Aplicação Web SoNDA: Social Network Data Analysis

Vinicius F. C. Ramos
Cesar Smaniotto Júnior

Introdução

A Web 2.0 estabeleceu uma série de novos conceitos além de ter promovido o usuário como gerador de conteúdo, ao invés de um mero consumidor da informação fornecida pelos *sites*. Esses avanços contribuíram para a evolução dos serviços de redes sociais *on-line*. Para este trabalho, foi escolhida a seguinte definição para *sites* de redes sociais:

> Serviços baseados na web que permitem que indivíduos (1) construam um perfil público ou semi-público dentro de um sistema limitado, (2) articulem uma lista de outros usuários com quem eles compartilham uma conexão, e (3) visualizem e atravessem sua lista de conexões e aquelas feitos por outros dentro do sistema. A natureza e nomenclatura dessas conexões podem variar de site para site (ELLISON et al., 2007, p. 211).

Serviços de *microblogging* normalmente têm as mesmas características de redes sociais citadas acima, porém costumam impor restrição ao tamanho do texto que o usuário publica. Publicações em *microblog* são, tipicamente, expressões sucintas do estado do usuário em relação a algum tema, podendo este ser qualquer mensagem relacionada ao usuário, como, por exemplo, algo do seu cotidiano, uma notícia, um evento ou outros interesses.

Java et al. (2007) diferenciam os *blogs* tradicionais de *microblogs* sob dois aspectos: primeiro, por incentivar a postagem de textos curtos, os *microblogs* requerem menos tempo do usuário para a geração de conteúdo; o outro fator que os difere é a frequência de atualização. Enquanto usuários de *blogs* costumam atualizar a sua página em um intervalo de dias, um usuário de *microblog* pode pu-

blicar várias atualizações por dia em seu perfil. Tornando, portanto, os serviços de *microblogging* uma ferramenta de comunicação extremamente ágil.

Uma das ferramentas de *microblogging* mais populares atualmente é o Twitter, o ranking Alexa[1] indica que é o oitavo *site* mais acessado no mundo. O Twitter permite aos usuários enviar mensagens de no máximo 140 caracteres, conhecidos como "tweets". Segundo dados divulgados pela própria empresa, o Twitter possui cerca de 313 milhões de usuários ativos mensalmente[2].

Não há dados atualizados sobre o número médio de *tweets* publicados por mês; a última vez que a empresa divulgou essa estatística foi em novembro de 2013, como informa Oreskovic (2015). Naquela altura, cerca de 500 milhões de *tweets* eram publicados por mês em média, e a tendência é que este número tenha crescido.

Ampofo et al. (2015) apontam que a imensidão de conteúdo gerada por usuários de plataformas de redes sociais aliado ao surgimento de ferramentas e técnicas para armazenamento dos dados em tempo real e análise automatizada foram fundamentais para que o potencial dessas informações fossem exploradas por diversas áreas, como da inteligência de negócio, da área da saúde e das ciências sociais.

A presente década também foi marcada por mobilizações sociais que ocorreram com amplo uso da Internet: Primavera Árabe, *Occupy Wall Street*, Indignados 15M, Jornadas de Junho de 2013 no Brasil, *Umbrella Revolution*, entre outras. A oportunidade de pulverização das fontes emissoras conectadas em rede elevou cada cidadão ao papel de partícipe ativo na produção de narrativas acerca dos eventos, e desestabilizou as relações de poder que antes predominavam. A comunicação não hierárquica e instantânea de redes sociais como Twitter e Facebook tem sido apontada como um fator decisivo para a expressão dessa vitalidade política.

A partir da incorporação de recursos de interconexão e compartilhamento aos processos comunicacionais entre todos os mem-

[1] http://www.alexa.com/topsites. Acessado em 08 de outubro de 2016.

