DETECÇÃO DE TÓPICOS DE INTERESSE EM TEXTOS DE FÓRUNS DE DISCUSSÃO DO PORTAL E-DEMOCRACIA

Eixo Temático: Tecnologias da Informação Aplicadas
Modalidade: Apresentação Oral

Alexandre Ribeiro Afonso
Renata dos Santos Brandão
Jeyce Ferreira dos Santos

1 INTRODUÇÃO

A identificação do sentimento ou característica de opinião majoritária dos internautas sobre um produto, evento ou indivíduo em uma rede social, considerando centenas ou milhares de usuários, é algo de interesse às organizações, pois permite levar à elaboração de estratégias de investimento em um sistema mercadológico competitivo (LOPES et al., 2008), ou mesmo à elaboração de políticas públicas.

É possível alcançar o conhecimento genérico considerando um conjunto de opiniões sobre um tópico (indivíduo, produto, evento) de diversas maneiras. Na literatura, a descoberta da polaridade das opiniões (positiva, negativa, neutra) tem sido um tema recorrente, com diferentes técnicas empregadas (BECKER; TUMITAN, 2013). Vários fatores, complexos, devem ser considerados na busca e análise dos dados: primeiramente, é evidente que as línguas diferem entre si nos níveis morfológico, léxico, sintático, semântico e discursivo-pragmático, logo, um mecanismo de busca construído especificamente para o inglês, por exemplo, pode não trabalhar de forma satisfatória com outras línguas. Recentemente, considerando o português do Brasil, Balage Filho e Pardo (2014) descrevem uma ferramenta para busca de opiniões na web; do ponto de vista linguístico, Scopim (2011) descreve algumas características lexicais de textos opinativos nesta língua.

Em relação ao conteúdo das postagens, é visível que os tópicos para os quais se buscam polaridades de opiniões podem ter estrutura formada com um simples nome, ou com estrutura mais complexa, contendo nomes, adjetivos, preposições e numerais. A posição do adjetivo em relação ao nome que ele qualifica apresenta maior flexibilidade no português do Brasil, sendo mais "livre" a posição que o adjetivo ocupa, se comparada à do inglês, por exemplo, que apresenta uma ordem bastante rígida: adjetivos sempre antepostos ao nome (CALLOU *et al*, 2003). Ainda, a expressão do sentimento poderia ser realizada por caracteres especiais e

desenhos, ao sair de um padrão linguístico formal puramente alfabético. Como outros exemplos de tal complexidade, somam-se a falta de padronização de escrita no texto coloquial, o uso de gírias e expressões regionais e populares, e ainda, a presença de elipses no texto opinativo em vez da identificação da entidade sujeita às opiniões.

Considerando tal complexidade, este texto descreve um primeiro experimento com postagens de um portal com fóruns de discussão on-line, onde internautas opinam e dialogam sobre tópicos, em português brasileiro; tópicos também levantados pelos próprios internautas. Neste estágio inicial de estudos, o objetivo foi verificar uma maneira de identificar tópicos de interesse (indivíduos, produtos, eventos, etc.) dos internautas que comentam. Após a identificação de tais interesses pretende-se focar em outras questões, por exemplo, de polaridade das opiniões sobre tais tópicos (se a postagem for opinativa), baseando-se em algum estudo linguístico mais aprofundado.

2 QUESTÃO DE PESQUISA

Neste primeiro conjunto de experimentos, consideram-se buscas em um *corpus* montado, contendo centenas de postagens textuais do portal citado. Procuraram-se tópicos de interesse utilizando softwares de etiquetagem morfossintática, softwares de busca de padrões linguísticos e anotando as palavras simples e compostas mais frequentes. A partir daí, verificam-se os resultados e espera-se localizar tais tópicos, de maior interesse público.

