"Bela, Recatada e do Lar": Base de Dados e Aspectos do Movimento Social Ocorrido na Rede Social Online Twitter

Marcela Mayumi Mauricio Yagui, Luís Fernando Monsores Passos Maia, Wallace Ugulino, Adriana Vivacqua, Jonice Oliveira

Programa de Pós-Graduação em Informática da Universidade Federal do Rio de Janeiro (PPGI/UFRJ) – RJ – Brasil

{marcelayagui,luisfmpm}@ufrj.br, ugulino@ppgi.ufrj.br, {avivacqua, jonice} @dcc.ufrj.br

Abstract. *Between April and May 2016 a virtual mass protest was observed in response to the publication of an article in a Brazil-wide distributed magazine. The present research consisted in the construction of a dataset on the messages published in the Online Social Network Twitter aiming at supporting researchers in the field of Opinion Mining by providing of a corpus of messages in Portuguese language. In addition to the dataset, which was characterized by techniques of Social Network Analysis, Sentiment Analysis techniques were applied with support from the Naive Bayes probabilistic algorithm to check the tweets' polarity.*

Resumo. *Entre Abril e Maio de 2016 um protesto virtual em massa foi observado como resposta à publicação de uma matéria em uma revista de circulação nacional. A presente pesquisa consistiu na construção de uma base de dados relativa às mensagens publicadas na Rede Social Online Twitter com o objetivo de apoiar pesquisadores da área de Mineração de Opiniões por meio da oferta de um corpus de mensagens em língua portuguesa. Além da base de dados, que foi caracterizada por meio de técnicas de Análise de Redes Sociais, foi aplicada a técnica de Análise de Sentimentos com o auxílio do algoritmo probabilístico Naïve Bayes para verificar a polaridade dos tweets.*

1. Introdução

As Redes Sociais Online (RSO) foram usadas de maneira marcante nos últimos eventos políticos do mundo. No Brasil, os cidadãos têm manifestado suas posições na internet por meio das RSO e há uma crescente polarização entre ideologias que está registrada nessas redes de forma não-estruturada: usuários se expressam em linguagem natural escrita, em fotos com mensagens, em charges de humor, entre outras formas de manifestação social online [Carvalho et al. 2016].

As interações sociais registradas nas RSO representam uma oportunidade de compreender melhor os movimentos sociais, suas ações e reações. As aplicações desse conhecimento são diversas e serviram de motivação para o surgimento de novas áreas de pesquisa, como a área da Mineração de Opiniões e da Análise de Sentimentos. O registro dessas interações, entretanto, é influenciado pela cultura do grupo social, sua língua, expressões linguísticas locais, entre outras características únicas de cada região que tornam o trabalho de mineração ainda mais desafiante. Uma revisão da literatura da área e uma discussão sobre trabalhos relacionados é apresentada na Seção 2.

Com o objetivo de gerar conhecimento sobre técnicas de mineração úteis para o contexto brasileiro, no presente trabalho foram coletadas mensagens postadas na RSO Twitter®, no período de 22/Abril até 03/Maio de 2016. Foram selecionadas as mensagens marcadas com a etiqueta (*hashtag*) #BelaRecatadaEDoLar ou que continham a mesma *string*. A etiqueta foi usada por muitos usuários do Twitter para posicionar-se em relação à uma matéria publicada em revista de circulação nacional.

Nesse trabalho buscou-se contribuir com a geração de uma base de dados comum para comparação de modelos por pesquisadores que atuem em mineração de opiniões a partir do Twitter. Na Seção 3 discute-se o procedimento adotado nessa pesquisa. Na Subseção 3.1 é apresentado o trabalho de análise de sentimentos realizado com o algoritmo Naïve Bayes. Obteve-se uma acurácia de 81% com o uso do '*bag of words*' como elemento de entrada. As subseções 3.2 (análise da rede de *retweets*) e 3.3 (nuvem de palavras) servem para caracterizar a base de dados. Uma discussão sobre esses dados é apresentada na subseção 3.4. Conclusão e Trabalhos Futuros são discutidos na Seção 4.

