Sistema Inteligente de Apoio à Análise Crítica da Atuação do Corpo Parlamentar

Luan Bruno de Souza, Hendrik Teixeira Macedo

Departamento de Computação – Universidade Federal de Sergipe (UFS)
49.100-000 – São Cristóvão – SE – Brasil

luanbrunos@gmail.com, hendrik@ufs.br

Resumo. *Neste trabalho é proposto um método para análise e processamento das notícias divulgadas pelo sítio Web da Assembleia Legislativa do Estado de Sergipe utilizando um processo de mineração de texto e sendo re-aplicável a câmaras de outros estados da federação. O objetivo é facilitar ao eleitor a documentação dessas notícias e fornecer apoio à decisão sobre qual candidato a deputado estadual depositar seu voto. A análise consiste na categorização automática das notícias em rótulos pré-determinados que, alinhado a outros dados, medem a atuação do parlamentar ao longo do seu mandato através da formalização de um índice de avaliação, denominado de QoP (Quality of Parlamentarian). Diferentes técnicas de categorização foram utilizadas e o algoritmo Naive Bayes Multinomial com frequência mínima de corte igual a 0 (zero) obteve a maior taxa de acerto, na ordem de 84%. O índice QoP mostrou-se eficaz para o ranqueamento dinâmico de deputados de acordo com a aferição de sua qualidade em termos quantitativos e qualitativos. Uma aplicação Web foi desenvolvida e permite que eleitores visualizem o ranqueamento padrão, bem como gerar ranqueamentos alternativos e personalizados através de uma interface para modificação de pesos da fórmula de cálculo do índice.*

Palavras-chave: Atuação parlamentar, Assembleia Legislativa, Categorização de textos.

Abstract. *In this work, we propose a method for analyzing and processing textual news from the Legislative Assembly of Sergipe State website, by means of a text mining process which can be applied to legislative chambers from different states. The main goal is to facilitate to the voter the documentation of such news, and provide decision support on which candidate to vote for state representative. The analysis consists mainly in the automatic categorization of news on predetermined labels that, aligned with other data, measure the performance of parliament during its mandate by generating an index, the QoP (Quality of Parlamentarian). Different techniques have been used to categorization and Multinomial Naive Bayes algorithm with a minimum cutoff frequency equal to 0 (zero) has obtained the highest success rate, in the order of 84%. The QoP index was effective for the dynamic ranking of deputies according to the assessment of their quality in both quantitative and qualitative terms. A Web application was developed and allows the voters to view the default ranking as well as generate alternative rankings through a custom interface where they can modify weights used to calculate the index.*

Key-words: Parliamentary action, Legislative Assembly, Text Categorization.

1. Introdução

A democracia brasileira permite que o eleitor escolha através do voto direto seus representantes nos parlamentos federal, estadual e municipal. Campanhas são feitas em época eleitoral para a conscientização da população de que o voto deve ser algo consciente e invendável. Sem informação apropriada, entretanto, é dificultoso para o eleitor selecionar um dentre vários candidatos disponíveis.

Aproveitando a facilidade de disseminação de informação da Web, assembleias legislativas estaduais divulgam informações a respeito da postura parlamentar. Uma dessas informações são as notícias diárias divulgadas no sítio Web da instituição. Uma equipe técnica interna diariamente edita e publica notícias referentes à atividade dos parlamentares. De posse de tais informações, o eleitor estaria mais apto a selecionar o candidato que mais lhe agradasse em termos de atuação, postura política, opiniões e ideias.

O volume de dados disponibilizado pelas assembleias, entretanto, é consideravelmente alto. Tomando-se como exemplo a *Agência de Notícias da Assembleia Legislativa do Estado de Sergipe* (ALESE), verifica-se a divulgação de uma média de 115 notícias mensais, totalizando mais de 5.500 notícias ao longo de um mandato de quatro anos. Tal volume de dados pode se tornar de difícil interpretação para um eleitor, mesmo que seja feito de forma gradativa.

Um outro problema está relacionado ao tipo de dado. Dados textuais carecem de um esforço de processamento consideravelmente maior em virtude da dificuldade inerente que existe em se interpretar adequadamente o significado (semântica) do texto em questão.

Para solução do problema, o presente trabalho visa prover uma estrutura para a análise desse volume de dados textuais, facilitando ao eleitor a documentação dessas notícias e, em seguir, fornecer apoio à decisão para escolha do melhor candidato a deputado estadual além de permitir verificar aqueles que não estariam aptos a uma possível reeleição. A abordagem consiste de (1) etapas de pré-processamento das notícias publicadas pela ALESE, alinhado a uma abordagem estatística para *Mineração de Textos* [LEHAL e GUPTA 2009] responsável pela representação de um texto como *vetor de atributos (pesos)*, (2) uma etapa de categorização de dados aplicada sobre este vetor [RIZZI et al. 2000], e (3) a elaboração de um índice para aferição da qualidade de um parlamentar a partir do quantitativo e qualitativo de sua atuação, obtidos na etapa anterior.

