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

Falar da metodologia de busca dos artigos

2 TÉCNICAS BÁSICAS E FERRAMENTAS UTILIZADAS

Introduzir o capítulo quando tudo já estiver escrito.

2.1 NAÏVE BAYES

O Naïve Bayes é um classificador que se baseia no Teorema de Bayes e assume a **independência condicional** entre as palavras[1] contidas nos documentos (LEWIS, 1998). Isso significa que, para o Naïve Bayes, as palavras em qualquer documento ocorrem independentemente umas das outras. Além disso, o classificador desconsidera a ordem das palavras nos textos: *casa de aline* e *aline casa de* são interpretados da mesma forma. Apesar dessas suposições simplificarem bastante a estrutura linguística dos documentos, o Naïve Bayes foi um dos classificadores mais explorados nos trabalhos revisados para essa monografia. **Sete de treze** estudos, voltados para classificação, fazem uso dele em pelo menos uma parte de seus experimentos.

Dado um documento d_i pertencente a um corpus D com um vocabulário V, o Naïve Bayes deve buscar o valor x para a classe c, $x \in \{0,...,|C|\}$, que maximize a seguinte aplicação do Teorema de Bayes (LEWIS, 1998)

$$P(c = x \mid d_i) = \frac{P(c = x)P(d_i \mid c = x)}{P(d_i)} \tag{2.1}$$

Na prática, como esse é um problema de maximização e a probabilidade $P(d_i)$ independe de qualquer classe, ela pode ser abstraída.

Há $|C|$ valores possíveis para a classe de d_i, distribuídos de acordo com uma distribuição $Binomial(|C|, \pi)$. O parâmetro π é um dos valores que uma variável aleatória ϕ pode assumir, e eles são distribuídos de acordo com uma distribuição $Beta(\alpha, \beta)$. α e β são parâmetros fixados

[1] Outros elementos dos documentos podem ser considerados, como sequências de n palavras (n-gramas).

antes de se iniciar o processo de classificação. Diante disso, a probabilidade de se obter o número real π é dada por (RESNIK; HARDISTY, 2009)

$$P(\pi \mid \alpha,\ \beta) = \frac{1}{B(\alpha,\beta)}\pi^{\alpha-1}(1-\pi)^{\beta-1} \tag{2.2}$$

A função B é aplicada aos valores α e β para garantir que a distribuição de probabilidade *Beta*, quando integrada, some um (EVANS; HASTINGS; PEACOCK, 2000)[2]. Amostrado o valor π de acordo com *Beta*, a probabilidade de se obter c_j tal que $c = c_j$ é dada portanto por (RESNIK; HARDISTY, 2009)

$$P(c_j \mid \pi,\ |C|) = \binom{|C|}{c_j}\pi^{c_j}(1-\pi)^{|C|-c_j} \tag{2.3}$$

Ainda considerando o lado direito da Equação 2.1, assume-se que c_j foi o valor amostrado e deve-se estimar $P(d_i \mid c = c_j)$. A probabilidade de se obter o documento d_i, na prática, não depende de c_j, mas de um parâmetro θ_j amostrado **especificamente** para essa classe. θ_j é um dos valores que uma variável aleatória ε pode assumir, e eles são distribuídos de acordo com uma distribuição $Dirichlet(\gamma_j)$. γ_j é um parâmetro fixado antes de se iniciar o processo de classificação. Sendo assim, a probabilidade de se obter o número real θ_j é dada por (RESNIK; HARDISTY, 2009)

$$P(\theta_j \mid \gamma_j) = \frac{1}{B(\gamma_j)}\prod_{k=1}^{|V|}\theta_{j,k}^{\gamma_{j,k}-1} \tag{2.4}$$

A função B também é utilizada para garantir que a distribuição *Dirichlet*, quando integrada, some um (EVANS; HASTINGS; PEACOCK, 2000). Fixado o valor θ_j, tem-se que todos os documentos possíveis de se amostrar para uma classe c_j estão distribuídos de acordo com uma distribuição $Multinomial(V,\theta_j)$. O primeiro parâmetro dessa distribuição é V porque apenas as palavras dos documentos são consideradas na construção da distribuição. A probabilidade de se obter o documento d_i, dado que a classe amostrada foi uma c_j qualquer, é dada por (RESNIK; HARDISTY, 2009)

$$P(d_i \mid V,\ \theta_j) = \prod_{k=1}^{|V|}\theta_{j,k}^{N(w_k,d_i)} \tag{2.5}$$

$N(w_k,d_i)$, por sua vez, é o número de vezes que a k-ésima palavra do vocabulário V, w_k,

[2]Toda distribuição de probabilidade, quando integrada, deve totalizar exatamente um. (EVANS; HASTINGS; PEACOCK, 2000)

ocorre no documento d_i (RESNIK; HARDISTY, 2009)[3]. A esse número, dá-se o nome de **contagem** de w_k em d_i (NIGAM, 2001).

Os parâmetros α, β e os $|C|$ γ, um para cada classe, recebem o nome de **hiperparâmetros**. Isso se deve ao fato de que eles são diferentes de parâmetros como π e os $|C|$ θ, pois modelam distribuições de probabilidade *antes* de serem feitas observações sobre as classes e palavras de D. Essas distribuições - no caso do Naïve Bayes apresentado, *Beta* e *Dirichlet* - recebem o nome de distribuições ***a priori*** (BISHOP, 2006). Os valores dos hiperparâmetros são fixados antes de se iniciar o processo de classificação. No caso dos dois primeiros, uma estratégia comum é escolher o mesmo valor para ambos, favorecendo uma distribuição uniforme para a variável aleatória ϕ (NIGAM, 2001).

Como o Naïve Bayes assume que as palavras de d_i, $\langle w_{d_{i,1}}, ..., w_{d_{i,|d_i|}} \rangle$, são condicionalmente independentes, a Equação 2.1 pode ser reescrita como (LEWIS, 1998)

$$P(c = x \mid d_i) = \frac{P(c = x) \prod_{k=1}^{|d_i|} P(w_{d_{i,k}} \mid c = x)}{P(d_i)} \tag{2.6}$$

Neste caso, a probabilidade de se amostrar cada palavra $w_{d_{i,k}}$ corresponde ao k-ésimo termo do produtório da Equação 2.5.

