UNIVERSIDADE FEDERAL DA BAHIA

INSTITUTO DE MATEMÁTICA

DEPARTAMENTO DE CIÊNCIA DA COMPUTAÇÃO

Aline Duarte Bessa

PROVISORIO: Um estudo sobre *Opinion Mining*

PROVISORIO: Aspectos teóricos e práticos

Salvador

2010

Aline Duarte Bessa

PROVISORIO: Um estudo sobre *Opinion Mining*

Monografia apresentada ao Curso de graduação em Ciência da Computação, Departamento de Ciência da Computação, Instituto de Matemática, Universidade Federal da Bahia, como requisito parcial para obtenção do grau de Bacharel em Ciência da Computação.

Orientador: Alexandre Tachard Passos
Co-orientador: Luciano Porto Barreto

Salvador

2010

RESUMO

Nonono nonono nonono, nonono, nonono nonono nonono nononono nonno. Nonono nonono nonono, nonono, nonono nonono nonono nononono nonno. Nonono nonono nonono, nonono, nonono nonono nonono nononono nonno. Nonono nonono nonono, nonono, nonono nonono nonono nononono nonno. Nonono nonono nonono, nonono, nonono nonono nonono nononono nonno. Nonono nonono nonono, nonono, nonono nonono nonono nononono nonno. Nonono nonono nonono, nonono, nonono nonono nonono nononono nonno.

Palavras-chave: monografia, graduação, projeto final.

ABSTRACT

Nonono nonono nonono, nonono, nonono nonono nonono nononono nonno. Nonono nonono nonono, nonono, nonono nonono nonono nononono nonno. Nonono nonono nonono, nonono, nonono nonono nonono nononono nonno. Nonono nonono nonono, nonono, nonono nonono nonono nononono nonno. Nonono nonono nonono, nonono, nonono nonono nonono nononono nonno. Nonono nonono nonono, nonono, nonono nonono nonono nononono nonno. Nonono nonono nonono, nonono, nonono nonono nonono nononono nonno.

Keywords: monograph, graduation, final project.

LISTA DE FIGURAS

LISTA DE ABREVIATURAS E SIGLAS

SUMÁRIO

1 INTRODUÇÃO

1.1 MOTIVAÇÃO

A busca por opiniões sempre desempenhou um papel importante na geração de novas escolhas. Antes de optar por assistir a um filme, é comum ler críticas a seu respeito ou considerar os comentários de outras pessoas; antes de comprar um produto, muitas vezes procuramos relatos sobre a satisfação de outros consumidores. Com a disseminação da Web e da Internet, a geração de opiniões com impacto, sobre os mais diversos assuntos, foi finalmente democratizada: não é mais preciso, por exemplo, ser um especialista em Economia ou Ciência Política para manter um blog **deveria definir blog?** convincente sobre algum candidato às eleições.

Neste contexto, a busca por opiniões e comentários em sites, blogs, fóruns e redes sociais também se popularizou, passando a fazer parte do cotidiano dos consumidores online. Uma pesquisa feita nos Estados Unidos revela que entre 73% e 87% dos leitores de resenhas de serviços online, como críticas de restaurantes e albergues, sentem-se fortemente influenciados a consumi-los ou não a depender das opiniões contidas nessas resenhas (**??**). Diante da relevância que opiniões têm na geração de decisões e no processo de consumo, estudos com o intuito de extraí-las da Web e interpretá-las automaticamente tornaram-se mais frequentes na área de Ciência da Computação. Juntos, esses estudos compõem o que ficou conhecido como **Análise de Sentimento** ou **Mineração de Opinião**[1].