[2] https://about.twitter.com/company. Acessado em 08 de outubro de 2016.

bros conectados da sociedade, os *microblogs* vêm possibilitando um paradigma comunicacional baseado em um modelo mais plural que o da mídia de massa. Um inesperado avanço em termos sociais e políticos. No entanto, as Ciências Humanas e Sociais carecem, ainda, de estudos aprofundados sobre esses fenômenos sociais catalisados e fomentados no contexto singular da comunicação em rede. Eles trazem desafios enormes para a investigação de "traços", "cursos" ou "restos" deixados pelos usuários da Internet (BARTOLOMÉ PINA et al., 2013). Seja pela inexistência de ferramentas adequadas para a pesquisa com grande quantidade de dados, seja pela falta de métodos próprios para uma pesquisa qualitativa dessa comunicação.

Diante do interesse de um grupo de pesquisadores de estudar a ação política impulsionada pelo ativismo nas redes sociais, fez-se necessário o desenvolvimento de um desenho de pesquisa próprio e, consequentemente, uma ferramenta adequada para possibilitar tal investigação. A pesquisa trata de identificar fatores e circunstâncias relevantes para a formação crítica de cidadãos no ativismo político engendrado por movimentos sociais em momentos de protesto e manifestação nas cidades brasileiras.

O objetivo, portanto, deste trabalho é apresentar a aplicação Web SoNDA (*Social Network Data Analysis*) capaz de carregar, processar e analisar grande quantidade de dados publicados no Twitter para que os pesquisadores possam identificar espaços de possibilidades e processos relevantes apresentados na metodologia apontada por Lapa et al. (2015).

Destacamos que um dos desafios da criação dessa plataforma foi o de uma integração de meios quantitativos e automatizados postos a serviço de uma análise qualitativa. Com um interesse maior do que mapear pontos de vistas na rede, visualizar a formação de comunidades e perspectivas, cartografar nós em relação à pesquisa em questão, o foco do trabalho era, portanto, filtrar qualitativamente a grande quantidade de trocas comunicativas para que o trabalho de leitura e análise por parte dos pesquisadores pudesse ser feita reduzindo, significativamente, a quantidade de mensagens a serem lidas. O objetivo final, portanto, foi identificar e selecionar criteriosamente as trocas comunicativas onde houve

diálogo para realizar uma análise de conteúdo dos mesmos com vistas a identificar fatores e circunstâncias que contribuíram para a sua existência.

Diante desse desafio metodológico, que resultou em um desafio tecnológico – de haver ferramenta/plataforma que pudesse filtrar qualitativamente a grande quantidade de dados para que pudéssemos chegar a uma quantidade manipulável de dados, que também fosse significativa para a análise de conteúdos. Apresentamos, a seguir, a aplicação Web SoNDA: sua arquitetura e funcionalidades.

Contextualização

Apresenta-se aqui uma visão geral da metodologia criada pelo grupo de pesquisa Comunic. Introduzimos brevemente cada etapa e alguns conceitos definidos pelos autores. Os detalhes podem ser verificados no trabalho de Lapa et al. (2015).

Metodologia de análise de redes sociais do grupo Comunic

Com o propósito de investigar a ação política motivada pelo ativismo nas redes sociais, o núcleo UFSC do grupo de pesquisa Comunic, formado por pesquisadores da área de Ciências Sociais e Humanas, elaborou uma metodologia de análise qualitativa de postagens de redes sociais. A aplicação deste estudo pretende identificar fatores e circunstâncias relevantes para a formação crítica nas redes sociais. O objetivo final é a construção de um referencial de fatores e circunstâncias para a formação de professores e educadores.

A Figura a seguir apresenta o diagrama com as etapas da metodologia. Ao final de cada uma dessas fases o conjunto de dados de interesse é reduzido, de modo que após a última etapa de mineração permaneçam apenas os *posts* mais relevantes para a pesquisa, quando serão submetidos à análise de conteúdo nos diálogos. Os itens a seguir detalham as atividades de cada etapa.