Em um cenário de classificação, os valores de todos os parâmetros, classes e documentos devem ser reamostrados iterativamente, a fim de se aproximarem cada vez mais das reais distribuições contidas no corpus. Isso pode ser feito através de alguma técnica de amostragem, como **Gibbs Sampling** ou **Expectation-Maximization**. Por questões de escopo, elas não serão apresentadas nessa monografia, podendo ser consultadas no livro de Aprendizado de Máquina de Christopher Bishop (BISHOP, 2006). De todo modo, é importante saber que essas reamostragens consideram as contagens de palavras de **todos os documentos**. Intuitivamente, a ideia por trás da classificação de um documento é buscar a classe cuja *proporção* das contagens mais se aproxime da sua própria *proporção*. Ela deve ser a classe que maximiza a Equação 2.6.

Na prática, para melhorar o desempenho da classificação, cria-se um *perfil inicial* de contagens para cada classe, informando-se ao Naïve Bayes a classe verdadeira de alguns documentos. Ao conjunto desses documentos, dá-se o nome de **conjunto de treinamento** (BISHOP, 2006). O Naïve Bayes não precisa reamostrar as classes dos documentos desse conjunto, pois elas são informadas antes da classificação. Aos documentos cujas classes não são conhecidas, dá-se o nome de **conjunto de teste** (BISHOP, 2006). A classificação em si, apresentada no parágrafo

[3]É possível utilizar, alternativamente, um *bit*, representando a ausência (0) ou presença (1) da k-ésima palavra de V em d_i.

anterior, só se aplica a esse segundo conjunto, que pode corresponder a D ou a um subconjunto dele, caso parte de seus documentos pertença ao conjunto de treinamento.

Para esse projeto, o Naïve Bayes implementado[4] aplica a técnica de Gibbs Sampling, amostrando valores para as classes e documentos do conjunto de teste até a classificação estabilizar. O número de iterações, fixado em 500, se mostrou mais do que suficiente para estabilizar a classificação em todos os experimentos conduzidos, apresentados nos Capítulos 4 e 7.

2.2 *SUPPORT VECTOR MACHINES* (SVMS)

Support Vector Machines, ou simplesmente **SVMs**, são uma família de métodos que utilizam uma abordagem geométrica para classificação. Eles são fundamentalmente utilizados em problemas de classificação envolvendo duas classes, mas podem ser adaptados para problemas mais complexos. Nesta seção, serão apresentados apenas os princípios de funcionamento de SVMs para duas classes, mais comuns na literatura. Para um aprofundamento sobre SVMs aplicados a problemas com mais de duas classes, recomenda-se a leitura do livro de Aprendizado de Máquina de Christopher Bishop (BISHOP, 2006).

Dado um conjunto de documentos D e um conjunto M de elementos[5] de D, tem-se que cada documento $d \in D$ é representado como um vetor $x \in \mathbb{R}^{|M|}$. Cada entrada de x contém um valor associado a um dos elementos de M. Se M corresponde ao vocabulário de D, por exemplo, cada entrada de x pode corresponder à **contagem** de uma palavra de M em d. Sem perda de generalidade, assume-se que D divide-se em dois conjuntos: **treinamento** e **teste**, de forma semelhante ao apresentado na seção 2.1. Ainda neste sentido, assume-se também que a classe de cada documento é um inteiro: 1 ou -1 (OGURI, 2006). Um SVM deve, portanto, utilizar as representações do conjunto de treinamento em $\mathbb{R}^{|M|}$ para construir os hiperplanos θ_1 e θ_{-1}, conforme as Equações 2.7 e 2.8 (OGURI, 2006)

$$\theta_1 \equiv \mathbf{x} \cdot \mathbf{w} + b = 1 \tag{2.7}$$

$$\theta_{-1} \equiv \mathbf{x} \cdot \mathbf{w} + b = -1 \tag{2.8}$$

O objetivo inicial de um classificador SVM é escolher os parâmetros $\mathbf{w}$ e b que maximizem a

[4] A implementação está disponível no repositório *online* de Aline Bessa: http://github.com/alibezz.

[5] Esses elementos podem ser, por exemplo, o vocabulário do corpus D, como no Naïve Bayes padrão apresentado na seção 2.1.

distância entre esses hiperplanos. Eles podem ser definidos minimizando-se o valor de (OGURI, 2006)

$$\frac{1}{2}||\mathbf{w}||^2 \tag{2.9}$$

Para realizar essa otimização mais facilmente, o problema pode ser remodelado com multiplicadores de Lagrange $\{\alpha_i\}$, $1 \leq i \leq n$, levando à Equação 2.10. Busca-se, então, a minimização desta equação com relação a $\mathbf{w}$ e b e maximização com relação a $\{\alpha_i\}$, com todo $\alpha_i \geq 0$ (OGURI, 2006)

$$L(\alpha, b, \mathbf{w}) = \frac{1}{2}||\mathbf{w}||^2 - \sum_{i=1}^{n} \alpha_i \left[y_i(x_i \cdot \mathbf{w} + b) - 1 \right] \tag{2.10}$$

onde n é a cardinalidade do conjunto de treinamento. Após a obtenção dos valores de $\{\alpha_i\}$ que maximizam a Equação 2.10, a obtenção da classe $y = 1$ ou $y = -1$ de um documento do conjunto de teste, representado por um vetor x_j, é dada pelo sinal do somatório (OGURI, 2006)

$$y(x_j) = sinal\left(\sum_{i=1}^{n} \alpha_i y_i (x_i \cdot x_j) + b \right) \tag{2.11}$$

Esta solução funciona em casos nos quais as representações do conjunto de treinamento, $\{x_1, ..., x_n\}$, são linearmente separáveis - ou seja, obedecem à restrição (OGURI, 2006)

$$y_i(x_i \cdot \mathbf{w} + b - 1) \geq 0, \; i = 1, ..., n \tag{2.12}$$

Quando esses pontos não são linearmente separáveis, essa metodologia precisa ser ajustada, modelando a classificação errônea de documentos. Isto envolve a introdução de n variáveis frouxas[6] ε_i, uma para cada ponto (x_i, y_i). $\varepsilon_i = 0$ se $y(x_i) = y_i$ e $\varepsilon_i = |y_i - y(x_i)|$ em caso contrário (BISHOP, 2006). O SVM deve, neste caso, minimizar (OGURI, 2006)

$$C \sum_{i=1}^{n} \varepsilon_i + \frac{1}{2}||\mathbf{w}||^2 \tag{2.13}$$

onde C é um parâmetro responsável por controlar o compromisso entre a penalidade das variáveis frouxas e a distância máxima dos hiperplanos θ_1 e θ_{-1} ao hiperplano θ_0 (OGURI, 2006). Na modelagem com multiplicadores de Lagrange, a Equação 2.10 deve ser otimizada de tal forma que todo α_i deve ser maximizado obedecendo à restrição $0 \leq \alpha_i \leq C$. Desta forma,

[6] Do inglês *slack variables*.

também se obtém a Equação 2.11 para determinação da classe de um novo documento.

