de coleta. A ferramenta SentiStrength [Thelwall et al. 2010] foi utilizada para extrair o sentimento dos *tweets*. Dessa forma, foi possível analisar se um usuário tem um sentimento favorável (positivo), desfavorável (negativo) ou neutro em relação a cada candidato. Enfim, a principal contribuição deste trabalho é propor uma nova abordagem para a análise da homofilia política no Twitter utilizando a análise de sentimentos como característica para a classificação dos usuários.

Os *tweets*, perfis de usuário e sua rede de contatos foram coletados ao longo de 122 dias (01/08/2016 até 30/11/2016). A coleta teve como ponto de partida a identificação de *usuários sementes*, ou seja, pessoas que estavam comentando sobre a eleição presidencial americana no Twitter em tempo real. Ao todo foram coletados aproximadamente 3,6 milhões de *tweets* de 18.450 usuários distintos do Twitter.

Os resultados obtidos indicaram que a homofilia é um fenômeno presente na base de dados analisada e que ela é mais evidente quando se considera apenas as conexões recíprocas entre os usuários. A homofilia entre usuários que manifestaram sentimento negativo aos candidatos é alta, sendo mais considerável no caso do candidato Donald Trump. Também ocorreu homofilia entre usuários com sentimento positivo em relação aos dois candidatos, entretanto, em menor intensidade. Os resultados mostraram a existência do inverso da homofilia (heterofilia) entre os usuários que não manifestaram opinião em relação aos candidatos.

O trabalho está dividido da seguinte maneira. Na Seção 2 serão apresentados os conceitos básicos para compreensão deste trabalho. Os trabalhos que já propuseram estudar homofilia relacionada ao posicionamento político de usuários de redes sociais *online* são apresentados na Seção 3. A metodologia aplicada na análise de homofilia é explicada na Seção 4 e os resultados obtidos são apresentados e discutidos na Seção 5. Por fim, as conclusões e trabalhos futuros são apresentados na Seção 6.

2. Referencial Teórico

Nesta seção serão apresentados e discutidos os conceitos necessários para o entendimento da metodologia aplicada no presente trabalho. Na Seção 2.1, será apresentado o funcionamento da rede social Twitter e contextualizado seu uso na eleição presidencial americana de 2016. Em seguida, na Seção 2.2 será definido o fenômeno da homofilia e também como são calculados os indicadores necessários para avaliar sua ocorrência.

2.1. Twitter e a Eleição Presidencial Americana de 2016

O Twitter é uma rede social que possibilita a publicação de pequenas mensagens de até 140 caracteres que são chamadas de *tweets* [Giachanou and Crestani 2016]. Lançada em 2006,

[1]https://www.nytimes.com/2016/10/28/us/politics/donald-trump-voters.html
[2]http://www.reuters.com/article/us-usa-election-twitter-idUSKCN12R2OV

possui, em maio de 2017, cerca de 320 milhões de usuários ativos e é uma das redes sociais mais populares do mundo[3].

No Twitter, quando um usuário A cria um vínculo com um usuário B, diz-se que A está seguindo (*following*) B, ou que B tem A como seguidor (*follower*). Diferente de outras redes sociais, no Twitter as conexões entre usuários não são obrigatoriamente recíprocas, ou seja, mesmo que A siga B, isso não implica que B siga A.

O perfil de um usuário no Twitter é composto basicamente pelos seguintes atributos: nome, descrição do perfil, foto e sua localização. A *timeline* de um usuário é o conjunto de *tweets* que ele publicou. Os dois candidatos à presidência nos EUA possuem contas no Twitter e fizeram uso constante da rede durante a eleição de 2016. O candidato republicano Donald Trump é identificado no Twitter pelo nome de usuário @realDonaldTrump e em novembro de 2016 tinha cerca de 17,1 milhões de seguidores, enquanto que a democrata Hillary Clinton, identificada pelo usuário @HillaryClinton, era seguida por aproximadamente 11,6 milhões usuários em novembro de 2016.