Figura
Etapas da metodologia de análise do grupo Comunic

Fonte: Elaborado pelos Autores

- *Etapa 1*: consiste na coleta de postagens de redes sociais por um determinado período de tempo. São obtidos *posts* que contenham palavras-chaves previamente escolhidas e relacionadas ao tema da análise. Além da mensagem em si, outras informações como: nome de usuário, data de publicação e geolocalização (quando disponível), são capturadas. Após a aquisição dos dados, o processo de análise inicia. As próximas três etapas visam obter resultados cada vez mais qualitativos ao término do processo.
- *Etapa 2*: nessa fase é realizado o primeiro filtro no conjunto de dados. Os *posts* são separados em categorias especificadas pelos pesquisadores, que a metodologia define como Espaços de Possibilidade. Segundo os autores, Espaços de Possibilidade são momentos com potencial para revelar processos que devem ser observados de acordo com os propósitos da pesquisa. Essa etapa se divide em duas atividades. A primeira é a identificação das categorias. A partir da leitu-

ra de amostras extraídas do conjunto de dados, os pesquisadores formulam as categorias. São obtidas três amostras, do início, meio e fim do *dataset* ordenado por *posts* em ordem crescente de data de publicação. Cada categoria é composta pela descrição do seu significado e por uma biblioteca de termos e palavras-chave. No segundo estágio todo o *dataset* é filtrado com base no dicionário de palavras-chave de cada Espaço de Possibilidade. Essa filtragem no conjunto de dados é feita verificando a ocorrência de um termo ou palavra de um Espaço de Possibilidade no texto da postagem. É importante ressaltar que as categorias não são mutuamente exclusivas, portanto, um *post* pode pertencer a zero ou mais categorias. A metodologia inicialmente propõe esse método de ocorrência de termo no texto para realizar a mineração por Espaços de Possibilidade. O presente trabalho introduz um segundo método para a filtragem por Espaços de Possibilidades, não previsto na metodologia, que é através da aplicação de técnicas avançadas de mineração de textos.

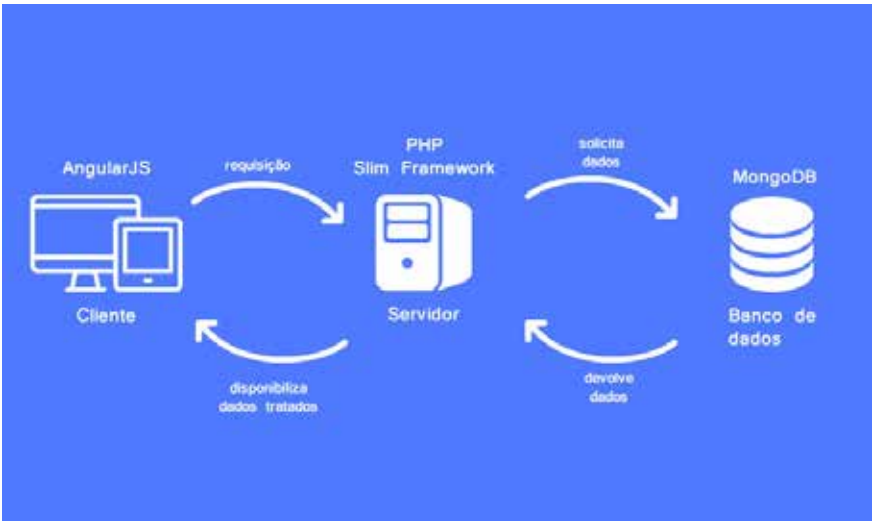
- *Etapa 3*: nesse estágio o conjunto de dados ainda é volumoso e necessita de mais um processo de filtragem até atingir uma quantidade de *posts* manipuláveis para proceder à análise de diálogos. A metodologia define essa etapa como mineração por Fatores e Circunstâncias, que são um segundo conjunto de categorias, definidos anteriormente através de revisão bibliográfica, que descrevem os processos buscados dentro dos Espaços de Possibilidade. Cada Fator e Circunstância é composto pela sua descrição e pelas métricas, que servem como guia para o analista identificar se um *post* dentro dos Espaços de Possibilidades pertence ou não a um dado fator e circunstância. Diferente da mineração por Espaços de Possibilidade, onde a categorização é feita de forma automática, nessa etapa o processo de categorização é feito manualmente pelo analista.
- *Etapa 4*: definida pelos autores como fase de compilação dos diálogos. Nessa fase da análise o conjunto de dados foi reduzido de modo a permanecer apenas os *posts* de maior interes-

se para a pesquisa. A partir de *posts* potenciais identificados e anotados na fase anterior, é recuperado todo o rastro de diálogo no qual um dado *post* faz parte. Com os diálogos recuperados, o analista pode efetuar análise de conteúdo e buscar respostas para suas investigações.