De acordo com um estudo de Ng e Jordan, o Naïve Bayes atinge bom desempenho com um conjunto de treinamento menor do que aquele requerido pelos SVMs (NG; JORDAN, 2002). Considerando essa observação e o fato de que alguns dos *datasets* estudados nesse projeto não são muito grandes, todos os experimentos desenvolvidos nos Capítulos 4 e 7 envolvem apenas o Naïve Bayes. Apesar disso, SVMs fazem parte da metodologia de boa parte dos trabalhos revisados no Capítulo 6 e nos **ANEXOS BLA e BLI** - por esse motivo, fez-se necessário apresentá-los nessa seção.

2.3 MÉTRICAS PARA MENSURAR O DESEMPENHO DOS CLASSIFICADORES

A classificação de documentos, de acordo com suas perspectivas, é um dos principais objetivos dos trabalhos revisados neste projeto. Para medir a qualidade dessa classificação, dado um conjunto de documentos D e um conjunto de classes C, normalmente são utilizadas as seguintes métricas: taxa de acerto, precisão, rechamada ou métrica F1. A **taxa de acerto**[7] é definida por (MANNING; RAGHAVAN; SCHUTZE, 2008)

$$\frac{\#documentos\ classificados\ corretamente}{\#total\ de\ documentos} \tag{2.14}$$

A taxa de acerto não evidencia o quanto o classificador está *errando*, apresentando apenas uma medida de seu sucesso. Para este caso, indica-se o uso da **precisão**. Essa métrica, medida para uma classe $c \in C$ qualquer, é definida por (MANNING; RAGHAVAN; SCHUTZE, 2008)

$$\frac{\#documentos\ classificados\ corretamente\ como\ c}{\#total\ de\ documentos\ classificados\ como\ c} \tag{2.15}$$

A **rechamada**[8], medida também para uma classe $c \in C$ qualquer, é definida como (MANNING; RAGHAVAN; SCHUTZE, 2008)

$$\frac{\#documentos\ classificados\ corretamente\ como\ c}{\#total\ de\ documentos\ pertencentes\ a\ c} \tag{2.16}$$

A rechamada também evidencia o quanto o classificador está *acertando* - mas por classe.

[7] Do inglês *accuracy*.
[8] Do inglês *recall*.

A **métrica F1**[9], também medida para uma classe $c \in C$ qualquer, é dada por (MANNING; RAGHAVAN; SCHUTZE, 2008)

$$2 \times \frac{precisao \times rechamada}{precisao + rechamada} \tag{2.17}$$

A métrica F1 pondera os valores obtidos para a precisão e para a rechamada de uma classe c qualquer. Essas métricas revelam aspectos diferentes do desempenho de um classificador. Por este motivo, é comum encontrar mais de uma delas sendo utilizada no mesmo contexto. Além de medirem desempenho, elas estabelecem critérios objetivos para a comparação entre métodos de classificação, como pode ser visto nos artigos de Lin et al. sobre o conflito Israel-Palestina (LIN et al., 2006) e de Efron sobre orientação cultural (EFRON, 2004).

2.4 VALIDAÇÃO CRUZADA DE *K* DOBRAS

A **validação cruzada de k dobras** é uma técnica estatística que pode ser associada a classificadores que utilizam conjuntos de **treinamento** e **teste** em suas metodologias - caso dos SVMs e do Naïve Bayes, por exemplo. A ideia é estimar o quanto um certo modelo generaliza para um conjunto aleatório de dados (REFAEILZADEH; TANG; LIU, 2009); nesse caso, o modelo é um classificador e os dados são documentos de teste. Nesse tipo de validação, divide-se, aleatoriamente, um conjunto de documentos D em k subconjuntos mutuamente exclusivos, denominados **dobras**[10] (KOHAVI, 1995). O conjunto de teste corresponde a uma das k dobras e as $k-1$ restantes, unidas, compõem o conjunto de treinamento.

A classificação deve ser executada k vezes, com uma dobra diferente como conjunto de teste por vez (KOHAVI, 1995). As métricas associadas ao desempenho de cada classificação podem ser consideradas conjuntamente, através de uma média aritmética, ou em separado (REFAEILZADEH; TANG; LIU, 2009). Os resultados obtidos evidenciam o quanto a classificação conserva seu desempenho, independentemente do conjunto de teste selecionado. Para os trabalhos estudados nessa monografia, os valores mais comuns para k foram **cinco** e **dez**.

2.5 *LATENT DIRICHLET ALLOCATION* (LDA)

O modelo *Latent Dirichlet Allocation*, ou simplesmente **LDA**, fundamenta-se na ideia de que um documento pode tratar de **múltiplos tópicos**, refletidos nas palavras que o compõem

[9] Do inglês *F1-measure*.

[10] Para os experimentos realizados neste projeto, cada dobra tem aproximadamente a mesma cardinalidade.

(GRIFFITHS; STEYVERS, 2004). A finalidade desse modelo não é classificar documentos, mas sim evidenciar como as palavras se relacionam com seus tópicos. O LDA associa as palavras dos documentos a tópicos diferentes, com maior ou menor probabilidade, criando agrupamentos que compartilham uma semelhança semântica/temática.

O LDA interpreta um conjunto de documentos D como uma mistura de tópicos, representada por uma distribuição de probabilidade sobre um conjunto de tópicos T. Cada tópico, por sua vez, é visto como uma mistura de palavras, representada por uma distribuição sobre o vocabulário de D, V. Para cada documento $d \in D$, é fixada uma distribuição de probabilidade sobre tópicos representada por θ_d, dada por $Dirichlet(\alpha)$. De forma análoga, para cada tópico $t \in T$ é fixada uma distribuição de probabilidade sobre palavras, representada por ϕ_t e dada por $Dirichlet(\beta)$ (GRIFFITHS; STEYVERS, 2004). Em seguida, para cada palavra de d, um tópico t específico é escolhido de acordo com a distribuição $Multinomial(T, \theta_d)$. Por fim, uma palavra $w \in V$ é amostrada de acordo com a distribuição $Multinomial(V, \phi_t)$ (BLEI; NG; JORDAN, 2003).