De acordo com (**??**), a área envolve o emprego de diversas técnicas computacionais com o intuito de atingir algum - ou alguns - dos objetivos abaixo:

1. **Identificação de opinião –** Dado um conjunto de documentos, separe fatos de opiniões;
2. **Avaliação de polaridade -** Dado um conjunto de documentos com caráter opinativo e uma palavra-chave (figura pública, empresa etc), classifique as opiniões como positivas ou negativas, ou indique o grau de negatividade/positividade de cada uma delas;

[1] Os dois termos, por serem considerados sinônimos, serão utilizados de forma intercambiável no decorrer desta monografia

3. **Classificação de pontos de vista ou perspectivas -** Dado um conjunto de documentos contendo perspectivas ou pontos de vista sobre um mesmo tema/conjunto de temas, classifique-os de acordo com essas perspectivas/pontos de vista;

4. **Reconhecimento de humor -** Dado um conjunto de textos com caráter emotivo/sentimental, como posts de blogs pessoais, identifique que tipos de humor permeiam os textos e/ou classifique-os de acordo com as diferentes emoções encontradas.

A ideia de utilizar metodologias computacionais para identificar e analisar opiniões é muito anterior à popularização da Web **Citar artigos do fim da década de 60 e começo de 70 que provam isso**. Motivos: pouco dado, IR e ML imaturas. Explicar os 3 e como se relacionam com Natural Language Processing.

1.2 PROPOSTA

Falar de Mineração de Perspectiva. Definir todos os termos correlatos utilizados, fechar os problemas da área e explicar como isso se diferencia de Opinion Mining clássica, que é basicamente Análise de Polaridade.

1.3 ESTRUTURA DA MONOGRAFIA

2 TÉCNICAS BÁSICAS E FERRAMENTAS UTILIZADAS

Neste capítulo, serão descritas técnicas básicas de Aprendizado de Máquina e Processamento de Linguagem Natural importantes para a compreensão dos capítulos seguintes desta monografia. Na seção **Classificadores**, será discutido o funcionamento dos classificadores Naïve Bayes e *Support Vector Machine* (**SVM**), comuns em métodos de Mineração de Perspectiva baseados em frequências de palavras. Na seção **Modelos Generativos**, **serão discutidos os modelos de tópicos *Latent Dirichlet Allocation* (LDA) e *Labeled Latent Dirichlet Allocation* (L-LDA). O segundo, que consiste em uma pequena modificação do primeiro, é utilizado em parte dos experimentos conduzidos neste projeto.**

2.1 CLASSIFICADORES

Diga algo

2.1.1 NAÏVE BAYES

Dado um conjunto de documentos D e um conjunto de classes C, o classificador Naïve Bayes estima, através da aplicação do Teorema de Bayes, a probabilidade de cada $d \in D$ ser de cada uma das classes $c \in C$. Com estas probabilidades calculadas, o classificador determina, para todo d, qual é a classe c a que ele estará associado. Esta classe pode, por exemplo, ser aquela para a qual a probabilidade obtida foi a mais alta (**??**) (**??**).

Este tipo de classificador assume que as informações presentes em um documento, utilizadas na determinação de sua classe, são independentes entre si. No caso de um documento de texto, assume-se que a presença ou ausência de um termo - uma palavra ou uma sequência de palavras - é independente da presença ou ausência de qualquer outro. Definida esta hipótese, a probabilidade de que um documento d seja de uma classe c é computada como

$$P(c|d) \propto P(c) \prod_{k=1}^{t_d} P(t_{dk}|c) \tag{2.1}$$

onde $P(t_k|c)$ é a probabilidade condicional do termo t_k ocorrer em um documento da classe c, $P(c)$ é a probabilidade a priori de um documento qualquer pertencer à classe c e n_d é o número de termos em d (**??**).

As probabilidades envolvidas na equação 2.1 contêm integrais díficeis ou mesmo impossíveis de se calcular. Para calcular $P(c|d)$, portanto, utiliza-se aproximações obtidas através de técnicas de amostragem. Uma destas técnicas, comum na literatura de Aprendizado de Máquina e empregada neste projeto, é a amostragem de Gibbs. Em uma iteração da amostragem, a técnica condiciona as probabilidades calculadas para um documento k às classificações obtidas para os $k-1$ documentos anteriores (**??**). O algoritmo básico da técnica está definido na **figura X**.