Arquitetura da Aplicação Web SoNDA

A SoNDA (*Social Network Data Analysis Platform*) é uma plataforma Web livre e de código aberto, desenvolvida usando-se o *framework* Slim (http://www.slimframework.com/), um micro *framework* de código aberto, escrito em PHP 5 e distribuído sob a licença do MIT (https://opensource.org/licenses/MIT). A Figura a seguir apresenta a arquitetura tecnológica da plataforma.

Figura
Arquitetura tecnológica da aplicação Web SoNDA



Fonte: Elaborado pelos Autores

Em termos do *back-end* da plataforma, foi construída uma API baseada em alguns princípios da arquitetura REST. O termo REST foi definido em 2000, na tese de doutorado de Roy Fielding e é acrônimo para *Representational State Transfer* (Transferência de Estado

Representacional). É um estilo arquitetural para projetar aplicações em rede.

A própria Web em si pode ser vista como uma arquitetura baseada em REST.

Em termos práticos, uma das vantagens da utilização de uma API para o acesso e modificação dos dados é a possibilidade de desenvolver paralelamente o *back-end* e *front-end* da aplicação.

Devido ao fato de os dados coletados nas redes sociais serem não estruturados, optamos pelo seu armazenamento em uma base de dados NoSQL.

O MongoDB é um sistema de gerenciamento de banco de dados categorizado como uma base de dados de armazenamento de documentos (*document store*) ou base de dados orientada a documentos. Sua principal característica é a utilização de um esquema livre de organização de dados, que significa: registros diferentes podem ter diferentes colunas, colunas podem ter mais de um valor (vetores), os registros podem ter estruturas aninhadas e os tipos de valores de cada coluna individualmente podem ser diferentes para cada um dos registros. Outra característica do MongoDB é a utilização de documentos do tipo da Notação de Objetos JavaScript (uma livre tradução para *JSON-like*) com esquemas dinâmicos de armazenamento de dados. O MongoDB é livre e de código aberto, com licença de utilização que é uma combinação da GNU Affero General Public License e da Apache License.

O *front-end* foi implementado usando o AngularJS, um *framework* Javascript *open-source*, mantido pelo Google, que auxilia na construção de *single-page applications*. Uma das principais características do *framework* é a ligação bidirecional dos dados (*two-way data binding*). Funciona da seguinte forma: o navegador renderiza o *template* em HTML puro, com os dados contidos em um escopo definido na *model*; essa renderização produz uma "visualização ao vivo", onde quaisquer mudanças na *view* são refletidas na *model* e da mesma forma alterações realizadas na *model* são sincronizadas automaticamente na *view*. O AngularJS está sob licença do MIT. O código-fonte da aplicação está disponível no Github sob licença MIT.

Funcionalidades da Aplicação Web SoNDA

O desenvolvimento da SoNDA partiu do objetivo de fornecer uma plataforma Web para usuários que querem observar o ativismo social nas redes sociais.

Para atingir esse objetivo baseamos o desenvolvimento da plataforma a partir da metodologia de pesquisa proposta pelo grupo de pesquisa Comunic, apresentada sucintamente na Seção 2.

Requisitos de *software* "são as descrições do que o sistema deve fazer, os serviços que oferece e as restrições a seu funcionamento" (SOMMERVILLE, 2011, p. 57).

Os requisitos de *software* se caracterizam como requisitos funcionais e não funcionais. Ainda segundo o autor, requisitos funcionais são especificações dos serviços que devem ser oferecidos pelo sistema e de como o sistema deve se comportar a determinadas entradas e situações. Requisitos não funcionais são restrições impostas aos serviços existentes no sistema. Desempenho, segurança e usabilidade são exemplos de requisitos não funcionais.