Na seção 2.1, como o Naïve Bayes se fundamenta em uma aplicação do Teorema de Bayes, foram apresentadas as probabilidades de se determinar uma classe e um documento específicos, evidenciando a relação entre classificação e contagens de palavras. Na presente seção, considerou-se mais importante enfatizar outros aspectos do LDA. Para um aprofundamento sobre as probabilidades associadas às escolhas de tópicos e palavras, portanto, sugere-se consulta ao artigo de Blei, Ng e Jordan, no qual o LDA é proposto (BLEI; NG; JORDAN, 2003). Por ora, é suficiente informar que quão maior é a probabilidade de se obter uma palavra w dado um tópico t, mais forte é a relação entre eles. De forma semelhante, quão maior é a probabilidade de se obter um tópico t dada uma distribuição θ_d, mais importante é o tópico no contexto do documento d (GRIFFITHS; STEYVERS, 2004).

No LDA, não se define de antemão a semântica que cada tópico deve ter. Após o processamento, observando-se as palavras que mais frequentemente foram amostradas para cada um deles, é que é possível analisar subjetivamente o significado de todos os tópicos. Para ilustrar como as associações de palavras a um tópico evidenciam seu significado, um experimento envolvendo receitas culinárias extraídas do *site* allrecipes.com[11] foi executado. Cada receita é um documento $d \in D$ e o vocabulário V é a união de todos os ingredientes contidos em D. Na Tabela 2.1, constam as cinco palavras mais frequentemente associadas a quatro tópicos, de acordo com o LDA.

Considerando que todas as receitas pertencem à culinária tradicional dos Estados Unidos,

[11] htpp://allrecipes.com

Tópico	Palavras
Tópico 1	beef, cheese, tomato, sauce, pepper
Tópico 2	chicken, breast, pastum, broth, tomato
Tópico 3	flour, sugar, butter, powder, egg
Tópico 4	cream, cheese, butter, milk, cake

Tabela 2.1: As cinco palavras mais fortemente associadas a quatro tópicos gerados por um LDA.

as palavras listadas na Tabela 2.1 indicam um bom funcionamento do LDA. O primeiro tópico pode ser interpretado como **ingredientes para *cheeseburger***; o segundo, ao associar *chicken*, *pastum* e *broth*, remete a receitas de sopas e caldos comuns em climas frios, podendo ser interpretado como **ingredientes para sopa**; o terceiro pode ser interpretado como **ingredientes para bolo**; o quarto, ao associar *cream*, *cheese* e *cake*, pode ser interpretado como **ingredientes para *cheesecake***. É importante frisar que estas interpretações, apesar de subjetivas, indicam perspectivas culinárias distintas e coerentes internamente. Seria diferente de encontrar, por exemplo, um tópico fortemente associado às palavras *sugar*, *pepper* e *potato*, dificilmente encontradas em uma mesma receita típica dos Estados Unidos.

No LDA, assim como no Naïve Bayes, alguma técnica de amostragem deve ser utilizada para, iterativamente, aproximar as distribuições de probabilidade apresentadas nessa seção o máximo possível daquelas presentes no corpus. A implementação do LDA utilizada na obtenção da Tabela 2.1 utiliza Gibbs Sampling e está disponível no repositório *online* de Alexandre Passos[12]. O número de iterações para o modelo foi fixado em 100, número suficiente para estabilizar as amostragens de tópicos e palavras.

2.5.1 *LABELED-LATENT DIRICHLET ALLOCATION* (L-LDA)

O modelo *Labeled-Latent Dirichlet Allocation*, ou simplesmente L-LDA, é uma variação do LDA em que se restringe o número de tópicos associados a cada documento. Ou seja, as distribuições fixadas para os tópicos de cada documento não necessariamente abrangem todos os tópicos $t \in T$. Além disso, os tópicos presentes em cada documento são identificados antes da execução do modelo, o que diminui a subjetividade envolvida na interpretação de seus significados após o processamento.

Um bom exemplo para ilustrar a aplicação deste modelo envolve um *blog*, em que cada *post* é marcado com um conjunto específico de *tags*. Se cada *tag* é interpretada como um tópico, é possível informar ao L-LDA em que *posts* cada uma delas está presente, processar os *posts* com o modelo e saber, após o processamento, quais palavras se associam mais fortemente a cada

[12] http://github.com/alextp

tag. Neste exemplo, existe um mapeamento direto entre os tópicos e as *tags*, conduzindo a uma interpretação mais imediata do significado de cada agrupamento de palavras.

Experimentos com o L-LDA foram desenvolvidos ao longo deste projeto nos Capítulos 4 e 7, associando cada tópico, por exemplo, a uma perspectiva a ser minerada. Após a execução do modelo, as palavras mais frequentemente associadas a cada perspectiva ilustram como os assuntos discutidos são enfocados por elas. Quanto mais duas perspectivas se distanciam, mais diferentes são as palavras que se associam com destaque a cada uma delas. É válido ressaltar que esse uso do L-LDA não foi encontrado em **nenhum** trabalho revisado para esta monografia.

O L-LDA foi discutido pela primeira vez em um artigo de Ramage et al. (RAMAGE et al., 2009), aplicado ao problema de atribuição de crédito em páginas do *site* del.icio.us[13], marcadas com múltiplas *tags*. O artigo parte da hipótese de que, embora um documento possa estar marcado com várias *tags* diferentes, nem sempre elas se aplicam igualmente a todas as palavras nele contidas. A ideia da atribuição de crédito, portanto, consiste em associar cada palavra do documento às *tags* mais apropriadas e vice-versa.

A implementação de L-LDA utilizada neste projeto utiliza Gibbs Sampling como técnica de amostragem e também está disponível no repositório *online* de Alexandre Passos. O número de iterações para o modelo, em todos os experimentos dos Capítulos 4 e 7, foi fixado em 100. Esse valor se mostrou suficiente para estabilizar as amostragens de tópicos e palavras.

[13] http://www.delicious.com/

3 PRINCIPAIS TRABALHOS E DATASETS *ESTUDADOS*

3.1 BITTERLEMONS

Falar de tudo e dos pre-processamentos

4 CLASSIFICAÇÃO POR PERSPECTIVA BASEADA EM CONTAGENS DE PALAVRAS

A classificação de um documento de acordo com sua perspectiva é o problema mais discutido nos artigos revisados para este projeto. Partindo da hipótese de que documentos escritos sob perspectivas diferentes costumam enfatizar termos distintos (TEUBERT, 2001), classificadores podem ser empregados para, considerando apenas as diferentes ocorrências de palavras em documentos, identificar suas perspectivas. O número de ocorrências de uma palavra em um documento é comumente denominado de **contagem**, e grande parte dos artigos estudados para esta monografia classificam textos baseando-se nessa informação. De treze trabalhos que envolvem classificação, nove se baseiam em contagens de palavras - seis exclusivamente e três combinando essa informação com outras propriedades dos documentos. Pela relevância que a contagem de palavras tem na Mineração de Perspectiva, portanto, esse capítulo é dedicado à revisão e discussão de seu uso na identificação automática da perspectiva de documentos.