A hipótese de independência entre os termos, razão pela qual o Naïve Bayes tem este nome[1], simplifica bastante a estrutura da informação contida nos documentos. Ainda assim, o classificador costuma apresentar boas performances em categorização de textos, sendo utilizado, por exemplo, como base metodológica para alguns filtros de *spam* (**??**). Para melhorar o desempenho do Naïve Bayes, é comum fixar um conjunto de documentos previamente classificados de forma correta e utilizar a informação sobre suas classes na determinação das classes de outros documentos. Ao conjunto de documentos previamente classificados, dá-se o nome de **conjunto de treinamento**; ao conjunto de documentos a serem classificados, **conjunto de teste**.

Todos os experimentos com Naïve Bayes conduzidos neste projeto utilizam a implementação disponível em (**??**). O número de iterações para a Amostragem de Gibbs foi fixado em 500.

2.1.2 SVMS

2.2 MODELOS GRÁFICOS

Modelos gráficos utilizam um grafo para representar relações entre variáveis aleatórias, provendo uma maneira simples de se representar distribuições de probabilidade e propriedades de independência condicional (**??**). Cada vértice corresponde a uma variável aleatória, ou a um

[1] Naïve é uma palavra de origem francesa que significa "ingênua"

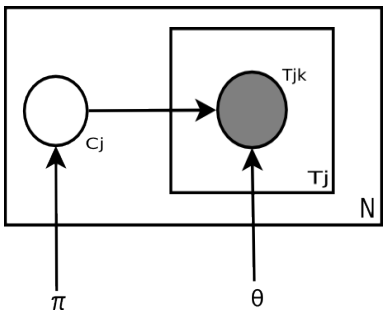

Figura 2.1: Modelo gráfico correspondente ao classificador Naïve Bayes.

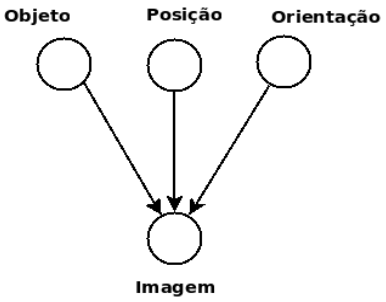

Figura 2.2: Modelo generativo que representa o processo pelo qual objetos são gerados. A posição e a orientação têm probabilidades independentes a priori, algo facilmente identificável, pois não há arestas entre elas. A distribuição de probabilidade para a imagem depende da identidade do objeto, de sua orientação e de sua posição (**??**).

conjunto de variáveis aleatórias, e cada aresta corresponde às relações probabilísticas entre dois vértices. Modelos gráficos são categorizados como Redes Bayesianas ou **Markov Random Fields (traduzir)**. Por questões de escopo, apenas as Redes Bayesianas serão discutidas nesta seção.

Redes Bayesianas, ou Modelos Gráficos Direcionados, utilizam grafos direcionados acíclicos para representar as relações estabelecidas entre variáveis aleatórias. No caso do classificador Naïve Bayes - modelo gráfico que, por conveniência, é discutido mais detalhadamente na seção 2.1.1 -, tem-se variáveis aleatórias que representam a classe de cada documento e, para todos os documentos, uma variável aleatória para cada um de seus termos distintos. Este modelo contém dois parâmetros, π e θ, que ajustam as distribuições de probabilidade das variáveis aleatórias.

A figura **X**, correspondente ao classificador Naïve Bayes, contém toda a notação necessária para as discussões deste projeto. Assume-se que o classificador será aplicado a um conjunto D contendo N documentos, cada um deles com um conjunto de palavras distintas. A classe de cada documento j, C_j, é uma variável aleatória com distribuição de probabilidade $P(C_j|\pi)$. A relação entre C_j e π, evidenciada em $P(C_j|\pi)$, é representada na figura **X** pela aresta que sai de π e incide em C_j. C_j está no interior de um retângulo rotulado com N, o que significa que há N variáveis deste tipo, uma para cada documento, com j variando de 1 a N.