Um *retweet* é um *tweet* publicado por um usuário A que foi compartilhado por um usuário B. Um dos recursos mais utilizados pelos usuários do Twitter são as *hashtags*, que consistem em expressões iniciadas pelo caractere "#" e tem a função de rotular ou resumir um tema em discussão [DeMasi et al. 2016]. Por exemplo, apoiadores de Donald Trump utilizaram as *hashtags* #VoteTrump e #TrumpWon, enquanto que apoiadores da Hillary Clinton utilizaram as *hashtags* #VoteHillary e #NeverTrump.

2.2. Homofilia

O princípio da homofilia é um tema muito estudado pela sociologia [McPherson et al. 2001]. Vários estudos observam o fenômeno da homofilia por características como etnia, idade, gênero, lugares onde viveu, profissão, entre outros [McPherson et al. 2001, Currarini et al. 2009, Easley and Kleinberg 2010]. [Ribeiro and Bastos 2014] propôs um estudo sobre a homofilia entre alunos cotistas e alunos não cotistas de uma Universidade Federal brasileira.

De acordo com [Colleoni et al. 2014] o indicador de homofilia de indivíduos do tipo *i* é dado por:

$$H_{\mathrm{i}} = \frac{s_{\mathrm{i}}}{s_{\mathrm{i}} + d_{\mathrm{i}}} \tag{1}$$

Onde H_{i} é o indicador de homofilia, s_{i} representa o número conexões que ligam indivíduos do tipo i (conexões homogêneas) e d_{i} representa o número conexões que ligam indivíduos do tipo i com indivíduos de outros tipos (conexões heterogêneas).

[Currarini et al. 2009] utilizam a Equação 1 para o cálculo da homofilia. Porém, eles ressaltam a dificuldade de medir a homofilia simplesmente considerando H_{i}. Os autores apresentam o seguinte exemplo: seja um grupo A que corresponde a 95% de uma população de uma rede e um grupo B que corresponde aos 5% restantes. Considere que cada grupo tenha um percentual de 96% de amizades homogêneas (H_{i} = 0,96). Comparando o grupo A com o grupo B, embora ambos tenham o mesmo H_{i}, a homofilia entre os membros do grupo B é maior que a do grupo A, pois a probabilidade de amizades entre membros do grupo B é menor que do a probabilidade de A.

[3]http://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users

Dessa forma, [Currarini et al. 2009] recomenda calcular o indicador de *inbreeding homophily* que é dado por:

$$IH_{\mathrm{i}} = \frac{H_{\mathrm{i}} - w_{\mathrm{i}}}{1 - w_{\mathrm{i}}} \tag{2}$$

Onde IH_{i} é o indicador de *inbreeding homophily*, H_{i} é o indicador de homofilia definido na Equação 1 e w_{i} é probabilidade da ocorrência de indivíduos do tipo i, que consiste no total de indivíduos do tipo i dividido pelo total de indivíduos de uma rede T. Caso H_{i} seja maior que w_{i}, então há homofilia.

Retornando ao exemplo anterior, o valor de IH_{i} para os grupos A e B, são 0,2 e 0,96, respectivamente. Esse resultado demonstra que a métrica *inbreeding homophily* pode ser utilizada para comparar a homofilia relativa entre diferentes populações. Sendo que, quanto maior o valor de IH_{i}, mais forte é a ocorrência da homofilia.

Ao se analisar o IH_{i}, o valor zero corresponde a linha de base determinante da homofilia. O inverso da homofilia é chamado de heterofilia, ou seja, quando há predominância de relacionamento entre indivíduos de tipos diferentes. Portanto, neste trabalho a ocorrência de homofilia ou heterofilia é dada pela seguinte condição:

$$\begin{cases} IH_{\mathrm{i}} > 0 & homofilia \\ IH_{\mathrm{i}} < 0 & heterofilia \end{cases}$$

3. Trabalhos Relacionados

Alguns trabalhos já propuseram análises da homofilia política no Twitter. [Colleoni et al. 2014] investigaram a homofilia política em uma base de dados de eleitores americanos democratas e republicanos. Utilizou-se aprendizado de máquina e análise da rede social para fazer a classificação da preferência partidária dos usuários. Nesse trabalho foi percebido que em geral os democratas apresentam maior nível de homofilia política quando comparado com os republicanos. Também foi constatado que os níveis de homofilia são maiores quando se considera conexões recíprocas.