Variações no uso de contagem de palavras foram encontradas em parte dos trabalhos estudados para este projeto: alguns deles utilizam valores *booleanos* para indicar a presença (1) ou ausência (0) de palavras nos documentos, como o estudo de Klebanov, Beigman e Diermeier (KLEBANOV; BEIGMAN; DIERMEIER, 2010), e outros empregam as contagens normalizadas em relação ao corpus, como o trabalho de Hirst, Riabinin e Graham (HIRST; RIABININ; GRAHAM, 2010) De todo modo, a hipótese linguística assumida por esses trabalhos é a mesma: textos escritos sob perspectivas diferentes empregam palavras de forma distinta.

Os três artigos que associam contagens de palavras a outras informações não serão apresentados neste capítulo, por uma questão de escopo. Quanto aos outros seis, dois deles são revisados detalhadamente na seção 4.1 e os outros quatro foram fichados e constam no **ANEXO BLA**. Dessa forma, a estrutura do capítulo permanece leve, sem que se diminua o nível de detalhamento da seção 4.1. O critério utilizado para a seleção dos dois artigos foi o número de

citações por outros trabalhos[1]. Na seção 4.2, a relação entre o desempenho da classificação baseada em contagens de palavras e o emprego das mesmas nos documentos é discutida, com o apoio de alguns experimentos. Eles envolvem a aplicação de um modelo de tópicos L-LDA a documentos classificados com um Naïve Bayes, metodologia que não foi encontrada **em nenhum trabalho** estudado para este projeto. Cada perspectiva corresponde a um tópico e, como a classificação se baseia na variação das contagens de palavras por perspectiva, a informação gerada pelo L-LDA amplia a compreensão sobre os resultados obtidos com um Naïve Bayes. Por fim, a seção 4.3 apresenta considerações sobre as informações apresentadas nas seções 4.1 e 4.2.

4.1 TRABALHOS REVISADOS

Dentre os seis trabalhos que se baseiam exclusivamente em contagens de palavras, o artigo de Lin et Al. é o único a utilizar um classificador diferente de um Naïve Bayes ou SVM, introduzindo o método *Latent Sentence Perspective Model* (LSPM) (LIN et al., 2006). O artigo de Mullen e Malouf, por sua vez, é o **único** que não atinge uma boa taxa de acerto na classificação (MULLEN; MALOUF, 2006). Esses artigos são apresentados, nessa ordem, nas subseções 4.1.1 e 4.1.2.

4.1.1 *WHICH SIDE ARE YOU ON? IDENTIFYING PERSPECTIVES AT THE DOCUMENT AND SENTENCE LEVELS*

O trabalho de Lin et al. analisa um conjunto de artigos sobre o conflito Israel-Palestina, escrito por especialistas no assunto e disponibilizado no *site* Bitterlemons[2]. O corpus contém **594** artigos separados entre pró-Israel (297) e pró-Palestina (297), feitos pelos editores do *site* e mais de 200 convidados entre os anos 2001 e 2005. Os autores realizam experimentos com classificadores Naïve Bayes e SVM para classificar os artigos de acordo com suas perspectivas, obtendo taxas de acerto mais altas com o Naïve Bayes (84.85% a 93.46% *versus* 81.48% a 88.22%). Todas as taxas de acerto obtidas nesse trabalho variam de acordo com divisões entre os conjuntos de treinamento e teste: ora é feita validação cruzada de dez dobras, ora os autores alternam os conjuntos entre os artigos escritos por editores e convidados. O ponto mais importante do estudo de Lin et al., entretanto, é a proposição do modelo generativo *Latent Sentence Perspective Model* (**LSPM**), empregado também para classificação.

[1] Os números de citações foram verificados no dia 22 de setembro de 2010, com auxílio do Google Scholar (http://scholar.google.com).

[2] http://www.bitterlemons.org/

Diferentemente do Naïve Bayes, o LSPM associa uma variável latente a cada sentença de um documento, cujos valores indicam se ela carrega ou não a perspectiva do documento. O modelo, portanto, parte do pressuposto de que, até mesmo nos textos mais opinativos, é possível encontrar frases neutras que pouco colaboram para a identificação de sua perspectiva. No processo generativo, amostra-se uma classe para o documento; em seguida, para cada uma de suas sentenças, amostra-se um valor que indica se ela carrega (1) ou não (0) a perspectiva correspondente à classe; por fim, as palavras da sentença são geradas, baseando-se nessas duas informações (classe e presença de perspectiva).

O ponto chave do funcionamento desse modelo tem a ver com o fato de que todas as sentenças que não carregam uma perspectiva são tratadas da mesma forma: admite-se que elas poderiam ocorrer em qualquer documento, independentemente de sua classe. Isso faz com que as palavras geradas para essas sentenças sejam, na maioria das vezes, comuns em todo o corpus. As sentenças que carregam a perspectiva de seus documentos, por sua vez, tendem a gerar termos mais específicos, evidenciando as particularidades do vocabulário de cada ponto de vista. O modelo é adaptado para classificação de forma semelhante ao Naïve Bayes. A diferença principal é que, em vez de considerar o documento como um todo, apenas as palavras geradas por sentenças com perspectiva contribuem para a decisão de qual é a classe do documento.

No trabalho de Lin et al., a taxa de acerto obtida na classificação dos artigos com o LSPM foi ligeiramente superior àquela obtida com o Naïve Bayes - 86.99% a 94.93% *versus* 84.85% a 93.46%. De todo modo, não foi encontrado nenhum outro artigo que faça uso do LSPM - provavelmente pelas dificuldades envolvidas em sua implementação, quando comparado ao Naïve Bayes. O tutorial de Resnik e Hardisty sobre Naïve Bayes e LSPM sugere algumas equações para a implementação dos modelos (RESNIK; HARDISTY, 2009), mas o próprio Resnik afirmou, em *e-mail* endereçado à autora desta monografia, nunca ter conseguido replicar os resultados apresentados por Lin et al.. Segundo ele, o modelo parece ser extremamente sensível ao valor dos hiperparâmetros escolhidos para as distribuições de probabilidade envolvidas (**??**). Essas questões fazem do LSPM uma opção de classificador relativamente complexa.