3 METODOLOGIA

3.1 DADOS COLETADOS

A base de postagens deste trabalho foi retirada do portal de relacionamentos online *e-Democracia*[1]. O portal objetiva incentivar a participação da sociedade no debate de temas importantes para o país. O portal possui uma forma de interação entre internautas denominada Comunidades Legislativas com temas pré-definidos (*Sistema Único de Saúde, Reforma Política, 1ª CONSOCIAL VIRTUAL*, etc.). As Comunidades Legislativas possuem fóruns, por exemplo, (*#Tema 1 - Transparência:*

[1] http://edemocracia.camara.gov.br/

~~acesso às informações do~~ poder público, #Tema 3 - Controle: fortalecimento *dos conselhos de políticas públicas*), cada fórum possui postagens iniciais, por exemplo, (*Combate à Corrupção, Biblioteca*) com título e conteúdo textual anexado. As postagens vindouras à postagem inicial também contêm título e conteúdo textual e podem ser respostas a comentários postados previamente, o que leva a um ambiente de interação social e discussão.

Para esta pesquisa, a base de postagens descrita foi filtrada, somente as postagens emitidas foram selecionadas, sem seus títulos, e reunidas em arquivo texto único. O resultado desta filtragem constitui exatamente o *corpus* onde se trabalhou as ferramentas computacionais, com centenas de postagens.

3.2 PROCEDIMENTOS EM LINGUÍSTICA COMPUTACIONAL

De posse do arquivo texto (*corpus*), com as postagens dos internautas nas comunidades, foi executado o etiquetador *MxPost* para o português do Brasil sobre tal arquivo. O etiquetador, descrito por Aires (2000), tem como objetivo agregar uma *tag* (etiqueta morfossintática) a cada palavra [2] simples do arquivo texto de mensagens. As seguintes etiquetas são consideradas pelo *MxPost*:

```
Contrações
---------------

PREP+ART - preposição + artigo
PREP+PD * - preposição + pronome demonstrativo
PREP+PPR * - preposição + pronome pessoal reto
PREP+N - preposição + nome
PREP+PPOT * - preposição + pronome pessoal obliquo tônico

Tagset Reduzido
-------------------------
I - interjeição; LOCU - locução; PREP - preposição; N - nome; NP - nome
próprio; VERB - verbo; ADJ - adjetivo; AUX - verbo auxiliar; ADV -
advérbio; PRON - pronome; CONJ - conjunção; NUME - numeral; ART -
artigo; RES - resíduo; PDEN - palavra denotativa.
```

Após tal etiquetagem, foi executado um software para listagem de palavras (simples ou compostas) e suas frequências no *corpus*, utilizando-se expressões regulares. Estas buscas são exatamente a essência da pesquisa realizada, cada

[2] Considera-se palavra simples uma cadeia de símbolos que está entre dois espaços em branco, desconsiderando-se sinais de pontuação, no texto. A palavra composta surge a partir da junção e palavras simples.

arquivo gerado a partir do *corpus* original contém uma tabela descrevendo palavras e suas frequências, a tabela de palavras/frequências de cada arquivo é gerada pelo resultado de uma expressão regular sobre o *corpus*. No quadro a seguir, leia o símbolo "_" como "concatenação" e "%" como "união". Cada expressão regular descrita a seguir gera um arquivo/tabela distinto.

Palavras Simples:

Arquivo **N%NP**: contém elementos com estrutura: substantivo ou nome próprio.

Palavras Compostas:

Arquivo **N%NP_N%NP%ADJ**: contém elementos com a estrutura (substantivo ou nome próprio seguido de (substantivo ou nome próprio ou adjetivo)).

Arquivo **N%NP_PREP%PREP+ART_N%NP**: contém elementos com a estrutura (substantivo ou nome próprio seguido de (preposição ou contração de preposição com artigo) e em seguida substantivo ou nome próprio).

Arquivo **N%NP_PREP%PREP+ART_N%NP_ADJ**: contém elementos com a estrutura (substantivo ou nome próprio seguido de (preposição ou contração de preposição com artigo) e em seguida nome próprio ou substantivo seguido de um adjetivo).

É visto que a busca gerada por cada expressão regular não é perfeita. O software de etiquetagem falha, por exemplo, ao classificar uma palavra como nome (N), sendo que na realidade sua função morfossintática na oração é de verbo. Isso acontece, pois a taxa de acertos do etiquetador é alta (97%), mas este tem problemas de classificação pela restrita abrangência do treinamento dado, pois foram utilizados *corpora* de treinamento ao etiquetador que não são originários de textos de redes sociais.