Categorização automática de textos é um objeto de estudo e investigação científica recorrente atualmente, claramente impulsionado pelo aumento significativo do volume de dados textuais disponíveis na Internet. Assume-se a existência de um rótulo (ou classe), pertencente a um conjunto de rótulos pré-definidos, para qualquer texto existente do domínio em questão. Baseado em um conjunto de treinamento, onde textos com categorização conhecida estão manualmente rotulados, busca-se descobrir de forma automática o rótulo mais apropriado para um novo texto fornecido e ainda não devidamente categorizado. Em [PAK 2010], o autor aborda o problema de categorização automática de mensagens eletrônicas recebidas no Centro de Tratamento de Incidentes de Segurança em Redes de Computadores da Administração Pública Federal (CTIR Gov). Ele afirma que relatos de incidentes reportados têm crescido a um ritmo preocupante, decorrendo uma necessidade por formas de tratar grandes volumes de notificações de incidentes. [CHAN et al. 2001] construíram um classificador utilizando um mecanismo automatizado de classificação personalizada de notícias online através de *SVM* (*Support Vector Machine*). Na classificação personalizada, o usuário pode definir suas próprias categorias utilizando algumas palavras-chave e fica a

cargo do próprio classificador obter o conjunto de treinamento. Esta liberdade dada ao usuário de criar suas próprias classes é aplicável principalmente a soluções mais genéricas, como é o caso do trabalho de CHAN. [STEINBRUCH 2006] propõe dois algoritmos de classificação automática de textos baseados no algoritmo multinomial *Naïve Bayes* e sua utilização em um ambiente *on-line* de classificação automática de textos com realimentação de relevância pelo usuário. O autor aponta a classificação automática de textos como solução para o problema de organização e aquisição de informação.

O trabalho descrito acima se assemelha com o presente trabalho proposto. Entretanto, por se tratar de uma solução genérica, a solução de STEINBRUCH não aborda quesitos específicos do problema em questão, como por exemplo, a associação das notícias a um determinado parlamentar.

O restante do artigo está organizado como se segue. A seção 2 descreve o sítio Web da ALESE, como se dá o pré-processamento das notícias divulgadas e a confecção do classificador. Na seção 3 são apresentados e discutidos os índices propostos para representação numérica da atividade parlamentar. Experimentos realizados para cada uma das etapas são apresentados e os resultados discutidos na seção 4. A seção 4 traz ainda o sistema Web desenvolvido segundo a abordagem proposta neste trabalho. Por fim, a seção 5 traz conclusões do trabalho.

2. Categorização Automática de Notícias

A ALESE é formada por uma equipe interna que diariamente divulga no sítio Web da assembleia notícias referentes à atividade parlamentar. O sítio também disponibiliza informações como contratos, licitações, processos, informações sobre os parlamentares, bancadas e outros. São disponibilizadas em torno de 4 notícias diárias, totalizando mais de 5.500 notícias durante um mandato de 04 (quatro) anos. Em Assembleias Legislativas de outros estados, essa quantidade chega a ser bem maior. Em levantamento feito entre os dias 05 e 09 de março de 2012, em 12 Assembleias Legislativas, obteve-se uma média de 14 notícias diárias, ou 15.400 notícias em um mandato de 04 anos. A estrutura de análise a ser apresentada é aplicável também a esses outros sítios.

Para devida extração das notícias do sítio da ALESE, um *focused crawler* [KONCHADY 2006] implementado. No nível 0 (zero) é determinada a URL base do sítio da ALESE. O último link de cada nível é a página de notícias mais antigas. A partir deste, o *crawler* busca o conteúdo das notícias. Esta estrutura, assim como a URL base, é correspondente à estrutura do site da ALESE e poderá sofrer alterações de acordo com o site da assembleia de cada estado. Após a notícia ser encontrada, ela é pré-processada e categorizada apropriadamente em um dos rótulos a serem descritos.

2.1 Pré-processamento das notícias

As notícias que são extraídas pelo crawler passam por duas etapas de pré-processamento importantes. Primeiramente, palavras que não agregam significado relevante no texto são removidos (eliminação de *stopwords*). Por fim, as palavras restantes selecionadas são reduzidas ao seu radical (atividade de *stemming*).

O vetor de palavras resultantes relativo a cada notícia original é então indexado em um *vetor de atributos*, onde cada atributo é representado por uma palavra e sua respectiva ocorrência no texto. Para evitar que palavras comuns a todos os textos sejam consideradas, foi utilizado o método de indexação *Term Frequency Inverse Document Frequency* (TF-IDF). Este método reduz o peso de termos que são populares em todos os textos do corpus considerado, ainda que a quantidade de ocorrências deste termo seja alta em um texto qualquer. A fundamentação lógica é que termos que são populares em

toda a coleção possuem baixo poder para caracterizar unicamente um determinado texto desta coleção. Termos como "deputado", "câmara" e "Sergipe", por exemplo, possuem pouca relevância, apesar de ocorrerem com grande frequência.