Os principais requisitos funcionais do sistema são:

1. Gerência de projetos: o sistema deve permitir a criação de novos projetos e a respectiva gerência por parte do administrador. Ações de administração de um projeto consistem na edição e remoção do projeto, além do controle de acesso, que envolve a adição/remoção de usuários de um projeto, bem como a alteração do nível de acesso. A gerência de um projeto envolve ações básicas como o cadastro, edição e remoção de um projeto. Além disso, compreende as ações de convidar/remover um usuário de um projeto e a alteração dos seus níveis de privilégio.
2. Importação de *dataset* do Twitter: a aplicação deve prover mecanismos para possibilitar o *upload* de um arquivo texto no formato CSV contendo *posts* do Twitter.
3. Geração de recortes temporais do *textit*{*dataset*}: o sistema deve permitir a extração de fragmentos do início, meio ou fim do *dataset*, seguindo o critério de ordem de data de publicação crescente.

4. Gerência de categorias: o sistema deve permitir a adição de novas categorias, bem como a edição e remoção das previamente cadastradas.
5. Mineração por Espaços de Possibilidade: o sistema deve ser capaz de filtrar os *datasets* por categorias de Espaços de Possibilidade pela ocorrência de termo ou palavra-chave no *tweet.*
6 Codificação de *tweets* em Fatores e Circunstâncias: a aplicação deve prover mecanismos para o usuário atribuir um Fator e Circunstância a um *tweet*.
7. Gerência de matrizes: o sistema deve permitir a adição de novas matrizes, bem como a edição e remoção das previamente cadastradas.
8. Questionar *tweets* codificados: a aplicação deve fornecer mecanismos para interrogar a base de dados, como consultas e execução de matrizes.
9. Recuperar o diálogo a partir de um *tweet*: o sistema deve permitir a recuperação de um rastro de conversação a partir de um *tweet* existente no *dataset*.
10. Exportação de resultados: o sistema deve ser capaz de exportar os resultados das consultas e cruzamentos no *dataset* e dos diálogos recuperados para um arquivo CSV.

A seguir descrevem-se como os requisitos funcionais foram implementados pela aplicação e suas respectivas etapas da metodologia.

Etapa 1: coleta e definição do dataset

Para se começar uma análise do conteúdo utilizando-se a plataforma é necessário criar um projeto, descrevê-lo e cadastrar os usuários que fazem parte dele (os usuários também podem fazer o cadastro posteriormente). Cada projeto possui um ou mais *dataset*.

Um *dataset* é uma coleção de dados coletados de redes sociais durante um período de tempo delimitado e importado para o sistema. Cabe ressaltar que a coleta dos dados não é feita pela plataforma. Essa escolha deve-se ao fato de já existirem diversas ferramentas de extração/coleta para as diferentes redes sociais existentes.

Uma categoria, por sua vez, pode ser um Espaço de Possibilidade ou um Processo Relevante, ambos os termos foram definidos em (COELHO et al., 2015) e são apresentados com mais detalhes a seguir.

Para ter acesso ao sistema o usuário necessita estar logado. Ao realizar seu cadastro no sistema, aguarda-se a aprovação de um usuário com papel de administrador. Após aprovação, o usuário poderá acessar o sistema e participar de algum projeto de análise de dados. A participação em projetos está condicionada também à aprovação por usuários já cadastrados no projeto, ou seja, todo usuário cadastrado em um projeto poderá incluir um novo usuário.

Usuários possuem perfis para restringir o acesso a certas funcionalidades do sistema. Usuários com perfil de administrador têm permissão de acesso a todas as partes do sistema. Ele é o único capaz de aprovar/desativar o acesso de um usuário e modificar o seu perfil. Usuários caracterizados como moderador têm permissão de excluir e modificar alterações feitas por outros usuários, além de criar novos projetos e importar *datasets*. O usuário definido como pesquisador pode visualizar todas as informações referentes aos projetos, porém só pode modificar informações cadastradas por ele, como uma categoria, por exemplo.

A primeira etapa da metodologia, portanto, consiste na coleta de postagens do Twitter por um determinado período de tempo. Essa etapa captura os *tweets* de acordo com palavras-chave e/ou *hashtags* previamente escolhidas e relacionadas ao tema em que se deseja fazer a análise. Além da mensagem em si, são capturadas outras informações relevantes, tais como data de publicação e a geolocalização.