2. Revisão da Literatura

A mobilização das multidões virtuais em RSO é um fenômeno que vêm sendo estudado com o objetivo de prever comportamentos, eventos e acontecimentos [Bonabeau 2004]. Esse tipo de conhecimento tem potencial de aplicação em várias áreas, como política, comunicação, marketing científico e até marketing pessoal [Li and Li 2011] - um bom exemplo é a emergente fama dos sociólogos e filósofos regionais que veio a reboque da exposição online dos mesmos nessas redes sociais. Numa investigação desse tipo, é fundamental que as características locais sejam levadas em consideração: acontecimentos e notícias que afetam o grupo em estudo, a linguagem usada pelo grupo, o uso de figuras emblemáticas e bordões. Muitas dessas características podem ser difíceis de serem medidas de forma objetiva [Carvalho et al. 2016].

A Mineração de Opiniões é uma área de pesquisa dedicada à geração de conhecimento sobre como extrair, analisar e comparar dados objetivos a partir dos dados (geralmente) semisestruturados, registrados nesses sistemas. As técnicas de mineração de dados e aprendizagem de máquina são empregadas largamente com o objetivo de perceber padrões latentes nessa interação [Elmasri and Navathe 2005].

O conhecimento do domínio de aplicação, no entanto, é um fator importante na hora de projetar esses modelos de predição baseados em dados semiestruturados. Dessa forma, a pesquisa em mineração de opiniões distingue-se das demais pesquisas em mineração de dados especialmente pelas características próprias das redes e da aplicação do conhecimento dessas características na hora de montar e aprimorar o modelo preditivo [Marteleto 2001].

Outra vertente que surgiu da mineração de opiniões recentemente é a análise de sentimentos [Sarlan et al. 2014], que refere-se à ampla área da mineração de texto, processamento de linguagem natural e linguística computacional que envolve o estudo computacional e semântico dos sentimentos, opiniões e emoções expressados em textos. Este tipo de mineração permite detectar se uma opinião expressa em um texto é positiva, negativa ou neutra. Ela também permite detectar emoções características como felicidade, tristeza ou raiva. A análise de sentimentos vem se tornando rapidamente uma das áreas do processamento de linguagem natural mais investigada nos últimos anos.

O Twitter se tornou bastante popular na realização deste tipo de análise por seu formato único e característico de *microblog*, que permite o envio e compartilhamento de mensagens de texto, imagens e vídeos rapidamente e através de uma rede de milhões de usuários, mas excepcionalmente, por permitir mensagens de texto de, no máximo, 140 caracteres. Isso é determinante no modo como as pessoas se expressam na rede e facilita o trabalho de mineração e reconhecimento de padrões [Carvalho et al. 2016].

Neste sentido, diversos estudos vêm sendo realizados com foco na análise de sentimentos para classificar emoções na plataforma Twitter. Um dos trabalhos pioneiros foi o de Davidov *et al*. (2010), onde uma análise de sentimentos foi realizada por meio da classificação de *tweets* com base em *hashtags* para separar as categorias desejadas, e nos *emoticons* contidos nas mensagens de texto, que auxiliaram no treinamento do algoritmo classificador, entre outros fatores que foram usados na extração de sentimentos, como a pontuação e *strings* que formavam palavras-chave.

A classificação de *tweets* em polaridades também já foi previamente estudada por Pak & Paroubek (2010) que aplicou o algoritmo de classificação probabilístico Naïve Bayes para realizar o procedimento. No estudo de Ferreira (2012), o modelo preditivo montado com o classificador Naïve Bayes obteve o melhor desempenho na tarefa de análise de textos em português. Os resultados obtidos com o uso de Naïve Bayes na área são encorajadores, especialmente em função da simplicidade do modelo.

Outro estudo importante de ser mencionado é o de Nascimento *et al.* (2012), no qual é possível observar a polaridade de sentimentos da população em relação às notícias divulgadas pela mídia. Foram selecionados previamente três tópicos em português para a coleta de dados; em seguida, os autores realizaram a rotulação manual do conteúdo para analisá-lo através do emprego de três métodos, um deles, sendo o Naïve Bayes. Novamente, Naïve Bayes teve melhores resultados que os demais modelos avaliados.