Uma vez categorizado, o texto passa por um reconhecimento da autoria, para determinar qual o parlamentar responsável pela ação descrita na notícia. Este reconhecimento é feito através do uso de expressões regulares (*regex*) aplicadas ao título da notícia. Em um teste feito com 365 títulos de notícias, o reconhecedor descrito obteve uma porcentagem de acerto de 98,63%.

2.2 Treinamento do classificador

As notícias do sítio Web da ALESE foram categorizadas em quatro rótulos (classes). A escolha dos rótulos foi tomada com base na análise das notícias da própria ALESE, tendo como preocupação uma clara delimitação dos tipos de atividade:

LAD (Lançamentos, aprovações e defesas): *descreve a participação ativa do parlamentar na construção e defesa de projetos. Inclui lançamento de ideias, sugestões, pedidos, aprovações e apoios a projetos de autoria própria ou de outros parlamentares.*

ACP (Acusações, críticas e protestos): *descreve a participação do parlamentar na forma de acusações. Inclui críticas e protestos a projetos, comportamento de terceiros, ideias e fatos.*

PNC (Participação na comunidade): *descreve as ações populares do parlamentar. Inclui a promoção, participação ou simples apoio em inaugurações, homenagens, dias especiais e outros eventos.*

NEU (Neutro): *notícia que não descreve uma ação parlamentar, ou descreve uma ação parlamentar coletiva. Não será considerada no cálculo do índice.*

Um conjunto de treinamento composto por 365 notícias foi criado e essas foram manualmente rotuladas. A distribuição do conjunto para as classes LAD, ACP, PNC e NEU foram, respectivamente, 119, 93, 109 e 44 notícias.

A verificação da qualidade do classificador foi feita utilizando-se *validação cruzada* com a base sendo dividida em *n folds*, sendo *n* um número especificado. Cada *fold* é utilizado como conjunto de teste enquanto os outros *n-1 folds* são utilizados para construção do classificador. O processo é repetido *n* vezes e a precisão do classificador para cada classe juntamente com a *matriz de confusão* são obtidos [WITTEN, FRANK e HALL 2011].

3. Índice para Medição de Qualidade do Parlamentar

É importante destacar que cada notícia pode possuir um peso diferente a depender do eleitor. Se um parlamentar, por exemplo, proferir um discurso na assembleia defendendo um projeto de reforma agrária, a notícia gerada por este discurso terá uma importância diferenciada para um trabalhador rural. Com isso, como aditivo ao processo de classificação, o eleitor poderá determinar para cada notícia um peso especial, um valor que dê uma maior relevância àquela notícia e, consequentemente, ao parlamentar autor.

> **Definição 1.** *Relevância (λ)* é associada ao valor mensurado pelo usuário para representar numericamente a importância de uma notícia, onde $\lambda \in [0,5]$.

A média das relevâncias ($\overline{\lambda}$) sempre será vista levando em consideração um parlamentar autor específico e uma classe de notícias específica. Sendo assim, tendo disponível dados como quantidade de notícias, classe e relevância de cada notícia, será calculado um índice, denominado *QoP (Quality of Parliamentarian)*.

Definição 2. *Quality of Parliamentarian (QoP)* é o índice que expressa numericamente o nível de atuação parlamentar de cada deputado, e é dado por:

$QoP_c = \frac{20 * \bar{\lambda} * q_c}{qm_c}$, onde c é o índice que indica a qual classe o QoP está relacionado (1, 2 ou 3), q_c é a quantidade de notícias de um determinado deputado na classe em questão, qm_c é a quantidade do deputado que mais publicou na classe em questão e $\bar{\lambda}$ representa a média das relevâncias para o mesmo deputado na mesma classe.

O índice *QoP* pode ser utilizado para gerar um sistema de ranqueamento, onde parlamentares são ordenados de acordo com sua atividade exercida e perfil do eleitor. Este, por sua vez, poderá demonstrar preferência por um determinado tipo de atividade parlamentar, representado pelas classes rotuladas. O *QoP* seguirá duas vertentes, a saber: (1) *Personal QoP* e (2) *General QoP*.

O *Personal QoP* refere-se aos dados de cada usuário e remete a opinião pessoal do eleitor, ou seja, para cada deputado, os valores podem diferir. Cada usuário definirá pesos às classes (1-LAD, 2-ACP e 3-PNC). Esses pesos determinam a preferência do mesmo em relação aos tipos de atividades parlamentares, de modo que:

$$p_{c=1} + p_{c=2} + p_{c=3} = 1$$

Definição 3. *Personal Quality of Parliamentarian (p-QoP)* é o índice *QoP* calculado a partir dos dados de um determinado eleitor, e calculado como $p\text{-}QoP = p\text{-}QoP_1 * p_{c1} + p\text{-}QoP_2 * p_{c2} + p\text{-}QoP_3 * p_{c3} \in [0,100]$,

onde $p\text{-}QoP_c = \frac{20 * \bar{\lambda} * q_c}{q_{mc}}$ e c é o índice que indica a qual classe o p-QoP está relacionado (1, 2 ou 3). Na média das relevâncias ($\bar{\lambda}$), são levadas em consideração apenas as relevâncias definidas pelo eleitor, para o deputado em questão e classe em questão. Os quantitativos q_c e qm_c são, obviamente, comum a todos os usuários.