[Halberstam and Knight 2016] investigaram como a homofilia política interfere na difusão de informações nas redes sociais. Esse estudo foi desenvolvido utilizando uma base de dados de usuários do Twitter politicamente engajados. Os autores identificaram que usuários ligados a grupos políticos majoritários, quando comparados com usuários de grupos minoritários, têm mais conexões, estão expostos a um maior número de informações e estas informações são recebidas mais rapidamente.

Esses trabalhos demonstraram que a homofilia política é um fenômeno presente no Twitter. Em geral, esses trabalhos consideraram métodos de classificação para caracterizar os usuários como adeptos do Partido Democrata ou Partido Republicano, ou classificaram os usuários de acordo com a orientação política conservadora ou liberal. Diferentemente, neste trabalho, os usuários foram classificados conforme o sentimento que eles expressaram nos *tweets* que eles publicaram a respeito dos candidatos. Portanto, a aplicação da análise de sentimentos é o principal diferencial deste trabalho em relação aos trabalhos anteriores que investigaram o tema da homofilia política no Twitter.

4. Metodologia

Nesta seção, as etapas da metodologia aplicada no presente trabalho serão apresentadas. Primeiramente, a coleta dos dados do Twitter será descrita. Em seguida, o cálculo do sentimento dos *tweets* e os critérios de identificação dos sujeitos de cada *tweet* serão explicados. Por fim, o cálculo da homofilia entre os usuários analisados será apresentado.

4.1. Coleta dos Dados

Os dados considerados neste trabalho foram coletados na rede social Twitter, de 01/08/16 e encerrando em 30/11/16. Ressalta-se que a eleição presidencial americana ocorreu no dia 08/11/16 e que os debates televisionados entre Donald Trump e Hillary Clinton ocorreram nas datas 26/09/16, 09/10/16 e 19/10/16. A API oficial do Twitter[4] foi utilizada para coletar os *tweets*, perfis de usuário e suas redes de contatos. A API disponibiliza, no máximo, os 200 últimos *tweets* publicados por um usuário.

A coleta teve como ponto de partida a identificação de *usuários sementes*, ou seja, pessoas que estavam comentando sobre a eleição americana no Twitter. Essa identificação foi obtida através do método de *streaming* da API que possibilita a coleta em tempo real de *tweets* e *retweets*.

O principal objetivo da coleta em tempo real foi identificar os usuários que escreveram *tweets* em inglês e que fizeram um *retweet* de algum *tweet* de um dos candidatos. Esse objetivo se baseou na hipótese de que se um usuário faz um *retweet* de um *tweet* de um candidato, então ele leu aquele *tweet* e fez uma citação do que o candidato disse. Desta forma, esse usuário provavelmente está utilizando o Twitter para discutir, ou promover discussão sobre política.

Para cada *usuário semente* foi coletado seu perfil, seus *tweets* e sua rede de contatos (*followers* e *friends*). Também foram coletados os perfis, *tweets* e conexões de cada membro da rede de contatos vinculada a um *usuário semente*. Ao todo foram coletados cerca de 3,6 milhões de *tweets* de 18.450 usuários, sendo 185 *usuários sementes*.

4.2. Análise do Sentimento dos *Tweets*

O discurso político de um usuário foi caracterizado a partir do sentimento expresso em relação aos dois candidatos em seus *tweets* coletados. A ferramenta para a análise de sentimentos utilizada nesse trabalho foi a SentiStrength [Thelwall et al. 2010]. Essa ferramenta retorna três valores de sentimento associado com cada frase do texto: positivo, negativo e *scale* (diferença entre os valores positivo e negativo). O SentiStrength utiliza a abordagem de análise de sentimentos baseada em lexemas de dicionários. Seu dicionário é composto por 700 lexemas e são utilizas listas de *emoticons* e *boosting words* (*very, most, worst, best*, etc) para melhorar o desempenho da análise de sentimentos [Giachanou and Crestani 2016]. Contudo, uma das limitações do SentiStrength é não reconhecer sarcasmo [Ferrara and Yang 2015]. Neste trabalho utilizou-se o dicionário padrão do SentiStrength sem modificações.