Outra forma de estudar as RSO diz respeito aos recursos de análise que mapeiam as relações sociais em estruturas de grafo. Por meio deste mapeamento, é possível representar nós e suas conexões que, por intermédio da aplicação de técnicas e algoritmos já consolidados na área de teoria dos grafos, dão indícios do comportamento, da influência e da propagação de informações dentro de uma comunidade específica. Este tipo de mapeamento é conhecido como Análise de Redes Sociais (ARS) e também é bastante difundido em estudos envolvendo a RSO Twitter [Marteleto 2001].

Um exemplo recente do emprego de ARS no Twitter é encontrado no trabalho de Willis *et al*. (2015), no qual foram coletados *tweets* relacionados aos Jogos Olímpicos de Londres de 2012. Por meio do uso de técnicas de ARS e de correlações estatísticas, foram calculadas métricas como o *Betweeness* e *Pagerank* para identificar o comportamento de clusters sociais formados durante o evento. Enquanto perfis corporativos moldaram a rede e afetaram a sua conectividade, perfis pessoais aumentaram as interações e discussões na rede, o que mostrou-se importante para a identificação de usuários chave e seu engajamento com o público dos Jogos Olímpicos.

A criação de redes de *retweets* foi também explorada no trabalho de Weitzel & Oliveira (2012). Com base em *tweets* do domínio da saúde, os autores analisaram a rede sob a perspectiva da estrutura topológica e de redes ego dos usuários, como também investigaram o mecanismo de classificação dos nós baseados nos pesos dos *retweets*. Por meio de métricas de centralidade, eles concluíram que medidas como o *Pagerank*

indicam a popularidade de um usuário e medidas como o *Betweeness* indicam a posição chave de um nó dentro da rede.

Na presente revisão, verificou-se o recorrente uso de algumas técnicas de ARS para identificação de perfis influentes na RSO Twitter por meio de análises de redes de *retweets*. Há destaque especialmente para investigação da correlação entre as métricas *Betweeness* e *Pagerank*. Já no campo da análise de sentimentos, os resultados com Naïve Bayes têm sido particularmente encorajadores, uma vez que tem se mostrado mais eficaz quando empregado na análise de textos pequenos como os encontrados no Twitter.

3. Procedimento de Pesquisa

O presente trabalho é contextualizado no protesto virtual online "bela, recatada e do lar", especificamente nas manifestações ocorridas no Twitter. No trabalho, buscou-se identificar se (e quais) perfis tiveram mais influência durante o movimento. Foi também investigada a relação temporal entre eventos ocorridos durante as manifestações e eventuais mudanças na frequência de postagens. Para essa investigação, as seguintes questões de pesquisa foram formuladas: (i) Quais perfis tiveram mais influência no movimento "bela, recatada e do lar"?; (ii) Qual foi a opinião e os sentimentos observados pelos usuários em relação ao movimento orquestrado na RSO Twitter? e (iii) Quais foram os tópicos mais comentados durante o movimento?

Para investigar a relação entre eventos ocorridos e o comportamento das postagens com a *hashtag* #BelaRecatadaEDoLar, foram coletadas notícias divulgadas em diversos jornais e *blogs*. O objetivo foi investigar ser há uma relação entre as publicações da mídia e o volume de mensagens postadas, especialmente picos e declínios no volume. Para essa investigação foram selecionados dois eventos distintos, um a favor do movimento “bela, recatada e do lar”, e outro contra.

No dia 23/04, grupos de mulheres se mobilizaram contra a imposição do padrão “bela, recatada e do lar” e protestaram em Brasília numa clara demonstração de apoio ao movimento feminista no Twitter. No dia 25/04, a esposa de um conhecido pastor brasileiro se manifestou contra o movimento feminista no Twitter, lançando uma campanha contra o movimento feminista (e a favor da posição conservadora expressa na matéria da revista). O fato, porém, só ganhou notoriedade da mídia ao final do dia 27/04, ganhando publicidade na internet a partir do dia 28/04 (dia selecionado para o segundo evento). Na Figura 1 é possível observar dois picos na distribuição. Esses picos correspondem às datas em que ocorreram os dois eventos mencionados anteriormente, o protesto feminista em 23/04 e a divulgação em noticiários online de uma campanha a favor do estereótipo conservador em 28/04.