O *General QoP* refere-se ao agrupamento dos dados de relevâncias definidos por todos os usuários, ou seja, o *g-QoP* será único para todos os eleitores. Este índice remete a opinião geral dos eleitores.

Definição 4. *General Quality of Parliamentarian (g-QoP)* é o índice *QoP* calculado a partir dos dados de todos os eleitores usuários e calculado como

$$g\text{-}QoP = \frac{3}{\frac{1}{gQoP_{C1}+1} + \frac{1}{gQoP_{C2}+1} + \frac{1}{gQoP_{C3}+1}} - 1 \in [0,100],$$

onde $g\text{-}QoP_c = \frac{20 * \bar{\lambda} * q_c}{q_{mc}}$ e c é o índice que indica a qual classe o p-QoP está relacionado (1, 2 ou 3). Na média das relevâncias ($\bar{\lambda}$), é levada em consideração as relevâncias definidas por todos os eleitores, para o deputado em questão e classe em questão. É adicionado 1 (uma) unidade à cada $q\text{-}QoP_c$ para evitar que o *g-QoP* assuma valor 0 (zero) se apenas um dos $g\text{-}QoP_c$ for igual à 0 (zero).

A média harmônica possui a característica de penalizar valores discrepantes, ou seja, o candidato que obteve valores de $g\text{-}QoP_c$ com menor variação possuirão vantagem em relação ao candidato que obteve $g\text{-}QoP_c$ com valores de maior variação.

A partir das definições de *p-QoP* e o *g-QoP*, é possível extrair as seguintes informações:

i. O *QoP* busca o equilíbrio qualitativo e quantitativo. Qualitativo, porque leva em consideração a média das relevâncias e médias (ponderada e harmônica) dos *QoPs* das classes de uma forma geral. Quantitativo, porque leva em consideração a quantidade de notícias publicadas. Um fator poderá penalizar ou não o outro.

ii. A divisão $\frac{q_c}{q_{mc}}$ retorna valores entre 0 (zero) e 1 (hum). Na melhor das hipóteses, para uma determinada classe, o candidato foi aquele que publicou mais notícias. Neste caso, a divisão retornará o valor 1 (hum), e seu *QoP* será determinado pela média das relevâncias. Considerando que as relevâncias atingem valores inteiros entre 0 (zero) e 5 (cinco), a fórmula $(20 * \bar{x})$ atingirá valores reais entre 0 (zero) e 100 (cem).

iii. Em um cenário ideal, um parlamentar terá seu *QoP* (Personal ou General) igual à 100 (cem) (valor máximo) se:

a. o mesmo foi aquele que mais publicou notícias em todas as classes, e

b. todas as notícias (pessoais para *p-QoP* ou totais para o *g-QoP*) tiveram suas relevâncias definidas como 5 (cinco).

iv. O *QoP* também pode ser visto como um índice parcial: a escala (de 0 a 100) não muda com o passar do tempo. Ou seja, a avaliação dos índices dos deputados poderá ser feita de forma gradativa, numa análise com frequência periódica.

4. Resultados e discussão

O trabalho produziu resultados em três diferentes linhas: (1) confecção de um mecanismo para categorização automática de notícias em formato textual natural sobre atuação parlamentar disponibilizadas em sítios Web, (2) formalização de um índice para avaliação da qualidade de atuação de um parlamentar e (3) desenvolvimento de uma aplicação Web que permite o ranqueamento dinâmico de deputados segundo seu índice de atuação.

4.1 Categorização automática dos textos da ALESE

Na confecção do módulo de categorização, diferentes algoritmos de classificação foram considerados e a frequência de corte, denominada *frequência mínima de termo*, também sofreu variações. Os algoritmos utilizados foram *Naive Bayes, Árvore de decisão e IBk*.

Para seleção do algoritmo *Naive Bayes*, foram executados testes com as variações *Naive Bayes*, *Naive Bayes Multinominal*, *Naive Bayes Updateable* e *Naive Bayes Multinominal Updateable*. Sem o uso da frequência de corte, os percentuais de rótulos classificados corretamente foram 72,6%, 83,56%, 72,6%, 83,56%, respectivamente. O algoritmo *Naive Bayes Multinominal* obteve as melhores taxas de classificação e erro médio, além de menor tempo de processamento.