Os dados aqui capturados e, consequentemente, armazenados, formam o que definimos anteriormente como *dataset*. Cabe ressaltar que essa fase não está contemplada na ferramenta. Ao término da fase de coleta, o pesquisador faz a importação do *dataset* para a ferramenta via arquivo CSV (*Comma-Separated Values*) ou JSON (*JavaScript Object Notation*).

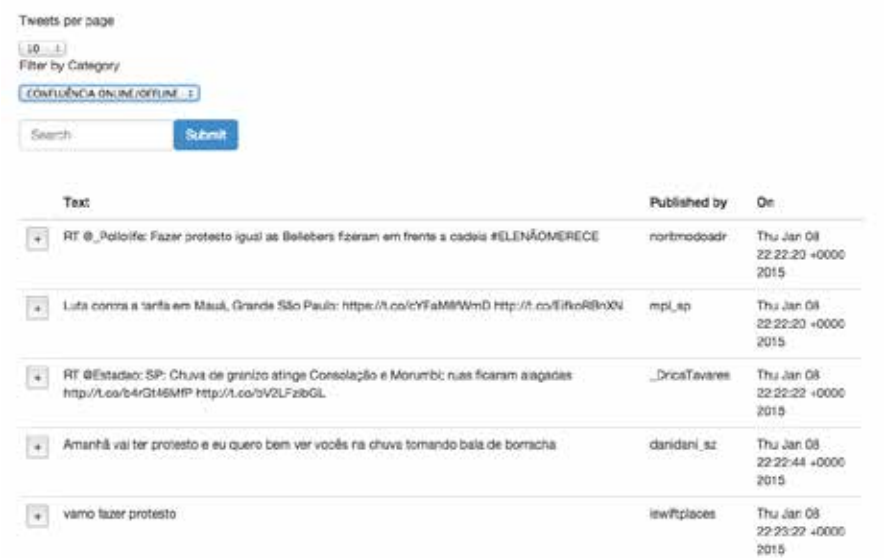
Etapa 2: Filtragem dos dados como Espaços de Possibilidade

Na segunda etapa são extraídas algumas amostras do conjunto de dados para que seja feita a leitura e definição de termos e palavras-chave para filtragem dos *datasets*. Essas amostras são recortes em ordem temporal do início, meio e fim do projeto. Com base na quantidade de *tweets* extraídos e do período de tempo de coleta, o usuário define o tamanho de cada recorte. A filtragem inicial tem como objetivo agrupar os *tweets* em categorias predefinidas. Os termos e palavras-chaves de uma categoria podem ser, além de palavras em si, *hashtags* ou expressões regulares. As expressões regulares devem estar descritas seguindo o padrão PERL. Por exemplo, ao definirmos a categoria "Universidade", seus respectivos termos ou palavras-chave poderiam ser: UFSC, UFRJ, universidade, ciência da computação, sistemas de informação e acadêmic[ao]. O uso de expressões regulares torna o conjunto de palavras-chave menos redundantes.

A partir do cadastramento das categorias e seus respectivos termos e palavras-chave, a plataforma SoNDA agrupa os *tweets* em cada uma das categorias, ou seja, os *tweets* que possuírem termos (ou palavras-chave) de uma determinada categoria serão agrupados naquela categoria. É importante ressaltar que as categorias não são mutuamente exclusivas, portanto, um *tweet* pode se apresentar em zero ou mais categorias.

Todas as categorias predefinidas, nesse momento são chamadas de Espaços de Possibilidades, que, segundo a metodologia proposta por Lapa et al. (2015), são momentos com potencial para revelar processos que devem ser observados de acordo com os propósitos da pesquisa; portanto, chamamos esse agrupamento de "filtragem por espaços de possibilidade". A Figura a seguir apresenta o resultado de uma filtragem por Espaços de Possibilidades.

Figura
Filtragem por Espaços de Possibilidade: categoria Confluência *on-line/ off-line* selecionada e as colunas Text (que representa os *Tweets,* Published by (que representa quem postou a mensagem) e On (com a data da postagem).