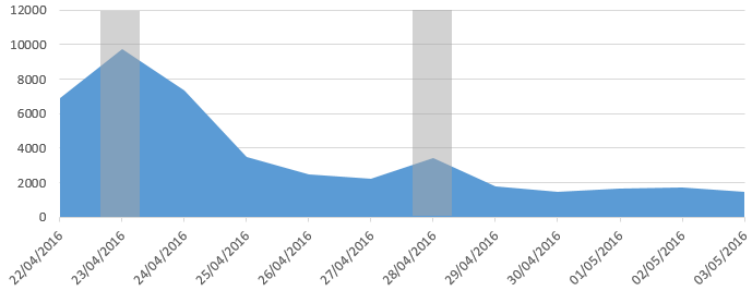

Fig. 1. Volume de *tweets* por dia com as datas dos eventos marcadas

Na primeira fase da pesquisa, foi realizada a coleta de dados, o pré-processamento e o etiquetamento manual. Foram coletadas postagens da RSO Twitter entre os dias de 22 de abril a 03 de maio. A coleta foi iniciada quatro dias após o início da manifestação virtual e encerrou-se após a constatação de que o movimento já havia atingido sua massa crítica e estava perdendo força e alcance (conforme o que pode ser observado na Figura 1). O pré-processamento consistiu na limpeza dos dados: fixou-se o idioma como "Português" e foram excluídas as mensagens em outras línguas. Além disso, foram eliminados caracteres especiais, pontuações e *stop words*. Como resultado, foi criada uma base com 43.647 mensagens em Português relacionadas ao movimento virtual "Bela, Recatada e do Lar". Para a realização dessa coleta foi usada a ferramenta Node-Red, do serviço Bluemix da IBM.

Ainda no pré-processamento, uma amostra aleatória de 1500 mensagens foi selecionada e anotada manualmente com uma das 3 etiquetas: 'positiva', 'neutra', 'negativa'. Das 1500 postagens etiquetadas, 523 foram positivas, 424 negativas e 553 neutras. A etiqueta qualifica a reação do autor da mensagem em relação ao movimento de protesto. Portanto, as mensagens etiquetadas como 'positivas' são aquelas que exprimem concordância com o protesto (e, por consequência, discordância da matéria original da revista). As mensagens cujo teor estava em desacordo com o movimento foram etiquetadas como 'negativas'. A produção e disponibilização dessa base de dados constitui a principal contribuição do presente trabalho.

Com relação à distribuição geográfica dos *tweets* coletados, apenas 1548 (3,54%) estavam georreferenciados, estando distribuídos em 460 cidades. O número baixo ocorre porque para serem georreferenciados, os *tweets* precisam ser publicados por meio de um dispositivo equipado com um receptor de sinal GPS e o recurso de localização, que consome bateria, deve estar ligado para ocorrer a marcação.

Na segunda parte da pesquisa, foi realizado o treinamento e teste de um modelo preditor feito que utiliza o algoritmo probabilístico Naïve Bayes, relatado em detalhes na Subseção 3.1, com o objetivo de definir uma linha de base para comparações futuras. A terceira parte da pesquisa foi focalizada em analisar os aspectos do referido movimento ocorrido na RSO, conforme detalhado nas Subseções 3.2 (análise da rede de *retweets*) e 3.3 (Nuvens de Palavras). Na sequência, na Subseção 3.4 são discutidos os resultados das análises.

3.1. Análise de Sentimentos

Para essa etapa foi usada uma implementação do algoritmo probabilístico Naïve Bayes em linguagem Python (versão 2.7). A estratégia de treinamento consistiu em usar a técnica de '*bag of words*' como elementos de entrada de dados. O modo de teste escolhido consistiu em separar um 1/3 das mensagens anotas para a etapa de testes, deixando 2/3 das mensagens anotadas para treinamento.