Os algoritmos de árvore de decisão utilizados foram o *Random Forest, C4.5 e REPTree*, obtendo os seguintes percentuais de rótulos classificados corretamente: 63,01%, 63,83% e 64,66%. Foi possível perceber que, dentre os algoritmos de árvore de decisão, o REPTree obteve a melhor taxa de categorização.

O algoritmo IBk foi testado com diferentes valores para o número de vizinhos mais próximos (campo k). O algoritmo IBk com k=1 obteve o melhor percentual de classificação correta, na ordem de 40,55%.

Para cada tipo de algoritmo, foi selecionado aquele com melhor desempenho e então implementamos variação na frequência de corte. Os resultados podem ser

visualizados no gráfico da figura 1, que relaciona a frequência mínima de corte com porcentagem de categorização correta. Observe que o algoritmo *Naive Bayes Multinominal* com frequência mínima de corte igual a 0 (zero) obteve a melhor porcentagem de categorização.

A tabela 1 traz a *matriz de confusão* obtida para os rótulos LAD, ACP, PNC e NEU, especificados para as notícias da ALESE. A partir dessa matriz é possível observar que:

i. Há um alto índice de confusão entre as classes a e b. No contexto da natureza das notícias, no âmbito político, essa confusão pode ser explicada considerando que ao realizar uma crítica, um deputado usualmente sugere uma mudança logo a seguir, lançando uma nova ideia. Nesses casos, é tolerável que ocorra a categorização do texto como quaisquer umas das classes a e b.

ii. Há também um alto índice de confusão entre as classes a e c. No âmbito político, essa confusão também possui uma fundamentação lógica. Em alguns casos, um deputado visita um determinado local onde funciona um projeto empenhado pelo próprio, por exemplo. Nesses casos, é tolerável que ele classifique o texto como quaisquer umas das classes a e c.

iii. Apesar de não ser aceitável, salvo em casos bastante específicos, que o classificador classifique como a, b ou c textos que estão rotulados com a classe d – NEU (Neutro), essas falhas não terão influência na qualificação de um deputado, pois textos por natureza neutros não possuem a autoria de um deputado específico, e mesmo sendo classificado como a, b ou c, este será descartado.

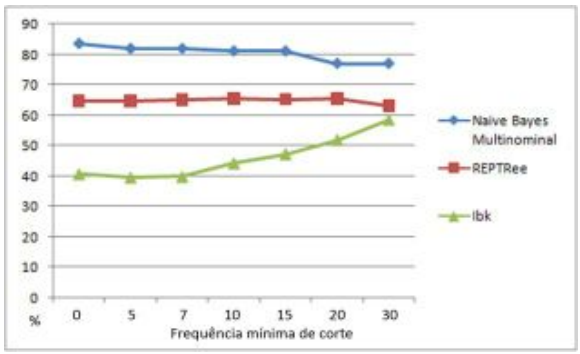

Figura 1: Frequência mínima de corte vs. Porcentagem correta de categorização.

Tabela 1: Matriz de confusão entre rótulos

	a: LAD	b: ACP	c: PNC	d: NEU
a: LAD	**95**	13	6	5
b: ACP	6	**85**	0	2
c: PNC	10	5	**91**	3
d: NEU	4	1	5	**34**

4.2 QoP – Quality of Parliamentarian

Para os testes de validação do índice QoP, foi implementado um sistema com as funcionalidades básicas descritas pelo trabalho em voga, sendo que 460 notícias divulgadas entre os dias 12 de dezembro de 2011 e 08 de julho de 2012 foram carregadas. As notícias foram classificadas e seus respectivos autores, determinados. Após esse processo, pode-se perceber a discrepância entre as quantidades de notícias divulgadas para cada deputado. Nesse período, as quantidades totais de cada deputado oscilaram entre 0 e 58 notícias, o que por certo afeta seus índices.

À medida que as notícias eram carregadas pelo *crawler*, um grupo de voluntários colaboradores realizavam a leitura de tais notícias e definiam para cada uma delas um valor de relavância. Ao todo, 1124 relevâncias foram definidas pelo grupo de voluntários, distribuídas entre as 460 notícias. A média da relevância dentre todas as notícias foi igual a 2,60. As dez primeiras posições do g-QoP resultante dos dados obtidos é representado na tabela 2.

Tabela 2 : G-QoP experimental

Pos	Deputado	QoP LAD	QoP ACP	QoP PNC	g-QoP
1	Au... B...	13.78	47.96	12.04	**17.2**
2	J... D...	37.23	3.67	46.67	**10.49**
3	C... S...	9.33	32.47	3.89	**8.06**
4	A... S.	70.86	3.13	11.56	**7.94**
5	M... M...	38.02	2.72	17.15	**7.58**
6	G... M...	16.0	2.92	8.59	**6.17**
7	V... F...	2.67	37.16	4.44	**5.22**
8	A... G...	4.67	2.57	14.96	**4.77**
9	Go... R...	3.11	3.21	6.42	**3.88**
10	C... V...	8.0	1.36	9.43	**3.76**

Um exemplo claro da implicação do uso da média harmônica dos g-QoP$_c$s para o cálculo do g-QoP é visto nas duas primeiras posições do ranking ilustrado na tabela 2. O deputado Au... B... possui g-QoP$_1$, g-QoP$_2$ e g-QoP$_3$ respectivamente iguais à 13.78, 47.96 e 12.04, com média harmônica igual a 17.2. Para o deputado J... D..., por sua vez, estes valores são respectivamente iguais à 37.23, 3.67 e 46.67 com média harmônica de 10.49. O uso da média aritmética ocasionaria uma inversão de posição entre estes dois deputados.