4 RESULTADOS

Os gráficos a seguir exibem as frequências de palavras (eixo y) para as 15 (quinze) palavras mais frequentes (eixo x). Cada gráfico representa um arquivo gerado por uma expressão regular a partir do *corpus* de postagens reunido.

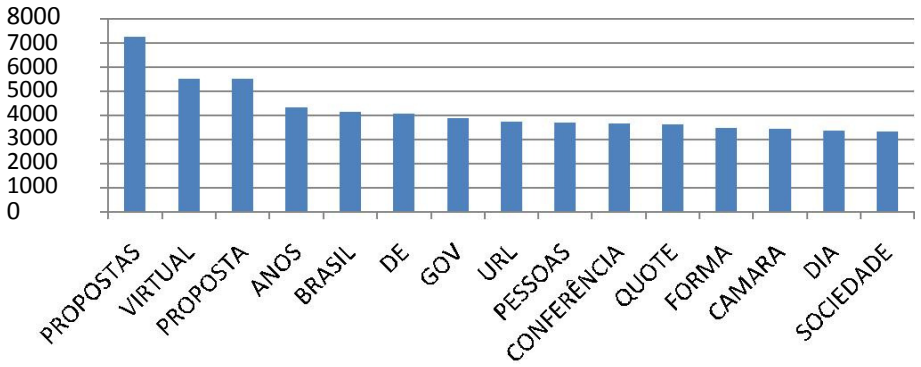

Figura 1 - Frequências (eixo y) para as 15 palavras mais frequentes (eixo x) buscadas pela expressão **N%NP** no corpus de testes.

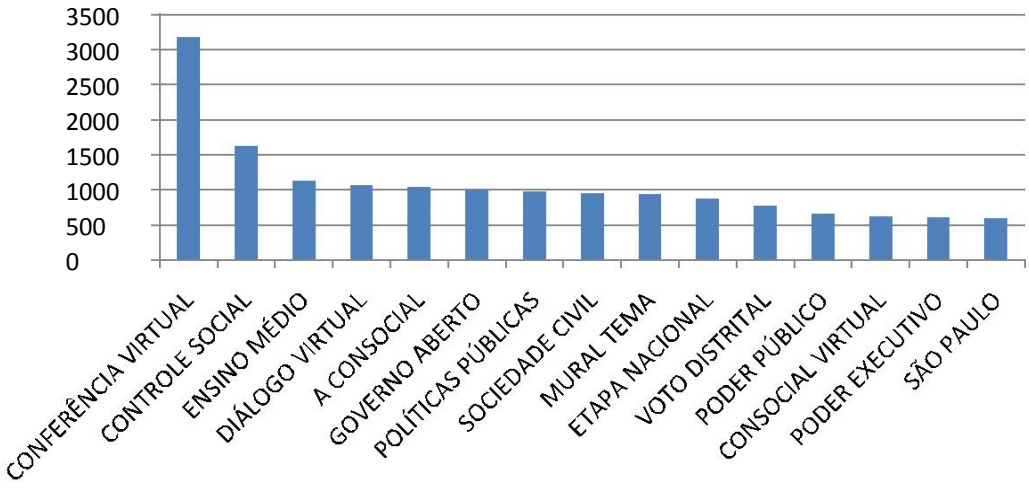

Figura 2 - Frequências (eixo y) para as 15 palavras mais frequentes (eixo x) buscadas pela expressão **N%NP_N%NP%ADJ** no corpus de testes.