O uso de "*bag of words*" com Naïve Bayes teve o objetivo de gerar uma medida de desempenho de base para comparação de resultados por outros pesquisadores. A simplicidade da implementação do Naïve Bayes e do '*bag of words*' (ambos bem documentados na literatura) possibilitam que o estudo seja mais facilmente reproduzido por outro pesquisador, o que também contribuiu na nossa decisão a respeito do uso desse algoritmo para a demarcação da *baseline* de desempenho.

Foi observada uma acurácia de 81% na classificação. O resultado é animador, uma vez que o índice de acerto humano, cuja capacidade de análise da subjetividade de um texto oscila entre 72% e 85% [Wiebe et al. 2005]. Finalmente, após a etapa de aprendizado do algoritmo foi realizada a classificação automática de todos os 43647 *tweets* da base de dados.

3.2. Análise da Rede de *retweets*

Para a construção da rede de *retweets* foi usado o *software* Gephi, que possibilita a realização de análise de redes com grande volume de dados e que oferece recursos para filtragem, manipulação e personalização de grafos. Além disso, há várias extensões que ampliam a capacidade de exploração e análise de dados [Bastian et al. 2009].

Um grafo não direcionado de *retweets* foi criado a partir dos dados coletados. Sua composição corresponde a 21246 nós e 20026 arestas. As métricas de centralidade *Degree*, *Betweeness* e *Pagerank* [Barrat et al. 2008] foram calculadas com o objetivo de investigar a influência de usuários na rede completa.

No contexto deste trabalho, o *Degree* equivale ao número de vezes que o *tweet* de um usuário foi compartilhado. Deste modo, quanto mais um perfil foi retweetado, maior sua importância na rede. Para o *Betweeness,* os usuários que apresentam os maiores valores têm maiores probabilidades de influenciar outros usuários próximos a eles. A partir disso é possível compreender como a informação flui através da rede e quais são os perfis que ligam comunidades. Para o *Pagerank*, os nós que apresentam os valores mais altos podem ser considerados usuários-chave da rede porque outros usuários importantes interagem com eles. Ou seja, é provável que esses usuários sejam altamente ativos na RSO e possivelmente são reconhecidos como perfis importantes por outros usuários. Para o cálculo de *Pagerank* deste trabalho, foi considerado que a probabilidade (p) de um usuário acessar uma página aleatória na web é de 0,85, onde p é um fator de amortecimento que varia entre 0 e 1, usualmente definido como 0,85 [Brin and Page 1998]. A Tabela 1 representa os 10 principais usuários e suas métricas.

Tabela 1. Métricas da rede completa

Usuário	Degree	Usuário	Betweeness	Usuário	Pagerank
BolsonaroZuero	454	andandrea	0.1752	BolsonaroZuero	0.0086
ClaudiaLeitte	392	Leo_oguarda	0.0661	ClaudiaLeitte	0.0080
Antoniotabet	374	gordaopvcs	0.0589	Gordaopvcs	0.0072
Gordaopvcs	351	gifsdegatinhos	0.0499	Antoniotabet	0.0072
Frasesdebebadas	340	BolsonaroZuero	0.0438	frasesdebebadas	0.0071
MidiaNINJA	284	antoniotabet	0.0430	Leo_oguarda	0.0054
Leo_oguarda	264	MidiaNINJA	0.0408	HugoGloss	0.0050
HugoGloss	237	maarinolasco	0.0404	MidiaNINJA	0.0049
Luscas	224	luscas	0.0383	iLovePut*****	0.0047
iLovePut*****	222	nadiardgs	0.0368	Falandocarioca	0.0046

Segundo Willis *et al*. (2015) os usuários-chave de um movimento podem ser inferidos a partir da correlação entre duas métricas (no caso desta pesquisa usamos *Betweeness* e *Pagerank*). A Figura 2 ilustra o diagrama de dispersão construído a partir das métricas *Betweeness* (eixo x) e *Pagerank* (eixo y). Cada ponto representa um perfil que está entre os 60 mais retweetados. Cada um dos quatro quadrantes do diagrama possui uma interpretação particular que define como cada usuário exerce sua influência: (i) o quadrante inferior esquerdo possui usuários comuns, pois os perfis nele contidos tendem a possuir nenhum papel específico; (ii) o quadrante superior esquerdo possui os usuários que, provavelmente, exercem influência iniciando discussões e compartilhando

tweets que os outros vão compartilhar. Isso se justifica pois eles tendem a estar localizados em um dos núcleos da rede; (iii) o quadrante inferior direito abriga usuários que são importantes para uma determinada comunidade/público, pois são considerados pontes entre a produção de conteúdo e o público no qual está se conectando; (iv) e finalmente, o quadrante superior esquerdo possui usuários que apresentam as características de (ii) e (iii) combinadas. São perfis raros e indicam uma forte influência na comunidade.