Para análise do p-QoP, tomemos como base um dos usuários do sistema. Como descrito anteriormente, cada usuário define pesos para cada classe, e o p-QoP é obtido calculando-se a média ponderada dos p-QoP$_c$s. Para uma melhor análise, assumamos três grupos de pesos distintos para o mesmo usuário. Chamaremos de *p1* o peso associado à classe 01-LAD (Lançamentos, aprovações e defesas), *p2* o peso associado à classe 02-ACP (Acusações, críticas e protestos) e *p3* o peso associado à classe 03-PNC (Participação na comunidade).

Para o grupo 1, assumamos **p1 = 0.5**, **p2 = 0.3**, e **p3 = 0.2**. A tabela 3 apresenta as cinco primeiras posições do resultado da média ponderada dos p-QoP$_c$'s para o grupo 1.

Tabela 3 : P-QoP experimental para o grupo 1

Pos	Deputado	QoP LAD	QoP ACP	QoP PNC	p-QoP
1	A... S...	80.0	0.0	4.44	**40.89**
2	Au... B...	16.0	38.75	14.81	**22.59**
3	J... D...	26.4	2.45	34.29	**20.79**
4	S... A...	40.0	1.22	0.74	**20.52**
5	M... M...	26.67	0.0	14.81	**16.3**

Para o grupo 2, assumamos **p1 = 0.2**, **p2 = 0.5**, e **p3 = 0.3**. A tabela 4 apresenta o resultado da média ponderada dos p-QoP$_c$'s para o grupo 2:

Tabela 4 : P-QoP experimental para o grupo 2

Pos	Deputado	QoP LAD	QoP ACP	QoP PNC	p-QoP
1	Au... B...	16.0	38.75	14.81	**27.02**
2	V... F...	0.0	39.11	0.0	**19.55**
3	A... S...	80.0	0.0	4.44	**17.33**
4	J... D...	26.4	2.45	34.29	**16.79**
5	C... S...	0.0	24.76	2.22	**13.05**

Para o grupo 3, assumamos **p1 = 0.2**, **p2 = 0.3**, e **p3 = 0.5**. A tabela 5 apresenta o resultado da média ponderada dos p-QoP$_c$'s para o grupo 3:

Tabela 5: P-QoP experimental para o grupo 3

Pos	Deputado	QoP LAD	QoP ACP	QoP PNC	p-QoP
1	J... D...	26.4	2.45	34.29	**23.16**
2	Au... B...	16.0	38.75	14.81	**22.23**
3	A... S...	80.0	0.0	4.44	**18.22**
4	M... M...	26.67	0.0	14.81	**12.74**
5	V... F...	0.0	39.11	0.0	**11.73**

É possível perceber que a atribuição de pesos às classes possibilita um ranqueamento diferenciado. Para cada grupo de pesos atribuído, obtem-se uma diferente ordem de ranqueamento. O parlamentar A... S... foi o que atingiu um melhor p-QoP no grupo pesos 1, onde deu-se uma maior preferência à classe 01-LAD. Para o grupo de pesos 2, onde deu-se maior preferência para a classe 02-ACP, o parlamentar Au... B... foi o que obteve um melhor índice. O parlamentar J... D... foi o que atingiu um melhor p-QoP no grupo de pesos 3, onde deu-se uma maior preferência à classe 03-PNC. O cálculo da média ponderada mostrou-se, portanto, eficiente na representação da preferência do usuário em relação a um tipo de atividade parlamentar.

4.3 Sistema ZeroUm

Baseado na estrutura descrita no presente trabalho, foi desenvolvido um sistema Web de gerenciamento das notícias publicadas pela ALESE, denominado *ZeroUm*. O sistema desenvolvido importa as notícias da ALESE através do *crawler*, classifica-as, determina os respectivos parlamentares envolvidos e armazena as notícias em um banco de dados. Cada usuário do sistema pode, então, visualizar a notícia e seu conteúdo e determinar sua relevância. O ZeroUm calcula os QoP's e disponibiliza o sistema de ranqueamento dos parlamentares.

O sistema foi desenvolvido seguindo o padrão MVC (Model View Controller), que divide as classes em três camadas distintas: interface, negócio e dados. Além dessas três camadas, o ZeroUm possui também a camada de integração, que contém o *crawler* que realiza a interface do sistema com o site da ALESE. O ZeroUm possui, ao todo, 23 classes e 7 arquivos de interface e seu banco de dados possui 5 tabelas. A página principal do ZeroUm é mostrada na figura 2.