Cada documento j, por sua vez, contém um conjunto de T_j termos distintos. A probabilidade associada a cada termo T_{jk}, k variando de 1 a T_j, é condicionada ao valor da variável C_j. Desse modo, e considerando o parâmetro θ, tem-se, por documento, T_j distribuições de probabilidade do tipo $P(T_{jk}|C_j,\theta)$. Uma vez mais, o rótulo do retângulo mais interno indica o número de variáveis, sintetizadas por um único círculo, que estabelecem o mesmo tipo de relações probabilísticas com θ no modelo. As arestas sempre incidem na variável (ou variáveis) cuja probabilidade está condicionada ao valor da variável - ou parâmetro - de onde elas saem.

Por fim, as variáveis representadas por círculos brancos são latentes - ou seja, seus valores não são observáveis diretamente no conjunto de dados ao qual o modelo é aplicado. As variáveis em cinza, por sua vez, correspondem a dados diretamente observáveis; neste caso, às palavras distintas de cada documento. A **Joint Distribution** do classificador Naïve Bayes, que corresponde ao produto de todas as distribuições de probabilidade do modelo, pode ser escrita como

$$\prod_{j=1}^{N}\left\{P(C_j|\pi)\left[\prod_{k=1}^{T_j}P(T_{jk}|C_j)\right]\right\} \tag{2.2}$$

Adequando-se a notação, observa-se que a relação 2.1, apresentada na seção **??**, pode ser facilmente obtida via a **Joint Distribution** 2.2.

Modelos Generativos

Uma categoria de modelos gráficos explorada neste projeto são os Modelos Generativos. Modelos deste tipo associam distribuições de probabilidade a todas as variáveis aleatórias envolvidas, permitindo a geração - i.e. simulação - de seus valores. Modelos generativos são úteis para expressar os processos através dos quais os dados observáveis são obtidos, como no seguinte exemplo: dado um conjunto de imagens de objetos, variáveis latentes podem ser utilizadas para representar suas posições e orientações; em seguida, para identificar o objeto contido em uma imagem particular observada, é preciso obter a distribuição de probabilidade a posteriori para objetos, o que envolve integrais sobre todas as posições e orientações possíveis (**??**). Com um modelo generativo, o valor das variáveis pode ser obtido através de técnicas de amostragem aplicadas à **Joint Distribution**, como a amostragem de Gibbs apresentada na seção **??**. É um mecanismo diferente do empregado em SVMs, por exemplo, em que variáveis observáveis não são associadas a nenhuma distribuição de probabilidade e as categorias são estimadas diretamente, condicionadas a seus valores. SVMs são exemplos de Modelos Discriminativos, e estão apresentados com mais detalhes na seção **??**.

Dois modelos generativos foram bastante explorados neste projeto: o classificador Naïve

Bayes, apresentado na seção **??**, e o modelo de tópicos L-LDA, que consiste em uma alteração do modelo LDA **explicar as siglas antes**. Estes últimos, discutidos na seção 2.2.1, foram empregados em experimentos e estudos de caso ao longo de todo o projeto, bem como o classificador Naïve Bayes.

figura do bayes figura bishop com legenda

2.2.1 LDA

O modelo LDA associa palavras a tópicos com maior ou menor probabilidade, de tal modo que é possível, após o processamento, atribuir um significado a cada um deles apenas observando as relações entre as palavras associadas. Neste sentido, o modelo é útil para identificar e agrupar padrões contidos em documentos de texto. Os tópicos não são, portanto, pré-identificados antes da execução do modelo, requerendo uma interpretação posterior e subjetiva de seus significados. O LDA se apóia na hipótese de que um documento pode tratar de múltiplos tópicos, e suas palavras refletem que tópicos são estes (**??**).