4.1.2 *A PRELIMINARY INVESTIGATION INTO SENTIMENT ANALYSIS OF INFORMAL POLITICAL DISCOURSE*

O trabalho de Mullen e Malouf analisa um conjunto de *posts* do fórum Politics.com[3], escrito por cidadãos comuns dos Estados Unidos. Apesar de admitirem a diversidade de posicionamentos contidos no fórum, os autores dividem os documentos em apenas duas perspectivas, por uma

[3] http://politics.com

questão de simplicidade: liberal e conservadora. Em seguida, eles condensam todos os *posts* de um mesmo usuário em um só texto, resultando em um corpus com apenas **185** documentos (96 liberais e 89 conservadores). Ao aplicar um Naïve Bayes para classificar os documentos de acordo com a orientação política de seus autores, a taxa de acerto obtida foi de 60.37%, via validação cruzada de dez dobras. O tamanho do *dataset* é apontado por Mullen e Malouf como um dos principais motivos para a obtenção desse resultado. De acordo com eles, a baixa taxa de acerto sugere que o Naïve Bayes é bastante sensível a uma pequena quantidade de artigos na etapa de treinamento.

Os autores discutem também outras hipóteses para o mau desempenho obtido na classificação desse corpus. Por ser composto de *posts* de um fórum, por exemplo, a linguagem do corpus é bastante informal. A probabilidade de se encontrar uma mesma palavra escrita de várias formas é relativamente alta, o que pode criar um certo ruído na classificação dos documentos. Essa hipótese, entretanto, não foi comprovada pelos autores: em alguns experimentos com as palavras corrigidas, a taxa de acerto obtida com o Naïve Bayes foi ligeiramente mais alta (64.48%); em outros, não (60.37%). O artigo também sugere que a presença de documentos menores, correspondentes a usuários do fórum que raramente postam, pode contribuir negativamente para o desempenho do Naïve Bayes. A ideia é que, como esses usuários participam muito pouco do fórum, as palavras em seus *posts* não são suficientes para se consolidar uma perspectiva, tornando-os mais difíceis de se classificar. Restringindo a classificação a usuários que postaram no fórum pelo menos 20 vezes, o número de documentos no corpus tornou-se ainda menor, mas a taxa de acerto obtida com o Naïve Bayes foi ligeiramente superior (61.38%).

Diante desses resultados, os autores sugerem que os usuários não estão se expressando de forma suficientemente diferente no nível das palavras, comprometendo o desempenho do Naïve Bayes. Uma última observação do artigo indica o que pode estar acontecendo: usuários liberais citam falas de usuários conservadores em 62.2% de seus *posts*; analogamente, conservadores citam liberais em 77.5% de seus *posts*. Embora o artigo não indique a proporção média entre essas citações e o restante dos documentos, é possível que elas estejam interferindo negativamente no aprendizado das perspectivas. A presença das perspectivas liberal e conservadora em um mesmo documento, com uma correspondendo às intenções do usuário e outra sendo citada, pode homogeneizar o uso de palavras no corpus, comprometendo a viabilidade da classificação. Uma discussão interessante sobre esse tipo de problema é apresentada por Polanyi e Zaenen em seu artigo sobre aspectos linguísticos que interferem na análise de sentimento (POLANYI; ZAENEN, 2004).

4.2 EXPERIMENTOS COM L-LDA E NAÏVE BAYES

Se um classificador utiliza apenas as contagens de palavras dos documentos para identificar suas perspectivas, sua taxa de acerto é tão mais baixa quanto menos essas contagens mudam de uma perspectiva para outra. Apesar dessa relação ser evidente, não se conhece nenhum método para quantificá-la. Como o seu entendimento amplia a compreensão dos resultados obtidos com esses classificadores, esta seção se detém a ilustrá-la através de alguns experimentos. A ideia é comparar a forma como as palavras são usadas em dois *datasets*: no primeiro, a taxa de acerto obtida com um classificador **Naïve Bayes**, considerando apenas contagens de palavras, deve ser alta; no segundo, baixa. Para a análise do uso das palavras, será utilizado o modelo de tópicos **L-LDA**. A escolha do Naïve Bayes advém do fato de que os *datasets* explorados nesta seção não são muito grandes, e o desempenho desse classificador, nesses casos, tende a ser superior ao obtido com SVMs (NG; JORDAN, 2002). O uso do LSPM não foi cogitado, devido aos pontos discutidos na seção 4.1.2.

O primeiro *dataset* estudado é o mesmo discutido na seção 4.1.1, composto de artigos sobre o conflito Israel-Palestina. A taxa de acerto obtida com um Naïve Bayes aplicado a esse corpus foi alta, variando entre 84.85% a 93.46%. Inicialmente, pensou-se em estudar também o corpus discutido na seção 4.1.2, composto de *posts* do fórum Politics.com. O Naïve Bayes não classificou muito bem seus documentos, atingindo taxas de acerto entre 60.37% e 64.48%. Infelizmente, não foi possível obtê-lo mediante solicitação aos autores do artigo. Por este motivo, o segundo *dataset* estudado provém de outro trabalho: o artigo de Thomas, Pang e Lee sobre classificação de perspectiva em debates políticos dos Estados Unidos (THOMAS; PANG; LEE, 2006). Esse trabalho, cujo corpus está disponível na página de Lee[4], é um dos três que não utilizam apenas contagens de palavras para a classificação, sendo detalhado no Capítulo 6. Por ora, é suficiente informar que ele é composto de **8126** trechos de discursos em debates da *House of Representatives*, um dos dois órgãos principais do poder legislativo federal dos Estados Unidos. Os documentos estão divididos de acordo com duas orientações políticas antagônicas: a republicana (4044) e a democrata (4046). Como no artigo original outras propriedades dos documentos foram consideradas, foi necessário aplicar o Naïve Bayes a esses documentos, considerando apenas as contagens de palavras dos mesmos. No trabalho de Thomas, Pang e Lee, os resultados apresentados são bons - no experimento efetuado com o Naïve Bayes para esse capítulo, entretanto, a taxa de acerto obtida, via validação cruzada de dez dobras, foi de 54.01%[5].

[4] http://www.cs.cornell.edu/home/llee/data/convote.html

[5] A precisão obtida foi de 54.47% e a métrica F1 foi de 51.42%.