Fonte: Extraído da plataforma SoNDA

Visto que os *datasets* podem conter *tweets* que fogem do assunto da pesquisa, esses podem ser ocultados, tanto pela etapa de categorização quanto por um usuário.

Caso um usuário oculte um *tweet*, ele não aparece em mais nenhum resultado (o usuário que ocultou esse *tweet,* ou o administrador, podem desfazer a ação).

Além da filtragem por Espaços de Possibilidade, é possível aplicar outros dois critérios de seleção sob os *datasets*: selecionar todos os tweets publicados por um determinado usuário ou fazer uma pesquisa por palavras ou conjunto de caracteres.

Por exemplo, usando a filtragem por conjunto de caracteres é possível isolar todas as menções a usuários do Twitter presentes no projeto. Todas as filtragens e critérios de seleção podem ser exportados para um arquivo CSV e serem carregados em um *software* de visualização de redes, como o Gephi.

A partir dos resultados da fase anterior, são definidas métricas,

conforme o propósito da pesquisa, para identificar as publicações mais relevantes dentro dos Espaços de Possibilidades. Essas publicações contêm o que os autores chamam de Processos Relevantes.

Etapa 3: Codificação por Fatores e Circunstâncias

O processo de codificação por Fatores e Circunstâncias compreende a terceira etapa da metodologia, onde os pesquisadores codificam os *tweets* segundo as métricas especificadas para cada uma dessas categorias. De modo que o SoNDA permite que múltiplos usuários compartilhem o mesmo projeto e manipulem as mesmas fontes de dados; a codificação de cada usuário é feita independentemente. Dessa forma, cada usuário codifica os *tweets* sem saber se eles já foram codificados por outra pessoa.

O ambiente para codificação é o mesmo onde são apresentados os resultados da Mineração por Espaços de Possibilidades. A Figura a seguir apresenta a mesma interface com algumas instâncias codificadas. A codificação é feita através das ações de arrastar e soltar (*drag-and-drop*) uma das categorias de Fatores e Circunstâncias listadas até o *tweet* onde deseja-se codificar.

Figura
Filtragem por Diálogo após a inclusão das categorias
(à esquerda) abaixo de cada um dos *tweets*

Fonte: Extraído da plataforma SoNDA

Ao encontrar um *tweet* em potencial o usuário pode marcar em

um ou mais processos relevantes previamente cadastrados. Nessa etapa da metodologia a análise é qualitativa; sendo assim, os *tweets* são anotados manualmente. Da mesma forma que foi feito anteriormente, é possível filtrar os *datasets* por processos relevantes e exportar os resultados. Por exemplo, é possível isolar todos os *tweets* que foram anotados com o processo relevante "PLURALIDADE". Nos *tweets* marcados com algum processo relevante, é necessário fazer uma análise no diálogo.

Ao final dessa fase todos os dados estão listados e organizados em suas categorias (Espaços de Possibilidades e Processos Relevantes). É nesse momento que a análise qualitativa se inicia, permitindo ao pesquisador encontrar mais facilmente as postagens relacionadas às categorias predefinidas. A análise, que leva aos resultados de uma pesquisa, é feita pelos pesquisadores que, no nosso primeiro estudo de caso, teve como objetivo principal a busca por elementos para reflexões acerca da formação crítica de sujeitos na ação política empreendida em espaços sociais *on-line*.

Etapa 4: Compilação dos Diálogos

A última etapa da metodologia prevê a análise dos diálogos recuperados a partir de algumas postagens de interesse. O Twitter possui uma API pública que disponibiliza diversas interfaces para acessar seus dados como, por exemplo, a recuperação dos *tweets* mais recentes de um usuário. Porém, a API não disponibiliza um método que, dado um *tweet*, recupere toda a sua *thread* de conversação, o que obriga a adotar outra abordagem para resolver esse problema.

Desse modo, foi implementado um *web crawler* para cumprir esse requisito. A partir de um *tweet*, independente da posição em que apareça em um diálogo, o *crawler* navega até o *tweet* "raíz", isto é, aquele que gerou todas as trocas comunicativas. Obtido o *tweet* raíz, o *crawler* recupera todos os *tweets* subsequentes, através da extração do conteúdo de *tags* HTML da página. Os itens extraídos são salvos no banco de dados da aplicação e apresentados ao usuário.