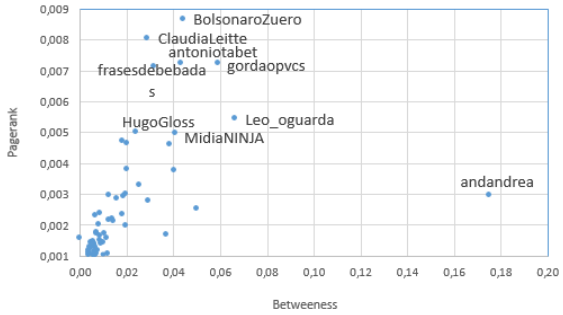

Fig. 2. Diagrama de dispersão da rede completa

Ainda na Figura 2, é possível identificar os usuários-chave do movimento. O perfil @andandrea é o único que se encaixa no item (iii) indicando sua influência dentro de comunidades específicas. Já os perfis @BolsonaroZuero, @ClaudiaLeitte, @antoniotabet, @gordaopvcs, @frasesdebebadas, @HugoGloss, @Leo_oguarda e @MidiaNINJA possuem características do item (ii) e representam usuários influentes com relação à publicação de conteúdo.

3.3. Construção de Nuvens de Palavras

A construção de nuvens de palavras foi realizada com o auxílio do *software* de mineração de dados Knime [KNIME 2017], que foi utilizado para criar nuvens de palavras por meio do processamento de grande volume de texto contido nos *tweets*.

O processo de criação de nuvens de palavras inicia com a inserção dos *tweets* no banco de dados MySQL. A partir disso, os dados são lidos, renomeados e carregados no módulo *'Text Processing'*, que processa e transforma o texto dos *tweets* (Figura 3).

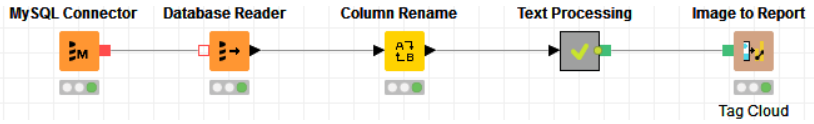

Fig. 3. Processo de criação de nuvens de palavras

Para disparar o processamento do texto, cada *tweet* foi transformado em vetores de documentos de texto associados a identificação de cada linha. Após a transformação, o processo executa o módulo '*Extract Hashtags*'. Nesse módulo, as *hashtags* são buscadas nos documentos por meio de uma expressão regular definida previamente. Após a busca, as *hashtags* são filtradas e inseridas em uma *'bag of words'* em conjunto com o documento relacionado a elas. O processo continua com o cálculo da frequência de termos relativos (TF) das *hashtags* contidas em cada documento e insere uma nova coluna que contém o valor do TF. Após o cálculo do TF, as *hashtags* semelhantes são contadas e ordenadas de forma decrescente. Logo após a ordenação, são filtradas as 10 *hashtags* mais populares. As *hashtags* são transformadas em *strings* e o processo

retorna a imagem da nuvem de palavras com as *hashtags* mais tweetadas. Foram criadas nuvens correspondentes ao movimento completo e aos dias dos eventos (23/04 e 28/04).