Durante a análise de sentimentos observou-se que alguns *tweets* continham menções aos dois candidatos ao mesmo tempo. Dessa forma, o texto poderia conter um sentimento muito positivo em relação a um candidato e um sentimento muito negativo em relação ao outro candidato. Essa situação poderia representar um problema para a análise dos sentimentos em relação à cada candidato, pois, os valores dos lexemas se neutralizavam e um *tweet* ficava com

[4]https://dev.twitter.com/overview/api

sentimento agregado de valor 0. Para resolver esse problema, foi utilizado o Stanford *Parser*[5], que é uma biblioteca para análise de linguagem natural desenvolvida pelo grupo de estudos sobre processamento de linguagem natural da Universidade de Stanford [Klein et al. 2003]. Essa ferramenta permitiu identificar os sujeitos do *tweet* e os lexemas associados a eles para fazer o cálculo do sentimento associado a cada sujeito do texto.

4.2.1. Cálculo da Média do Sentimento Manifestado pelo Usuário

A média do sentimento em relação aos candidatos foi calculada para cada um dos 18.450 usuários. O cálculo consistiu na seleção de cada frase dos *tweets* que o usuário havia escrito e que possuíam sujeitos referentes a um dos candidatos. Os candidatos foram identificados por nomes e pronomes associados a eles conforme apresentado na Tabela 1, pois observou-se que existiam vários termos utilizados para se referir aos candidatos além do nome de usuário no Twitter.

Os pronomes somente foram considerados quando no *tweet* em análise também era feita menção a um dos nomes dos candidatos. Por exemplo, se o *tweet* possuía o nome "@realDonaldTrump" e o pronome "He", então considerou-se que "He" estava associado com "@realDonalTrump".

Tabela 1. Nomes e pronomes considerados como menções aos candidatos

Candidato	*Nomes*	*Pronomes*
Donald Trump	@realDonaldTrump, Trump, Trumps, Trump's, DT	He, he, him
Hillary Clinton	@HillaryClinton, Hillary, Hillarys, Hillary's, Clinton's, HC	She, she, her

Fonte: Elaborado pelos autores

O processo de calcular a média do sentimento dos usuários em relação aos candidatos resultou em dois *scales*, sendo um referente ao Donald Trump e o outro à Hillary Clinton. Quanto maior o valor, mais positivo é o sentimento e quanto menor o valor mais negativo é o sentimento. A Tabela 2 apresenta a média e o desvio padrão do sentimento dos 18.450 usuários em relação aos candidatos Donald Trump e Hillary Clinton.

Tabela 2. Média e Desvio Padrão dos Sentimentos

	Donald Trump	*Hillary Clinton*
Média	-0,130	-0,055
Desvio Padrão	0,296	0,334

Fonte: Elaborado pelos autores

Para simplificar a análise dos resultados deste trabalho decidiu-se discretizar os valores numéricos dos sentimentos em relação aos candidatos em valores categóricos (Tabela 3). Usuários com média de sentimento maior que 1 ou com média de sentimento menor que 1 representaram apenas 0,87% dos usuários analisados.

[5]http://nlp.stanford.edu/software/lex-parser.shtml

Neste trabalho assume-se que um usuário tem um sentimento positivo em relação a um candidato quando a média do sentimento for maior que 0 e que um usuário tem sentimento negativo em relação a um candidato quando a média for menor que 0. Quando a média do sentimento tiver valor 0, considera-se que há sentimento neutro em relação ao candidato. Se um usuário não tiver *tweets* mencionando um dos candidatos (valor nulo), então diz-se que o sentimento não foi manifestado.