Na página inicial, o usuário pode ter acesso às notícias que ainda não foram visualizadas, além dos menus de acesso às notícias que já foram lidas, visualização dos QoP's, alteração de dados pessoais e obtenção de informação a respeito do processo envolvido. A figura 3 descreve a interface do sistema quando uma notícia é selecionada para visualização.

Na visualização de uma notícia, o sistema exibe seu título, autor, data, classe, relevância e conteúdo da notícia, além da média de relevâncias que outros usuários atribuíram à mesma. Ainda nesta página, o usuário deverá definir a relevância da notícia em questão. Como mencionado anteriormente, o sistema calcula os QoP's e exibe-os na tela ilustrada na figura 4.

Nesta mesma tela, é possível ao usuário visualizar o p-QoP e o g-QoP. Os parlamentares são, por padrão, ordenados pelos índices finais. Entretanto, o usuário pode, para cada QoP, ordenar os parlamentares de acordo com o QoP de uma determinada classe.

Figura 2 : Página inicial do ZeroUm

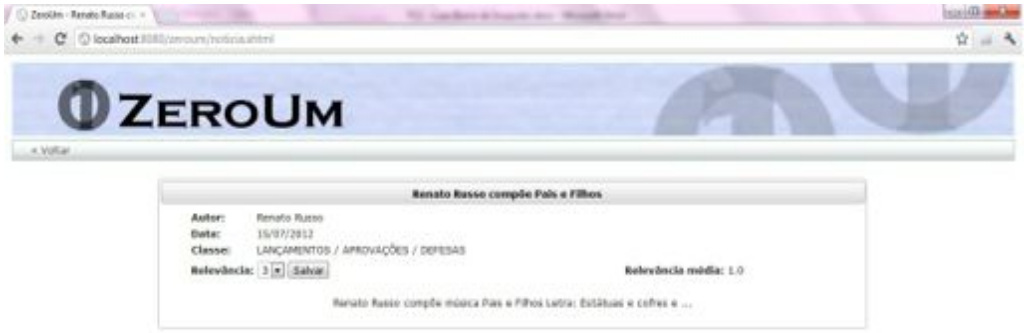

Figura 3 : Visualização de uma notícia

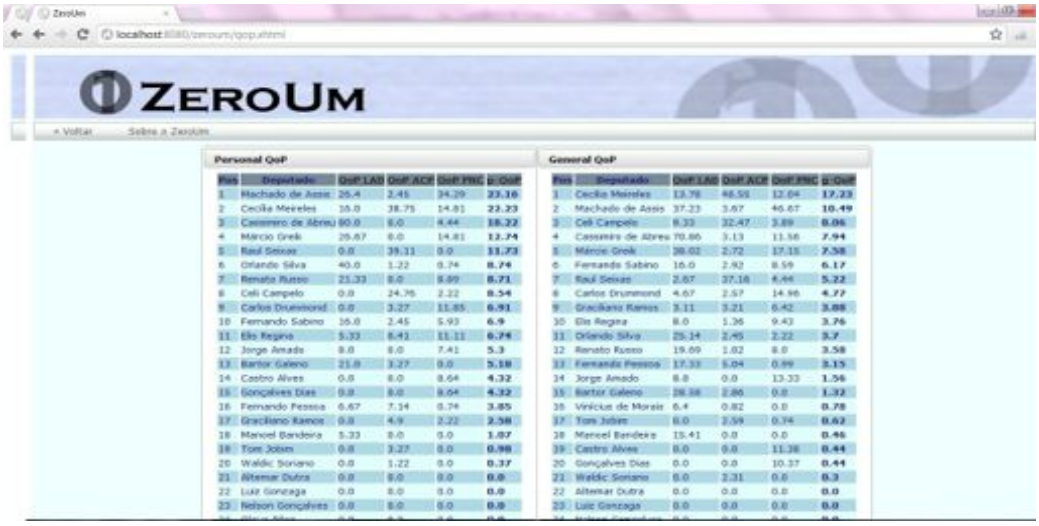

Figura 4 : Tela de QoPs

5. Conclusões

Neste artigo é proposta uma abordagem para aferição da qualidade de um parlamentar, em termos quantitativos e qualitativos, baseada em informações automaticamente extraídas de notícias de sítios Web oficiais de assembleias legislativas.

São três as linhas de contribuição do trabalho: (1) confecção de um mecanismo para categorização automática de notícias em rótulos pré-definidos, (2) formalização de um índice para avaliação da qualidade de atuação de um parlamentar e (3)

desenvolvimento de uma aplicação Web que permite o ranqueamento dinâmico de deputados segundo seu índice de atuação.

Resultados da avaliação do classificador mostraram ser possível a categorização automática de notícias sobre a atuação de deputados estaduais com um percentual de acerto em torno de 84%.