Em ambos os *datasets*, cada documento foi associado a dois tópicos: um genérico, idêntico para todos eles, e outro referente à sua perspectiva. No primeiro corpus, essas perspectivas são pró-Israel ou pró-Palestina; no segundo, republicana ou democrata. Há portanto, em cada corpus, três tópicos diferentes. O uso de um tópico genérico associado a todos os documentos ajuda a identificar palavras muito comuns nos *datasets*, independentemente de perspectiva. Essa é a diferença fundamental entre essa aplicação do L-LDA[6] e a simples contagem de palavras em documentos, dividida entre duas perspectivas. Esse tipo de contagem não evidencia que palavras são mais escolhidas em documentos escritos sob uma certa perspectiva e quais são muito utilizadas por todos eles - informação que colabora para um maior entendimento das taxas de acerto supracitadas, obtidas com um Naïve Bayes.

As dez palavras mais frequentemente associadas a cada tópico, retirando-se artigos, conjunções, preposições, advérbios e pronomes pessoais, estão listadas nas Tabelas 4.1 e 4.2. Apesar de pequenas, essas listagens sugerem interpretações sobre como as palavras de cada corpus são exploradas por suas diferentes perspectivas. Essas interpretações, essencialmente subjetivas, ajudam a entender o comportamento do classificador Naïve Bayes aplicado aos dois *datasets*.

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 frequentemente associadas aos tópicos pró-Israel, pró-Palestina e genérico, de acordo com um L-LDA.

Considerando o primeiro corpus, as palavras associadas às perspectivas pró-Israel e pró-Palestina remetem de imediato ao conflito travado entre essas duas nações. Parte delas, como *palestinian* e *israeli*, são bastante mencionadas em ambas as perspectivas, ainda que com propósitos diferentes:

> *"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."*
>
> Retirado de "Peace in peaces", de Ghassan Khatib (pró-Palestina) - 10/06/2002

[6] A implementação de L-LDA utilizada está disponível no repositório online de Alexandre Passos (http://github.com/alextp).

> *"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."*
>
> Retirado de "The lessons Israel should learn", de Meir Pa'il (pró-Israel) - 29/04/2002

Outras palavras, como *bush* e *occupation*, foram mais enfatizadas, respectivamente, pelas perspectivas pró-Israel e pró-Palestina, constando em apenas uma listagem. Os trechos abaixo evidenciam a importância, no período em que os documentos desse corpus foram escritos (2001 - 2005), do Governo Bush para Israel e da criação de um Estado para a nação palestina:

> *"**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."*
>
> Retirado de "Barak was willing, and so were US Jews", de Yossi Alpher (pró-Israel) - 15/07/2002

> *"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."*
>
> Retirado de "Guarding our legitimacy", de Samir Abdullah (pró-Palestina) - 14/10/2002

Associadas ao tópico genérico, também estão as palavras *palestinian* e *israeli*, bastante exploradas pelas duas perspectivas. Isso reflete a importância que esses termos têm para o corpus como um todo. Palavras relacionadas mais genericamente ao conflito Israel-Palestina, como *state* e *political*, também aparecem na listagem.

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 frequentemente associadas aos tópicos republicano, democrata e genérico, de acordo com um L-LDA.

Considerando o segundo corpus, tem-se que parte das palavras associadas às perspectivas republicana e democrata, como *bill*, *legislation*, *states* e *act*, tem mais a ver com o processo legislativo em si do que com alguma delas. Isso sugere que os debatentes não focam seus discursos na defesa de ideias republicanas ou democratas - ou, se o fazem, não utilizam as palavras de forma suficientemente diferente, para que um Naïve Bayes classifique-os corretamente. Reforçando a ideia de que esse corpus é pouco polêmico, dificultando a identificação de perspectivas diferentes, foi observado que algumas das palavras listadas na Tabela 4.2 são, muitas vezes, empregadas com conotações semelhantes por democratas e republicanos. Um exemplo é a palavra *security*, como pode ser observado nos trechos abaixo

> *"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."*
>
> Retirado de discurso de Jane Harman, democrata - 09/02/2005

> *"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****."*
>
> Retirado de discurso de John Boozman, republicano - 02/10/2005

Duas palavras listadas apenas para a perspectiva republicana, *11* e *9*, podem sugerir um maior enfoque no episódio 11 de Setembro e seus desdobramentos. A palavra *billion*, listada apenas para a perspectiva democrata, é muitas vezes utilizada para discutir gastos públicos, o que pode indicar um destaque maior para esse assunto

> *We know that all but one of the* ***9/11*** *hijackers acquired some type of U.S. identification documents. In fact, the 19 hijackers had 63 driver 's licenses among them.*
>
> Retirado de discurso de John Sullivan, republicano - 09/02/2005

> *The Pomeroy bill would cost the treasury $72* ***billion*** *over 10 years, compared with the $290* ***billion*** *price tag of a full repeal through 2015, according to the joint committee on taxation.*
>
> Retirado de discurso de Earl Pomeroy, democrata - 12/04/2005

As palavras associadas ao tópico genérico, assim como aquelas associadas às perspectivas republicana e democrata, têm mais a ver com o processo legislativo e a estrutura dos debates do que com posicionamentos políticos. Alguns exemplos são *bill*, *mr.*, *speaker* e *gentleman*. Isso reforça a ideia de que é difícil identificar as perspectivas dos documentos desse corpus.

As Tabelas 4.1 e 4.2 sugerem que os republicanos e democratas estudados se expressam de forma mais parecida do que os autores do primeiro corpus. Talvez isso advenha do fato de que seus discursos fazem parte de debates, o que pode concentrar as discussões em torno de um conjunto muito específico de termos. No primeiro corpus, os autores não escreveram seus artigos como resposta direta a outros, o que pode colaborar para que eles se expressem de forma mais autêntica e veemente, focando seus textos na apresentação de seus pontos de vista. A aplicação do L-LDA, associando cada tópico a uma perspectiva diferente, se mostrou válida, evidenciando uma certa homogeneização no uso de palavras no segundo corpus, quando comparado com o primeiro.

As palavras das Tabelas 4.1 e 4.2 foram representadas graficamente com o apoio do *software* wordle[7]. Apesar de se associarem aos tópicos de formas distintas, o *software* não foi sensível o suficiente para captar essas diferenças. Por este motivo, essas imagens não constam nessa monografia. O número de vezes que cada palavra se associou a cada tópico, entretanto, pode ser consultado no **ANEXO BLA**.