Tabela 3. Discretização dos valores dos atributos de média do sentimento

Valor da média do sentimento	*Atributo categórico*
Valor nulo	Sentimento não manifestado
Valor menor que 0	Sentimento negativo
Valor igual a 0	Sentimento neutro
Valor maior que 0	Sentimento positivo

Fonte: Elaborado pelos autores

A Tabela 4 apresenta a distribuição de usuários nas categorias definidas na discretização (Tabela 3). A categoria "Sentimento não manifestado" apresentou 55,50% e 65,30% para Donald Trump e Hillary Clinton, respectivamente. Isso mostra que mais da metade dos usuários não publicaram *tweets* em relação aos candidatos.

Usuários com sentimento negativo ao Donald Trump representaram 27,42% dos usuários analisados, mais que o dobro dos usuários com sentimento negativo em relação à Hillary Clinton (12,36%). A categoria "Sentimento neutro" compreendeu 12,65% e 15,54% dos usuários com sentimento em relação ao Donald Trump e Hillary Clinton, respectivamente. A categoria "Sentimento positivo" possui a menor quantidade de usuários (4,43% e 6,80% para Donald Trump e Hillary Clinton, respectivamente).

Tabela 4. Sentimento dos usuários em relação aos candidatos

	Donald Trump	*Hillary Clinton*
Sentimento não manifestado	10.240 (55,50%)	12.047 (65,30%)
Sentimento negativo	5.059 (27,42%)	2.281 (12,36%)
Sentimento neutro	2.334 (12,65%)	2.868 (15,54%)
Sentimento positivo	817 (4,43%)	1.254 (6,80%)

Fonte: Elaborado pelos autores

Observa-se que as categorias de sentimento em relação aos candidatos não são equilibradas. Por exemplo, existem mais usuários com sentimento que favorece a Hillary Clinton do que o Donald Trump. Contudo, isso não implica em um problema para a realização deste trabalho, pois, a *inbreeding homophily* (Equação 2) leva em consideração a proporção de cada tipo de usuário conforme apresentado na Seção 2.2.

4.3. Cálculo da Homofilia

Para cada valor dos atributos da média do sentimento em relação aos candidatos foi calculado o H_{i} (Equação 1). A partir de H_{i}, o indicador IH_{i} (Equação 2) também foi calculado. Esse indicador é utilizado para comparar o resultado da homofilia entre diferentes atributos. Por

exemplo, comparar se a homofilia entre os usuários com sentimento positivo em relação à Hillary Clinton é maior quando comparado com os usuários que têm sentimento positivo em relação ao Donald Trump.

Neste trabalho, dois cenários foram considerados para a análise da homofilia dos usuários. O primeiro cenário é uma análise a partir de quatro classes de sentimento e o segundo cenário é uma análise que considera seis classes de sentimento com o intuito de representar melhor o do domínio do problema.

As quatro classes do primeiro cenário foram escolhidas de acordo com as categorias definidas na discretização. As seis classes do segundo cenário (Tabela 5) foram definidas a partir da interseção dos usuários pertencentes à cada classe de sentimento em relação ao candidato do primeiro cenário .

Ressalta-se que no primeiro cenário, cada usuário pertence a duas classes de sentimento: uma em relação ao Donald Trump e a outra em relação à Hillary Clinton. No segundo cenário, cada usuário pertence a somente uma classe de sentimento. A Tabela 5 apresenta o nome, descrição e total de usuários de cada uma das seis classes definidas para o segundo cenário de análise.

Tabela 5. Definição das Seis Classes

Classe	**Descrição**	**Total**
Não manifestado	Sentimento não manifestado ao Donald Trump e Hillary Clinton	9.811 (53,18%)
Neutro	Sentimento neutro em relação ao Donald Trump e Hillary Clinton	2.278 (12,35%)
Crítico ambos	Sentimento negativo em relação ao Donald Trump e Hillary Clinton	1.750 (9,49%)
Pró Trump	Sentimento positivo em relação ao Donald Trump e sentimento diferente de positivo em relação à Hillary Clinton; ou sentimento negativo em relação a Hillary Clinton e diferente de negativo em relação ao Donald Trump	1.001 (5,43%)
Pró Hillary	Sentimento positivo em relação à Hillary Clinton e sentimento diferente de positivo em relação ao Donald Trump; ou sentimento negativo em relação ao Donald Trump e diferente de negativo em relação à Hillary Clinton	3.509 (19,00%)
Favorável ambos	Sentimento positivo em relação ao Donald Trump e Hillary Clinton	101 (0,55%)