Dado um conjunto de documentos D, o conjunto de todas as palavras distintas em D (W) e um conjunto de tópicos T, o modelo LDA trata cada $d \in D$ como uma mistura de tópicos, representada por uma distribuição de probabilidade $\theta^{(d)}$ sobre T. Cada tópico t, por sua vez, é tratado como uma mistura de palavras, representada pela distribuição $\phi^{(t)}$ sobre W. A probabilidade da i-ésima palavra de d, w_i, se associar a um tópico t é dada, portanto, pelo produto (**??**)

$$P(w_i|t)P(t) = \phi_{w_i}^{(t)}\theta_t^{(d)} \tag{2.3}$$

$P(w_i|t)$ indica o quanto a palavra w_i se associa a t; $P(t)$, por sua vez, funciona como uma medida do quanto o tópico t é importante no contexto do documento d. O cálculo destas probabilidades também envolve integrais difíceis, ou mesmo impossíveis, de resolver analiticamente, o que implica no uso de técnicas de amostragem e aproximação. A amostragem de Gibbs, descrita superficialmente na seção 2.1.1, é uma alternativa para a inferência dessas probabilidades (**??**), sendo utilizada nos experimentos deste projeto que envolvem o modelo LDA. Para todos os experimentos, a implementação utilizada está disponível em (**??**) e o número de iterações da amostragem é fixado em 100.

Quando um método de classificação é aplicado a um conjunto de documentos, a interpretação do resultado é imediata: cada documento estará associado a uma única classe. No caso

do modelo de tópicos LDA, a interpretação do resultado é mais subjetiva. É preciso observar as palavras que se associam com maior probabilidade a cada tópico, buscando algum tipo de semelhança entre elas, para inferir seus significados. Para ilustrar como as palavras evidenciam o significado de um tópico, um experimento envolvendo receitas culinárias extraídas do *site* **allrecipes.com** foi executado. Apenas os ingredientes de cada receita foram considerados. Na Tabela **??**, constam as quinze palavras mais fortemente associadas a quatro tópicos - ou seja, as quinze palavras que obtiveram mais alta probabilidade de acordo com 2.3.

Análise

L-LDA

Falar das receitas. Falar do uso de Gibbs Sampling para inferência

2.2.2 L-LDA

3 PRINCIPAIS TRABALHOS E DATASETS *ESTUDADOS*

4 *MÉTODOS BASEADOS EM FREQUÊNCIAS DE PALAVRAS*

4.1 INTRODUÇÃO

Métodos baseados em frequências de palavras se apóiam na ideia de que é possível identificar a perspectiva de um documento analisando o seu vocabulário. Eles partem da hipótese de que documentos escritos sob perspectivas diferentes costumam dar destaque a termos distintos, mencionando-os com maior ou menor frequência a fim de reforçar ideias particulares (**??**). O emprego de palavras diferentes para um mesmo propósito, outra hipótese linguística assumida por métodos desse tipo, evidencia pontos de vista diferentes sobre um mesmo assunto. Um exemplo popular no Brasil é o uso dos termos *Revolução* ou *Golpe* para o mesmo evento histórico: o começo do Regime Militar Brasileiro. Enquanto o primeiro termo reflete a perspectiva pró-Ditadura, o segundo reflete a anti-Ditadura. Palavras como *Revolução* e *Golpe*, exploradas por diferentes perspectivas, são chamadas de *banner words*, e têm o objetivo de facilitar a identificação entre adversários e aliados ideológicos (**??**).

Para os métodos revisados neste capítulo, a única informação extraída dos documentos é a frequência de suas palavras. Isto significa que a ordem das palavras em um documento, e as relações sintáticas que elas estabelecem entre si, não são consideradas. Os métodos também ignoram informações referentes ao domínio temático do *dataset*. Apesar de simplificar bastante a estrutura linguística dos documentos, essa informação é a mais explorada pelos trabalhos estudados para esta monografia. Este capítulo revisa casos em que ela foi suficiente para, quando interpretada por um classificador, identificar as perspectivas de um corpus com boa taxa de acerto.

Neste capítulo, artigos que utilizam métodos baseados em frequências de palavras, e obtêm bons resultados, são revisados na seção **Trabalhos Analisados**. Experimentos conduzidos com

um modelo de tópicos do tipo L-LDA[1] e um classificador Naïve Bayes padrão[2], apresentados na seção **Experimentos com L-LDA e Naïve Bayes**, ilustram a relação entre esse tipo de método e os vocabulários de diferentes perspectivas. Por fim, a seção **Conclusões** encerra a discussão sobre esse tipo de método, apresentando considerações sobre seu uso.