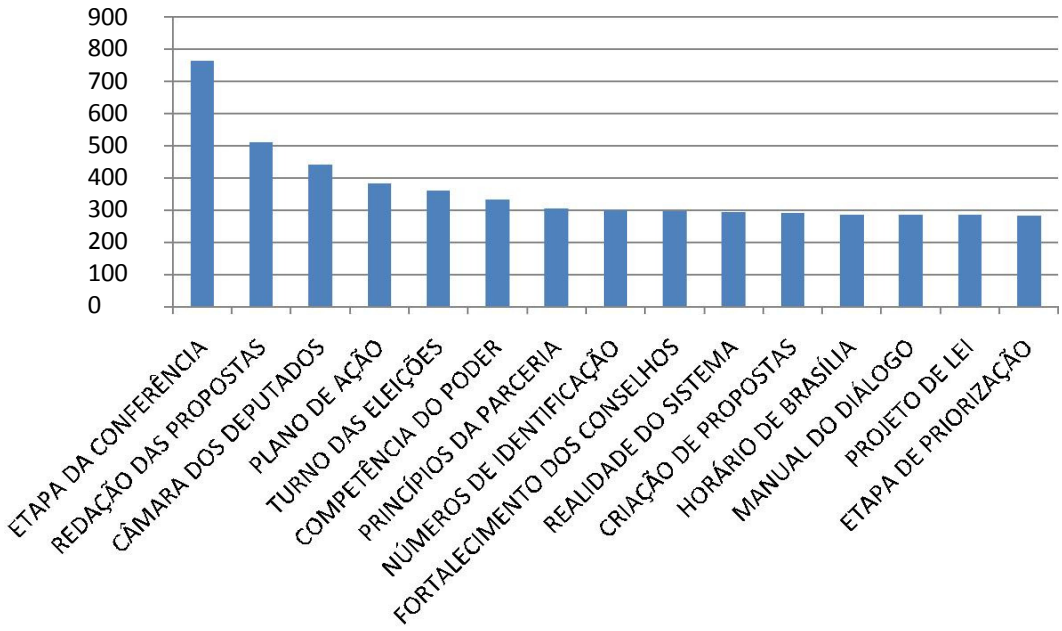

Figura 3 - Frequências (eixo y) para as 15 palavras mais frequentes (eixo x) buscadas pela expressão **N%NP_PREP%PREP+ART_N%NP** no corpus de testes.