Fonte: Elaborado pelos autores

A classe "Não manifestado" contém 53,18% dos usuários. Isso mostra que mais da metade dos usuários não publicaram nenhum *tweet* em relação a ambos candidatos e fazem parte da rede de contatos dos *usuários sementes* (conforme definido na Seção 4.1). A classe "Neutro" contém 12,35% do total de usuários. Essa classe corresponde aos usuários que tiveram sentimento neutro em relação a ambos candidatos.

A classe "Crítico ambos" representa os usuários com sentimento negativo em relação a ambos candidatos e contém 9,49% do total de usuários. A classe "Pró Trump" consistiu de

usuários com sentimento que favorece o candidato Donald Trump e desfavorece a candidata Hillary Clinton. Essa classe contém 5,43% do total de usuários. A classe "Pró Hillary" contém 19,00% dos usuários. Essa classe contém usuários que têm sentimento que favorece a candidata Hillary Clinton e desfavorece o candidato Donald Trump. A classe "Favorável ambos" consistiu de usuários com sentimento positivo em relação a ambos candidatos e compreendeu 0,55% dos usuários.

Para cada cenário analisado, dois valores de IH_i foram calculados. No primeiro cálculo considerou-se todos os tipos de conexões entre os usuários (conexões gerais). No segundo cálculo considerou-se somente as conexões recíprocas entre os usuários conforme definido na Seção 2.1. A escolha de utilizar os dois valores de IH_i foi motivada pelo resultado obtido por [Colleoni et al. 2014], onde percebeu-se que a homofilia entre usuários com conexões recíprocas é maior quando comparada com a homofilia de conexões gerais.

5. Resultados e Discussão

Nesta seção os resultados dos cálculos da homofilia obtidos através da metodologia explicada na Seção anterior serão apresentados e discutidos. As subseções seguintes apresentam discussões dos resultados nos dois cenários considerados neste trabalho e as possíveis limitações da metodologia aplicada.

5.1. Análise da Homofilia do Primeiro Cenário

Neste cenário a homofilia considerando as quatro classes de sentimento é analisada. Conforme apresentado na Tabela 6, nota-se que houve homofilia ente os usuários da classe de sentimento "Negativo" e "Positivo" (para ambos candidatos) e que houve heterofilia nas classes "Não manifestado" e "Neutro" (para ambos candidatos). Nota-se também que a intensidade da homofilia nas classes de sentimento "Negativo" e sentimento "Positivo" em relação aos candidatos aumenta quando se analisa apenas as suas conexões recíprocas.

Tabela 6. Resultado da homofilia nas quatro classes de sentimento

Candidato	*Classe*	*IH_i das conexões gerais*	*IH_i das conexões recíprocas*
Trump	Não manifestado	-1,160	-1,144
	Negativo	*0,377*	*0,518*
	Neutro	-0,109	-0,124
	Positivo	0,039	0,039
Clinton	Não manifestado	-1,721	-1,768
	Negativo	*0,116*	*0,144*
	Neutro	-0,056	-0,057
	Positivo	*0,039*	*0,052*

Fonte: Elaborado pelos autores

O maior grau de homofilia ocorreu entre usuários que publicaram *tweets* com sentimento negativo em relação ao Donald Trump (possíveis apoiadores da Hillary). Entre usuários que publicaram *tweets* com sentimento negativo em relação à Hillary Clinton (possíveis apoiadores do Trump) ocorreu o segundo maior grau de homofilia. Isso indica que a rede de usuários com sentimento negativo em relação a um candidato é formada por usuários mais conectados entre si.