4.2 TRABALHOS ANALISADOS

4.3 EXPERIMENTOS COM L-LDA E NAÏVE BAYES

Tópico	Palavras
Genérico	israel, palestinian, israeli, palestinians, state, one, two, israelis, political, right
Pró-Israel	sharon, palestinian, arafat, peace, israeli, prime, bush, minister, american, process
Pró-Palestina	palestinian, israeli, sharon, peace, occupation, international, political, united, people, violence

Tabela 4.1: As dez palavras mais fortemente associadas aos tópicos Pró-Israel, Pró-Palestina e Genérico.

Tópico	Palavras
Genérico	mr., speaker, bill, all, time, people, today, gentleman, federal, support
Democrata	bill, security, legislation, states, chairman, country, act, billion, million, law
Republicano	act, chairman, security, states, bill, legislation, 11, support, 9, system

Tabela 4.2: As dez palavras mais fortemente associadas aos tópicos Republicano, Democrata e Genérico.

Se um método utiliza apenas a frequência das palavras dos documentos para identificar suas perspectivas, é natural que a taxa de acerto seja tão mais baixa quanto menos essas frequências mudam de uma perspectiva para outra. Nesta seção, serão descritos experimentos que evidenciam o vocabulário contido em dois *datasets*, as taxas de acerto obtidas na classificação dos documentos com um Naïve Bayes e a relação entre estas informações.

O primeiro *dataset* estudado é o **Bitterlemons**[3], composto de artigos pró-Israel e pró-Palestina. Cada documento foi associado a um tópico referente à sua perspectiva e outro ge-

[1] Este modelo de tópicos está descrito na seção **X** desta monografia.
[2] Este classificador está descrito na seção **X.X** desta monografia.
[3] A descrição deste *dataset* encontra-se na seção **XXX** desta monografia

*"The recent **Israeli** government decision to begin building extensive walls around **Palestinian** is just one more example of how **Israeli** Prime Minister Ariel Sharon is unable to deal with **Israeli** problems save through his narrow security vision."* - Trecho extraído de artigo Pró-Palestina.
*"The first conclusion that the Israeli political and security establishment should learn and internalize after 18 months of **Palestinian** Intifada, concerns the intensity of **Palestinian** blind terrorism and guerilla warfare against the State of Israel."* - Trecho extraído de artigo Pró-Israel.

Tabela 4.3: Trechos com as palavras *palestinian* e *israeli*, extraídos do *dataset* **Bitterlemons**.

nérico, idêntico para todos eles. Um modelo de tópicos do tipo L-LDA foi aplicado aos documentos assim anotados, agrupando palavras genéricas em torno do tópico genérico, pró-Israel em torno do tópico pró-Israel e pró-Palestina em torno do tópico pró-Palestina. As dez palavras mais fortemente associadas a cada um dos tópicos, excluindo-se *stop words*, estão listadas na Tabela 4.3.

O uso de um tópico genérico ajuda a identificar palavras de *background*, comuns no corpus independentemente de perspectiva. Esta é a diferença fundamental entre o uso de um L-LDA e a simples contagem de palavras em documentos pró-Israel e pró-Palestina. Como esse tipo de contagem não considera palavras de *background*, a visualização de palavras mais específicas para cada perspectiva é prejudicada.

As palavras listadas na Tabela 4.3, para as perspectivas Pró-Israel e Pró-Palestina, remetem semanticamente às discussões entre Israel e Palestina. Parte delas, como *palestinian* e *israeli*, se associam às duas perspectivas, ainda que sejam empregadas nos documentos de forma diferente, como ilustrado pelos exemplos contidos na Tabela 4.3. Outras, como *bush* e *occupation*, funcionam como *banner words*, colaborando com a consolidação de pontos de vista diferentes. O exemplo na Tabela 4.4 ilustra a importância do Governo Bush para Israel à época, enquanto o exemplo na tabela 4.5 evidencia a principal luta Palestina do período: a criação de um Estado próprio. A alta frequência de palavras associadas às perspectivas, bem como a presença de *banner words* importantes, configuram um bom cenário para o uso de métodos baseados em frequências de palavras. O desempenho de um Naïve Bayes na classificação deste *dataset* será discutido mais à frente, ainda nesta seção.