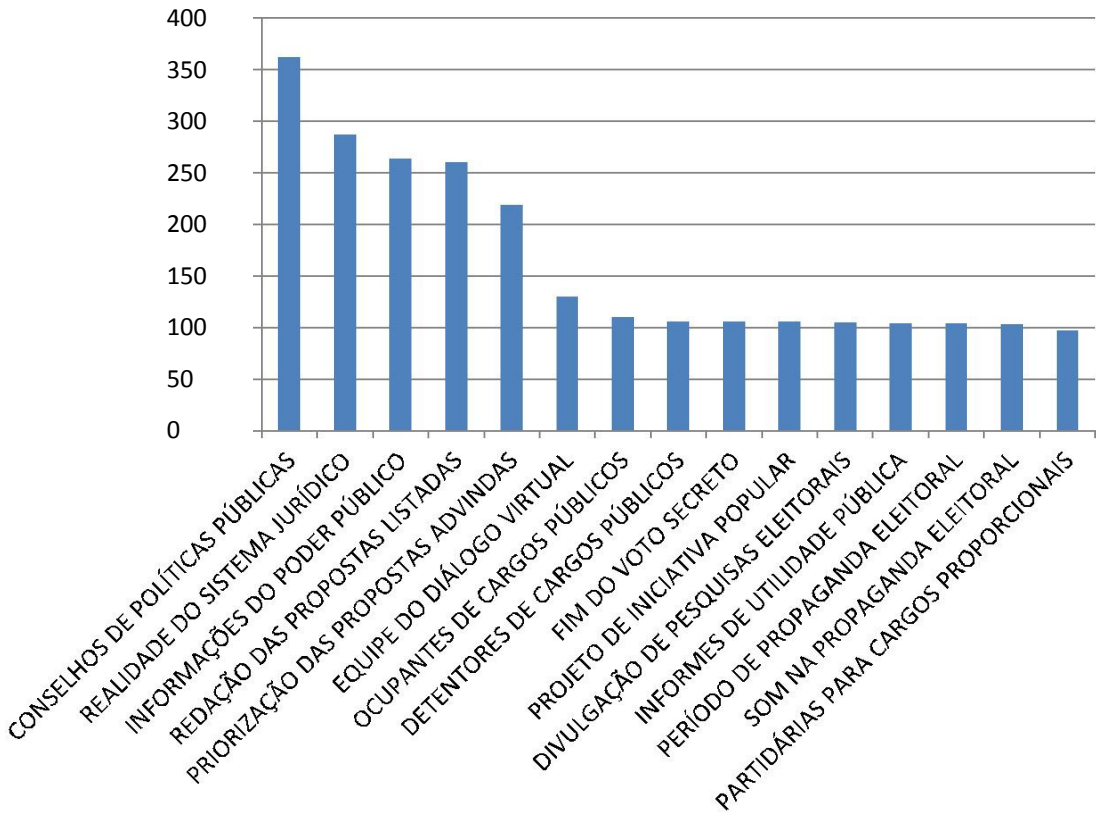

Figura 4 - Frequências(eixo y) para as 15 palavras mais frequentes (eixo x) buscadas pela expressão no corpus de testes.

5 ANÁLISE DE RESULTADOS E CONCLUSÕES

Como especificado na questão de pesquisa, o objetivo dos experimentos realizados foi procurar tópicos de interesse dos grupos de discussão (indivíduos, produtos, eventos, etc.) e verificar se o método descrito aplicado é eficaz ao localizar tais tópicos.

Observa-se no gráfico da figura 1, que as 15 (quinze) palavras simples mais frequentes não identificam claramente tópicos relevantes, considerando que o *corpus* é formado, geralmente, por assuntos da área de economia, política, educação, cultura e sociedade brasileira. A palavra "CAMARA", com mais de 3000 ocorrências, poderia referir-se a "Câmara dos Deputados", mas também poderia referir a "Câmara de Vereadores"; o mesmo acontece com "PROPOSTAS" que pouco identifica algum tópico de discussão (quais propostas?). Outras palavras com alta frequência mostradas no gráfico da figura 1, como "ANOS", "FORMA", "DIA" também precisariam de algum contexto para serem consideradas como tópicos de interesse.

Nas três figuras posteriores (figuras 2, 3 e 4) observam-se identificações mais consistentes dos tópicos de interesse, ainda que existam palavras que necessitam de maior contextualização, como, por exemplo, "CONFERÊNCIA VIRTUAL" no gráfico da figura 2 ou "ETAPA DA CONFERÊNCIA" no gráfico da figura 3 (qual conferência?).

Pode-se concluir que apesar das palavras compostas mais frequentes trazerem uma noção melhor sobre quais tópicos são mais comentados, a necessidade de melhor contextualização existe. Porém, as palavras apontadas, que necessitam de maior contextualização, não são totalmente excluídas como tópicos de interesse dos internautas. Seria possível, por exemplo, aplicar novos filtros nestes elementos e identificar em quais contextos discursivos tais palavras são utilizadas (ou seja, com quais outras palavras). Investigações mais elaboradas podem ser realizadas futuramente.

6 REFERÊNCIAS BIBLIOGRÁFICAS

AIRES , Rachel Virgínia Xavier. Implementação, adaptação, combinação e avaliação de etiquetadores para o português do Brasil. **Dissertação** (Mestrado em Matemática). São Carlos, SP: Universidade de São Paulo, 2000.

BALAGE FILHO, P. P.; PARDO, T. A. S. Busca Opiniões: searching for opinions over the internet. In: WORKSHOP OF SOFTWARE DEMONSTRATIONS, 2014, São Carlos. **Anais...** São Paulo: Universidade de São Paulo, 2014. p. 1-3.

BECKER, Karin; TUMITAN, Diego. Introdução à Mineração de Opiniões: conceitos, aplicações e desafios. Simpósio Brasileiro de Banco de Dados, 2013.

CALLOU, Dinah et al. A Posição do Adjectivo no Sintagma Nominal: duas perspectivas de análise. **Análise Contrastiva de Variedades do Português-Primeiros Estudos**, p. 11-35, 2003.

LOPES, Thomas Jefferson P. et al. Mineração de Opiniões aplicada à Análise de Investimentos. In: BRAZILIAN SYMPOSIUM ON MULTIMEDIA AND THE WEB. ACM, 2008. p. 117-120.

SCOPIM, D. Estudo de Padrões Lexicais em Textos Opinativos. **Dissertação** (Mestrado em Linguística). São Carlos, SP: Universidade Federal de São Carlos, 2011.