Também houve homofilia na classe de sentimento "Positivo" em relação a ambos candidatos. Entretanto, com menor expressividade em relação à classe "Negativo". Nas demais classes ("Não manisfestado" e "Neutro"), houve a ocorrência de heterofilia. Ressalta-se que houve maior grau de heterofilia na classe "Não manifestado". Isso demonstra que usuários com sentimento não manifestado são mais conectados com usuários que manifestaram algum sentimento em relação a um candidato. Ou seja, usuários pertencentes à classe "Não manifestado" possuem uma rede de contatos predominantemente composta por usuários de outras classes.

5.2. Análise da Homofilia do Segundo Cenário

A Tabela 7 apresenta os resultados da análise da homofilia considerando as seis classes de sentimento. Houve homofilia nas classes "Crítico ambos", "Pró Trump" e "Pró Hillary". O fenômeno da heterofilia ocorreu nas classes "Não manifestado", "Neutro" e "Favorável ambos".

Tabela 7. Resultado da homofilia nas seis classes de sentimento

Classe	*IH_i das conexões gerais*	*IH_i das conexões recíprocas*
Não manifestado	-1,056	-1,030
Neutro	-0,116	-0,138
Crítico ambos	*0,090*	*0,124*
Pró Trump	*0,053*	*0,057*
Pró Hillary	*0,075*	*0,106*
Favorável ambos	-0,002	-0,002

Fonte: Elaborado pelos autores

Ressalta-se que o maior grau de homofilia ocorreu na classe "Crítico ambos" e o maior grau de heterofilia ocorreu na classe "Não manifestado". Isso indica que usuários com sentimento negativo em relação ao Donald Trump e à Hillary Clinton são mais conectados entre si e, como no primeiro cenário, usuários com sentimento não manifestado possuem uma rede de contatos formada predominantemente por usuários com sentimentos diferentes em relação aos candidatos.

Nas classes em que ocorreram homofilia, quando são comparados os valores de IH_{i} das conexões gerais e das conexões recíprocas, nota-se que o maior grau de homofilia acontece quando são analisadas apenas conexões recíprocas dos usuários. Na classe "Não manifestado", a heterofilia possui um grau maior das conexões gerais. Na classe "Neutro", houve maior grau de heterofilia das conexões recíprocas e na classe "Favorável ambos" os dois valores de IH_{i} foram os mesmos.

5.3. Limitações

Uma das limitações deste trabalho é o risco da base dados utilizada ter muitos *tweets* que contém conteúdo político sarcástico, pois o SentiStrength não reconhece sarcasmo. Dessa forma, muitos usuários que foram classificados com sentimento positivo poderiam ser, na verdade, classificados como usuários com sentimento negativo.

Outra limitação é que a API do Twitter fornece no máximo, os 200 últimos *tweets* publicados. Dessa forma, algum usuário pode ter publicado *tweets* de sentimento positivo ou negativo sobre um candidato e estes podem não ter sido coletados. Portanto, esse usuário não seria classificado como Não manifestado ou como Neutro a um candidato.

6. Conclusão

Neste trabalho foi realizado um estudo sobre a ocorrência do fenômeno da homofilia política considerando o sentimento manifestado por usuários do Twitter em relação aos candidatos Donald Trump e Hillary Clinton na eleição americana de 2016. Os *tweets*, perfis de usuário e redes de contatos foram coletados ao longo de 122 dias (01/08/2016 até 30/11/2016). O discurso político de um usuário foi caracterizado a partir do sentimento expresso em relação aos dois candidatos a partir de seus *tweets* coletados.

A análise da homofilia foi realizada em dois cenários. O primeiro cenário considerou as classes de sentimento "Negativo", "Neutro" e "Positivo" de forma separada para cada candidato. O segundo cenário considerou a classificação do discurso do usuário como "Neutro", "Crítico ambos", "Pró Trump", "Pró Hillary" e "Favorável ambos". Nos dois cenários também foi considerado aqueles usuários que não manifestaram sentimento em relação aos candidatos.