A Figura 4(a) ilustra a nuvem criada com as *hashtags* mais usadas no movimento. Elas se destacaram pela frequência de utilização, sendo usadas para identificar novos eventos secundários, o que pode ter sido um fator decisivo para influenciar no conteúdo e volume dos *tweets*. A *hashtag* #SOSCoup foi um movimento contra o processo de impeachment de Dilma Rousseff. Já a #BTSisonFIRE se refere ao lançamento do videoclipe da música *Fire* da banda sul-coreana BangTan Boys. #askmagcult é uma *hashtag* que indica a produção de perguntas a serem respondidas em um evento chamado Magcon. #LEMONADE está relacionado ao lançamento do sexto álbum da cantora Beyoncé. Outra tendência é #VEDA, que significa "*Vlog Everyday in April*", ou seja, ao usar a *hashtag*, cada vlogueiro se compromete em publicar um vídeo por dia no mês de abril. #METGala é um evento anual de arrecadação de fundos para o Museu de Artes Metropolitano de Nova York que reúne diversas celebridades e artistas da atualidade. Outras *hashtags* relacionadas ao movimento "bela, recatada e do lar" também se destacaram, como: #bela, #recatada e #mulhernaogostadehomemque.

Na nuvem do dia 23/04, Figura 4(b), as seguintes *hashtags* se destacaram: #DemiteoZehdeAbreu, que refere-se à cusparada do ator Zé de Abreu em casal após uma discussão sobre política em um restaurante de São Paulo. A partir disso, vários usuários do Twiter se manifestaram contra o ator através da hashtag. #EtaMundoBom, que refere-se à uma novela brasileira bastante popular. #youngmaland, *hashtag* popular na RSO Instagram e que refere-se a *selfies,* principalmente de *cosplay,* tiradas no parque temático sul-coreano Youngma Land. As *hashtags* #vejamachista, #bela e #recatada estavam relacionadas ao movimento e foram alvo de *tweets* publicados.

No dia 28/04, Figura 4(c), as *hashtags* #VEDA e o perfil do jornal #OGlobo foram mencionados, além das *hashtags* relacionadas ao tema: #veja, #mulherperfeita, #MENTIRA, #feminista, #luta e #marcelatemer. Além disso, a hashtag #sqn (só que não) também se destacou durante todo o período analisado, demostrando um indício de sarcasmo nos *tweets* relacionados ao padrão social defendido na matéria da revista.

Fig. 4. Nuvens de palavras (a) rede completa, (b) dia 23/04 e (c) dia 28/04

3.4. Discussão

A série temporal de frequências diárias de *tweets* indica uma diminuição progressiva no ritmo das postagens com o passar do tempo. No entanto, pode-se observar que dois eventos principais ocorreram nos dias de pico no volume de *tweets* (23/04 e 28/04). Por

meio da análise das *hashtags* publicadas no movimento e nos dias dos picos de volume de *tweets*, foi possível identificar que diversos eventos secundários geraram impacto no volume de *tweets*. Deste modo, existem fortes indícios de que diversos eventos menores podem gerar impactos em movimentos populares orquestrados em RSO.

Por meio da correlação das métricas de centralidade *Betweeness* e *Pagerank*, foi possível identificar os usuários que mais influenciaram o movimento "Bela, recatada e do lar". Dentre esses usuários, um exerce influência dentro de comunidades específicas e oito possuem características de influência relacionadas à publicação de conteúdo.

Por meio da observação dos *tweets* mais compartilhados no movimento, também foi possível verificar que não há correlação desses *tweets* com os eventos principais identificados anteriormente. Manifestações populares em RSO como o Twitter são difíceis de serem estudadas e compreendidas, pois tendem a perder força e alcance rapidamente na ausência de fatos e eventos que "impulsionem" o movimento.

Por fim, após classificar a base de dados em relação ao posicionamento do autor (81% de acurácia), observou-se que as mensagens mais frequentes foram de posicionamento neutro (39%). Em seguida, os posicionamentos mais frequentes são de apoio ao movimento feminista (positivo, 35%) e - por fim - o posicionamento contrário à manifestação é menos frequente (26%). Entre as mensagens neutras, verificou-se que grande ocorrência da republicação de notícias online sem emissão de opinião pelo usuário que as compartilha.

4. Conclusão e Trabalhos Futuros

No presente artigo é oferecido à comunidade um novo *dataset* de mensagens em língua portuguesa da RSO Twitter. O *Dataset* foi manualmente anotado e uma linha de base de desempenho foi demarcada com o uso de Naïve Bayes. Como resultado, foi possível compreender melhor a atuação dos usuários, as relações existentes entre os movimentos populares em RSO e as opiniões expressadas pelos usuários.