O segundo *dataset* estudado é o **Convote-Menor**, composto de colocações em debates da *House of Representatives*, um dos dois órgãos principais do poder legislativo federal dos Estados Unidos. Os documentos foram marcados como sendo de parlamentares Republicanos ou Democratas, e como representando um posicionamento a favor ou contra a lei em pauta.

*"**Bush** and his advisers, who have been critical of Clinton's deep involvement in a failed peace process ever since taking office, nevertheless understood at the time that peace in the Middle East should be beyond politics in America, and that the US could not permit itself to turn its back on an Israeli leader who was determined to make peace."* - Trecho extraído de artigo Pró-Israel.

Tabela 4.4: Trecho com a palavra *bush*, extraído do *dataset* **Bitterlemons**.

*"But just as we were close to a complete package that would have ended the **occupation** and established a Palestinian state, Barak permitted Ariel Sharon's provocative visit to Al Aqsa mosque, and launched his "revenge"on Palestinians."* - Trecho extraído de artigo Pró-Palestina.

Tabela 4.5: Trecho com a palavra *occupation*, extraído do *dataset* **Bitterlemons**.

Para este experimento, apenas a divisão entre Republicanos e Democratas foi considerada. O L-LDA foi aplicado a este *dataset* de forma análoga ao primeiro experimento, e as dez palavras mais fortemente associadas a cada um dos tópicos - Genérico, Republicano e Democrata - estão listadas na Tabela 4.2. *Stop words* também foram excluídas desta listagem.

As listas de palavras da Tabela 4.2 indicam que o vocabulário do segundo *dataset* não é suficiente para distinguir as perspectivas Republicana e Democrata. Parte das palavras, como *bill*, *legislation*, *states* e *act*, estão mais associadas ao processo legislativo *per se* do que a alguma das perspectivas contidas nos documentos. A alta frequência de palavras como essas, empregadas pelos dois lados do debate, indica um cenário pouco polêmico, com menos *banner words* e divergências. A palavra *security*, por exemplo, fortemente associada às duas perspectivas, é utilizada de forma similar por ambas, como ilustrado na Tabela 4.6. Métodos baseados em frequências de palavras funcionam tão melhor quanto mais distintos forem os vocabulários empregados por cada perspectiva. Por este motivo, é esperado que suas taxas de acerto em *datasets* como este não sejam altas.

As palavras extraídas a partir da aplicação de um L-LDA provêem informações subjetivas sobre a linguagem empregada nos corpora. Ainda assim, essas informações ajudam a entender o comportamento do classificador Naïve Bayes aplicado aos dois *datasets*. Para o **Bitterlemons**, as taxas de acerto obtidas variaram entre 73.46% e 98.98%, a depender da divisão entre os conjuntos de treinamento e teste; para o **Convote-Menor**, entre 48.73% e 54.17%. Não é trivial quantificar a relação entre essas taxas de acerto e a linguagem dos corpora - mas, como o Naïve Bayes utiliza apenas a distribuição das palavras para inferir a perspectiva dos documentos, é evidente que a escolha do vocabulário contribui para a qualidade da classificação.

"Mr. speaker , I wholeheartedly agree that if we want to cut down on illegal immigration , we must improve border ***security****. Just 2 weeks ago, an astute crane operator at the port of Los Angeles discovered 32 Chinese stowaways in a container that had just been unloaded from a Panamanian freighter."* - Trecho de discurso Democrata.
"The fence remains incomplete and is an opportunity for aliens to cross the border illegally. This incomplete fence allows border ***security*** *gaps to remain open. We must close these gaps because they remain a threat to our national* ***security****."* - Trecho de discurso Republicano.

Tabela 4.6: Trechos com a palavra *security*, extraídos do *dataset* **Convote-Menor**.