No primeiro cenário, o maior grau de homofilia ocorreu entre usuários que publicaram *tweets* com sentimento negativo em relação ao Donald Trump (possíveis apoiadores da Hillary). Entre usuários que publicaram *tweets* com sentimento negativo em relação à Hillary Clinton (possíveis apoiadores do Trump) ocorreu o segundo maior grau de homofilia. Nas demais classes ("Não manisfestado" e "Neutro"), houve a ocorrência de heterofilia.

No segundo cenário houve homofilia entre usuários que são críticos dos dois candidato, apoiadores do Donald Trump e apoiadores da Hillary Clinton. O fenômeno da heterofilia ocorreu entre usuários que não manifestaram seu sentimento, que tiveram sentimento neutro e sentimento favorável a ambos candidatos. Ressalta-se que houve maior grau de heterofilia entre usuários com sentimento não manifestado nos dois cenários analisados.

Uma vez que mais da metade dos usuários nos dois cenários analisados foram classificados com sentimento "Não manifestado", sugere-se, como trabalho futuro, que sejam acrescidas novas características para classificação do discurso político dos usuários no Twitter. Por exemplo, características sintáticas do *tweet* como *hashtags* e menções além de características da rede de contatos do usuário como por exemplo se um usuário segue ou é seguido por um candidato.

7. Agradecimentos

Este trabalho foi financiado pela FAPEMIG, CAPES, CNPq, e FAPEMIG-PRONEX-MASWeb – Modelos, Algoritmos e Sistemas para a Web (processo APQ-01400-14).

Referências

Colleoni, E., Rozza, A., and Arvidsson, A. (2014). Echo chamber or public sphere? predicting political orientation and measuring political homophily in twitter using big data. *Journal of Communication*, 64(2):317–332.

Currarini, S., Jackson, M. O., and Pin, P. (2009). An economic model of friendship: Homophily, minorities, and segregation. *Econometrica*, 77(4):1003–1045.

DeMasi, O., Mason, D., and Ma, J. (2016). Understanding communities via hashtag engagement: A clustering based approach. In *Tenth International AAAI Conference on Web and Social Media*.

Easley, D. and Kleinberg, J. (2010). *Networks, crowds, and markets: Reasoning about a highly connected world*. Cambridge University Press.

Ferrara, E. and Yang, Z. (2015). Measuring emotional contagion in social media. *PLOS ONE*, 10(11):1–14.

Giachanou, A. and Crestani, F. (2016). Like it or not: A survey of twitter sentiment analysis methods. *ACM Comput. Surv.*, 49(2):28:1–28:41.

Halberstam, Y. and Knight, B. (2016). Homophily, group size, and the diffusion of political information in social networks: Evidence from twitter. *Journal of Public Economics*, 143:73–88.

Klein, D., Manning, C. D., et al. (2003). Fast exact inference with a factored model for natural language parsing. *Advances in neural information processing systems*, pages 3–10.

McPherson, M., Smith-Lovin, L., and Cook, J. M. (2001). Birds of a feather: Homophily in social networks. *Annual review of sociology*, pages 415–444.

Ribeiro, E. M. B. and Bastos, A. V. B. (2014). A homofilia por cotas em cursos de alta e baixa concorrência na universidade federal da bahia. In *XXIV Congresso da Sociedade Brasileira de Computação*. III Brazilian Workshop on Social Network Analysis and Mining.

Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., and Kappas, A. (2010). Sentiment in short strength detection informal text. *J. Am. Soc. Inf. Sci. Technol.*, 61(12):2544–2558.

Wang, H., Can, D., Kazemzadeh, A., Bar, F., and Narayanan, S. (2012). A system for real-time twitter sentiment analysis of 2012 u.s. presidential election cycle. In *Proceedings of the ACL 2012 System Demonstrations*, ACL '12, pages 115–120, Stroudsburg, PA, USA. Association for Computational Linguistics.

Yuan, G., Murukannaiah, P. K., Zhang, Z., and Singh, M. P. (2014). Exploiting sentiment homophily for link prediction. In *Proceedings of the 8th ACM Conference on Recommender Systems*, RecSys '14, pages 17–24, New York, NY, USA. ACM.