O sistema Web desenvolvido surge com o intuito de fortalecer as ferramentas de e-gov, que são responsáveis pela disseminação dos conteúdos governamentais para a população, mais especificamente, uma maior divulgação das atividades parlamentares. O tratamento das notícias das Assembleias Legislativas sugeridas neste trabalho deverá melhorar a fiscalização dos parlamentares em exercício por parte da população, além de ajudar na seleção de um parlamentar na hora do pleito e, consequentemente, a melhora na qualidade dos parlamentares eleitos.

Uma das limitações do trabalho é consequência de uma característica inerente aos sítios Web: de fato, não cabe às assembleias legislativas conceder notícias sobre prováveis candidatos na eleição seguinte. As notícias estão, obviamente, restritas aos parlamentares em exercício. O eleitor deverá então considerar o fato de que há outros candidatos pleiteando uma vaga na assembleia. Uma possível extensão é considerar o cruzamento de informação com outras fontes de dados. Deste modo, não só informações sobre novos candidatos estariam disponíveis, mas também outras informações sobre os parlamentares atuais seriam adicionadas as notícias da assembleia. O problema dessa solução é escolher tais fontes de dados, já que não podemos garantir que toda e qualquer mídia seja totalmente imparcial.

Uma outra limitação do trabalho é que algumas notícias possuem características de mais de um rótulo. Nesses casos, é tolerável que o classificador categorize o texto como quaisquer umas das classes aceitas. Como solução a esse problema, sugere-se uma classificação *multi-classe*, que daria ao classificador a liberdade para rotular um texto com mais de um rótulo, se assim for conveniente.

Como adicional ao processo definido nesse trabalho, o usuário poderá fazer uso de outras ferramentas para apoio à decisão do voto. Um delas é o sítio Web *Repolítica* [REPOLITICA]. Enquanto o presente trabalho busca a coleta de informações de maneira gradativa, o Repolítica recomenda um candidato com base na coleta de opiniões pessoais, e nas opiniões pessoais de outros usuários sobre o perfil geral que o eleitor espera de um parlamentar.

Finalmente, é importante ressaltar que o presente trabalho não se restringe à Assembleia Legislativa do Estado de Sergipe. Todas as assembleias legislativas estaduais brasileiras possuem sítios e divulgam diariamente suas movimentações. Da mesma maneira, as notícias não são a única fonte de informação presente nesses sítios, que contêm também informações sobre trâmites de processos, informações sobre os parlamentares, pautas diárias, entre outras. Dessa forma, outras técnicas de mineração de dados e de texto poderão ser integradas à técnica apresentada, ampliando e facilitando o assimilamento das informações disponíveis.

Referências

CHAN, Chee-Hong; SUN, Aixin; LIM, Ee-Peng. Automated Online News Classification with Personalization. *4th International Conference of Asian Digital Library*, Singapura, 2001.

KONCHADY, Manu. Text Mining Application Programming. Boston: Ed. Thomson, 2006. 412 p.

LEHAL, Gurppreet S. GUPTA, Vishal. A Survey of Text Mining Techniques and Applications. *Journal of Emerging Technologies in Web Intelligence*, Vol 1, No 1, 60-76, 2009.

PAK, Albert Frederico de Menezes Il. Aplicação de Técnicas de Mineração de Texto para Categorização de Eventos de Segurança no CTIR Gov. 2010. Xi, 71 f. *Dissertação (Mestrado em Informática) – Universidade Federal de Brasília*, Brasília, 2010.

RIZZI, Claudia Brandelero (2000). et al. Fazendo uso da categorização de textos em atividades empresariais. Disponível em: <http://leandro.wives.nom.br/pt-br/publicacoes/iskmdm2000-2.pdf>. Acesso em: 29 mar. 2011.

OLIVEIRA, C. J. S. & ARAÚJO, A. de A. Classificando Imagens Coletadas na World Wide Web em duas Classes Semânticas: Imagens Gráficas e Imagens Fotográficas, Relatório Técnico RT.DCC.006/2002, DCC/ICEx/UFMG, Belo Horizonte, MG, Brasil, Dezembro, 2002, 29p.

STEINBRUCH, David. Um estudo de Algoritmos para Classificação Automática de Textos Utilizando naive-Bayes. *Pontifícia Universidade Católica,* Rio de Janeiro, RJ, 2006.

TOMAZELA, Maria das Graças J. M.; DANIEL, Luiz Antônio. Uma Estratégia de Preparação de Dados para Aumento de Precisão de Modelos de Classificação da Produtividade de Cana-de-açucar. *Faculdade de Tecnologia de Indaiatuba,* Indaiatuba, SP, 2011.

WITTEN, Ian H.; FRANK, Eibe; HALL, Mark A. Data Mining: Practical Machine Learning Tools and Techniques. 3rd edition. Hamilton, New Zealand: Ed. Morgan Kaupmann, 2011.