É válido ressaltar que, a depender do *dataset*, outras questões podem colaborar para um mau desempenho na classificação. Um conjunto de documentos com poucos exemplares, ou contendo poucas palavras, é um cenário onde a classificação com Naïve Bayes pode não funcionar bem. Investigar o vocabulário de um corpus, quando não se obtém uma boa taxa de acerto com classificadores baseados em frequências de palavras, pode ser interessante para verificar se sua uniformidade, ainda que em parte, está relacionada à má classificação obtida. A depender da conclusão retirada, pode-se pensar em estratégias mais específicas resolver o problema.

4.4 CONCLUSÃO

Este capítulo apresentou duas hipóteses linguísticas assumidas por métodos baseados em frequências de palavras: 1) palavras específicas, denominadas *banner words*, costumam ser utilizadas para defender perspectivas diferentes e 2) a quantidade de vezes que uma palavra é mencionada em um documento está diretamente relacionada com seu enfoque (**??**). Como consequência, esses métodos funcionam melhor em *datasets* nos quais o emprego de palavras varia significativamente por perspectiva.

Para ilustrar a relação entre as palavras de dois *datasets* e o desempenho desses métodos, experimentos com o modelo de tópicos L-LDA foram executados. A extração das dez palavras mais fortemente associadas a cada tópico conduziu à visualização parcial de como o vocabulário dos *datasets* se agrupa em torno de suas diferentes perspectivas. A informação, apesar de subjetiva, auxilia na compreensão das taxas de acerto obtidas com um classificador Naïve Bayes padrão, aplicado aos dois corpora. Para o primeiro *dataset*, as taxas de acerto obtidas foram mais altas, o que pode ser explicado por uma presença maior de *banner words* em comparação com o segundo *dataset*.

Métodos baseados em frequências de palavras foram explorados pela maioria dos trabalhos revisados para esta monografia - mesmo fazendo parte de metodologias mais complexas. Apesar de outros fatores contribuírem para o mau desempenho destes métodos, como um número muito pequeno de documentos no *dataset*, é interessante investigar o vocabulário do corpus caso as taxas de acerto obtidas estejam aquém do desejado. O uso de um modelo de tópicos L-LDA, agrupando palavras por perspectiva, é útil para compreender como os autores dos documentos se expressam. Se as palavras são empregadas de modo muito parecido por todas as perspectivas, isto justifica, ainda que em parte, o mau desempenho obtido.

5 METODOLOGIAS QUE USAM INFORMAÇÃO EXTRA-DOCUMENTO

5.1 CONCORDÂNCIA E DISCORDÂNCIA ENTRE DOCUMENTOS

Falar do Get Out the Vote e artigos que seguem a linha

5.2 META-INFORMAÇÕES SOBRE OS AUTORES

6 METODOLOGIAS QUE USAM RELAÇÕES INTRA-DOCUMENTO

Falar de targets, uso de dicionários de polaridade, limitações importantes

7 ESTUDO DE CASO: ELEIÇÕES 2010

8 TRABALHOS RELACIONADOS

9 CONCLUSÃO

O trabalho onon ono non ono non ono non ono non ono non on n ono non ono non ono ono non ononon ono non ono non ono non ono non on

9.1 DIFICULDADES ENCONTRADAS

O trabalho onon ono non ono non ono non ono non ono non on n ono non ono non ono ono non ononon ono non ono non ono non ono non on

9.2 TRABALHOS FUTUROS

Pode-se indicar como trabalhos futuros:

n ono non ono non ono non ono non . n ono non ono non ono non ono non n ono non

ono non ono non ono non n ono non ono non ono non ono non **controlador** n ono non ono non ono ono non ononon ono non ono non ono non ono non on n ono no oo non ono ononon ono non ono non ono non ono non ono non on

ono non ono o non ono non ono ono non ononon ono non ono non ono non ono non on n ono no oo non ono ononon ono non ono non ono non ono non ononon o

APÊNDICE A – RESULTADOS EXPERIMENTAIS

No no nnononono no n ono o nn.

REFERÊNCIAS BIBLIOGRÁFICAS