Com base nos resultados obtidos foi possível analisar os perfis que influenciaram grupos distintos de usuários e os perfis que geraram os tópicos de maior discussão de conteúdo, que foram identificados por meio da seção de análise da rede de *retweets*. Também foi possível observar que eventos secundários ocorridos durante o movimento social aparentemente geraram oscilações no volume de mensagens, fato corroborado na seção de nuvens de palavras. Além disso, entre as opiniões emitidas pelos usuários observou-se a maioria de neutros (39%), seguidos de positivos (35%) e negativos (26%). O presente estudo também indica que a aplicação de técnicas de ARS, de mineração de dados e de análises estatísticas podem ser usadas para identificar padrões de comportamentos em manifestações populares nas mídias sociais.

Para trabalhos futuros espera-se criar a rede de *retweets* com arestas direcionadas. Considera-se também a inclusão das categorias sarcasmo e de *emoticons* para auxiliar na criação de conjuntos mais eficazes de treino e testes.

Referências Bibliográficas

Barrat, A., Barthélemy, M. and Vespignani, A. (2008). *Dynamical Processes on Complex Networks*. Cambridge University Press.

Bastian, M., Heymann, S., Jacomy, M. and Others (2009). Gephi: an open source software for exploring and manipulating networks. *Proceedings of the Third International ICWSM Conference*, v. 8, p. 361–362.

Bonabeau, E. (2004). The perils of the imitation age. *Harvard Business Review*, v. 82, n. 6, p. 45–54.

Brin, S. and Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. *Computer networks and ISDN systems*, v. 30, n. 1, p. 107–117.

Carvalho, C. de S., De França, F. O., Goya, D. H. and De Camargo Penteado, C. L. (2016). The People Have Spoken: Conflicting Brazilian Protests on Twitter. In *Proceedings of the 2016 49th Hawaii International Conference on System Sciences*.

Davidov, D., Tsur, O. and Rappoport, A. (2010). Enhanced sentiment learning using twitter hashtags and smileys. In *Proceedings of the 23rd international conference on computational linguistics: posters*. Association for Computational Linguistics.

Elmasri, R. and Navathe, S. B. (2005). *Sistemas de banco de dados*. 4. ed. São Paulo: Pearson Addison Wesley.

Ferreira, M. da C. da S. (2012). Classificação Hierárquica da Atividade Económica das Empresas a partir de Texto da Web. Universidade do Porto.

KNIME (2017). KNIME | About KNIME. https://www.knime.org/about, [accessed on Jan 6].

Li, Y.-M. and Li, T.-Y. (2011). Deriving marketing intelligence over microblogs. In *Proceedings of the 44th Hawaii International Conference on System Sciences*.

Marteleto, R. M. (2001). Análise de redes sociais: aplicação nos estudos de transferência da informação. *Ciência da informação*, v. 30, n. 1, p. 71–81.

Nascimento, P., Aguas, R., De Lima, D., et al. (2012). Análise de sentimento de tweets com foco em notícias. In *Anais do Brazilian Workshop on Social Network Analysis and Mining (BraSNAM)*.

Pak, A. and Paroubek, P. (2010). Twitter as a Corpus for Sentiment Analysis and Opinion Mining. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation*.

Sarlan, A., Nadam, C. and Basri, S. (2014). Twitter sentiment analysis. In *Proceedings of the 6th International Conference on Information Technology and Multimedia*.

Weitzel, L., Quaresma, P. and De Oliveira, J. P. M. (2012). Measuring node importance on twitter microblogging. In *Proceedings of the 2nd International Conference on Web Intelligence, Mining and Semantics*.

Wiebe, J., Wilson, T. and Cardie, C. (2005). Annotating expressions of opinions and emotions in language. *Language resources and evaluation*, v. 39, n. 2, p. 165–210.

Willis, A., Fisher, A. and Lvov, I. (2015). Mapping networks of influence: tracking Twitter conversations through time and space. *Participations: Journal of Audience & Reception Studies*, v. 12, n. 1, p. 494–530.