A Figura 5 mostra um *tweet* presente nos resultados gerados pelo módulo de questionamento do sistema. Ao clicar em "ver conversação", o sistema recupera o diálogo armazenado no banco de dados, caso a extração tiver sido previamente feita. Caso contrário,

o *crawler* tentará recuperar e armazenar os *tweets* envolvidos no diálogo e, por último, o sistema apresenta os resultados, como mostra a Figura a seguir. O bloco com a borda esquerda na cor azul identifica o *tweet* usado para a recuperar as demais interações.

Figura
Diálogo recuperado pelo sistema à partir do *tweet* da Figura 5.

Fonte: Extraído do Twitter

Essa estratégia para recuperação dos diálogos possui limitações e em pelo menos três situações identificadas, não é possível obter o diálogo: quando o autor da postagem utilizada para recuperar as outras interações apagou sua conta ou alterou a privacidade da conta de pública para privada, e quando a postagem em questão tiver sido excluída. Nesses casos, o sistema informa o erro ao usuário.

Considerações finais

O maior benefício da Aplicação Web SoNDA é de permitir a manipulação de grande quantidade de dados (Big Data) para análise qualitativa. Outro ponto importante é a possibilidade de inclusão de dados de diferentes redes sociais. Cabe ressaltar que a etapa de extração de dados das redes sociais não é contemplada nessa ferramenta.

Essa escolha deve-se ao fato de já existirem diversas ferramentas de extração para as diferentes redes sociais existentes. Assim, focamos apenas no processo de análise dos dados e não na integração de ferramentas e APIs proprietárias oferecidas pelas redes sociais.

Atualmente a filtragem por Espaços de Possibilidades é efetuada através de busca por termos de uma categoria em um *tweet*, onde algumas mensagens fora de contexto são incluídas nos resultados. A intenção é que esses ruídos sejam minimizados com o uso de técnicas de aprendizagem de máquina para a filtragem inicial do *dataset*. O pesquisador fornece um subconjunto do *dataset*, com os *tweets* anotados com seus respectivos Espaços de Possibilidade, para que a ferramenta construa um modelo para classificar os demais *tweets*, em função das instâncias de exemplo.

REFERÊNCIAS

AMPOFO, L.; COLLISTER, S.; O'LOUGHLIN, B.; CHADWICK, A. Text mining and social media: when quantitative meets qualitative and software meets people. **Innovations in Digital Research Methods**, pages 161–91, 2015.

BARTOLOMÉ PINA, A. R.; SOUZA, F. N.; LEÃO, M. C. Investigación educativa a partir de La información latente en internet. **Revista Eletrônica de Educação**, v.7, n. 2, p. 301-316, 2013.

COELHO, I. C.; LAPA, A. B.; RAMOS, V.; MALINI, F. **A research design for the analysis of contemporary social movements**. 5th Workshop on Making Sense of Microposts, p. 38-42, 2015.

ELLISON, N. B. et al. Social network sites: definition, history, and scholarship. **Journal of Computer-Mediated Communication**, 13(1):210–230, 2007.

JAVA, A.; SONG, X.; FININ, T. TSENG, B. **Why we twitter:** understanding microblogging usage and communities. In: Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis, pages 56–65. ACM, 2007.

LAPA, A. B.; COELHO, I. C.; RAMOS, V. F. C.; MALINI, F. **Fatores e circunstâncias para o empoderamento do sujeito nas redes sociais – um desenho de pesquisa**. CONGRESSO IBERO-AMERICANO EM INVESTIGAÇÃO QUALITATIVA (CIAIQ) 2015, volume 2, 2015.

ORESKOVIC, A. **Here's another area where twitter appears to have stalled**: tweets per day. 2015. Disponível em: http://www.businessinsider.com/twitter-tweets-per-day-appears-to-have-stalled-2015-6.

SOMMERVILLE, I. **Engenharia de software**. 9. ed. São Paulo: Pearson Brasil, 2011.