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 aplicação do Naïve Bayes pode não funcionar bem. Entretanto, esse não parece ser o caso dos corpora analisados nessa seção. De todo modo, quando não se obtém uma boa taxa de acerto com um classificador baseado em contagens de palavras, a investigação subjetiva de como cada perspectiva faz uso delas pode ajudar na compreensão do fenômeno. No segundo corpus analisado nesta seção, por exemplo, o uso de palavras que pouco têm a ver com as perspectivas republicana e democrata, muitas vezes destacadas pelas duas perspectivas, reforça a ideia de que contagens de palavras não são suficientes para classificação nesse cenário.

4.3 CONCLUSÕES

A classificação baseada em contagens de palavras, ou em alguma das variações mencionadas na introdução desse capítulo, assume a seguinte hipótese linguística: a quantidade de vezes

[7] http://wordle.net/

que uma palavra é mencionada em um documento está diretamente relacionada a seu enfoque (TEUBERT, 2001). Como consequência, esse método funciona melhor em *datasets* nos quais o emprego de palavras varia significativamente por perspectiva. Esse parece ser o caso da maioria dos artigos estudados para este capítulo: cinco de seis trabalhos apresentaram uma boa taxa de acerto, considerando apenas contagens de palavras. Outros três trabalhos, apresentados no capítulo 6, também fazem uso dessa informação - mas a associam a outras propriedades dos documentos classificados. De todo modo, a contagem de palavras mostrou ser a característica mais **essencial** à classificação por perspectiva, reforçando a ideia de que essa hipótese linguística é válida.

Nesse capítulo, foram revisados dois trabalhos que utilizam apenas essas contagens para classificar seus *datasets* de estudo - eles são, inclusive, os mais citados dentre os treze analisados nessa monografia.

O primeiro, de Lin et al., apresenta um novo modelo para classificação (o LSPM), destacando-se dos demais artigos por não enfocar em SVMs ou Naïve Bayes. Diferentemente desses classificadores, o LSPM considera que apenas uma parte das sentenças de cada documento realmente apresenta um ponto de vista, e gera palavras mais específicas para cada uma delas. As demais frases concentram palavras mais genéricas, que poderiam ser empregadas da mesma forma por todos os lados do corpus. Por esse motivo, elas não contribuem diretamente com a classificação. O modelo apresenta uma boa taxa de acerto, mas não é tão estudado - nem tão trivial de implementar - quanto um Naïve Bayes. O trabalho é muito citado por ser um dos primeiros a tentar classificar documentos de acordo com suas perspectivas, e o *dataset* analisado (artigos pró-Israel e pró-Palestina) também foi estudado por outros pesquisadores[8].

O segundo, de Mullen e Malouf, é muito citado por ser um dos poucos que trabalha com um *dataset* tão informal (*posts* em um fórum sobre política), apresentando discussões interessantes sobre as dificuldades envolvidas no processo de classificá-lo. As taxas de acerto obtidas na classificação foram baixas, e um dos motivos possíveis, de acordo com o exposto no trabalho, é a quantidade de citações a textos escritos sob uma perspectiva diferente daquela que o autor defende. Isso *mistura* contagens relativas a perspectivas distintas em um mesmo documento, homogeneizando-os sob o ponto de vista do classificador.

O artigo de Lin et al. também apresentou uma boa taxa de acerto utilizando um Naïve Bayes, diferentemente do estudo de Mullen e Malouf. Apesar do corpus analisado por esses últimos ser pequeno, os autores indicam que a forma como cada perspectiva se apropria das palavras interfere na qualidade da classificação. A fim de ampliar a compreensão sobre essa

[8] Verificar **ANEXO BLA**.

interferência, foram conduzidos dois experimentos envolvendo um Naïve Bayes e um L-LDA, aplicados a *datasets* para os quais a classificação funcionou de forma diferente. O primeiro corpus considerado foi aquele estudado por Lin et Al., para o qual as taxas de acerto obtidas com um Naïve Bayes foram altas; o segundo, dado que não foi possível ter acesso aos documentos analisados por Mullen e Malouf, foi um conjunto de trechos de discursos em debates da *House of Representatives*, órgão legislativo dos Estados Unidos. A taxa de acerto obtida com um Naïve Bayes nesse último corpus foi bastante baixa, tornando-o ideal para o objetivo dos experimentos.

Nos experimentos com o L-LDA, cada documento foi associado a dois tópicos: um genérico e outro correspondente à sua perspectiva. Mesmo sabendo que em um Naïve Bayes não há distinção entre palavras genéricas e específicas, optou-se por essa divisão de tópicos por conta do objetivo da aplicação do L-LDA: a visualização parcial de que termos são mais enfocados por cada perspectiva. Essa informação sugere que há uma certa homogeneização nos enfoques do segundo *dataset*, quando comparado com o primeiro. Considerando que quão mais homogêneo é o emprego de palavras por perspectiva, pior é o desempenho dos classificadores baseados em contagens de palavras, a informação obtida com o L-LDA ajuda a compreender porque a taxa de acerto obtida para o segundo *dataset* foi tão mais baixa que aquelas obtidas para o primeiro. É importante frisar que não se encontrou **nenhum outro trabalho** que faça uso de um L-LDA para compreender, ainda que parcialmente, como certos termos são enfocados por diferentes perspectivas. Apesar de outros fatores contribuírem para o mau desempenho de uma classificação, como um número muito pequeno de documentos no *dataset*, a investigação do emprego de palavras, quando as taxas de acerto obtidas não são boas, amplia a compreensão do corpus analisado - o que pode ser útil no momento de se pensar em outras estratégias para melhorar a classificação.

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: PERSPECTIVAS SOBRE O GOVERNO BRASILEIRO*

Muitos dos trabalhos revisados neste projeto analisam documentos que tratam de política. Em particular, boa parte deles estuda textos relacionados a governos federais - quer sejam discussões entre os próprios governantes, como nos estudos de Thomas et al. (THOMAS; PANG; LEE, 2006) ou Hirst et al. (HIRST; RIABININ; GRAHAM, 2010), quer sejam artigos opinativos escritos por cidadãos ou especialistas, como nos artigos de Mullen e Malouf (MULLEN; MALOUF, 2006) (MULLEN; MALOUF, 2008). Considerando essa tendência, e o fato de que 2010 é ano de eleições para presidente no Brasil, decidiu-se realizar um estudo de caso que aproveitasse a abundância de artigos opinativos, disponíveis na *Web*, que tratam do governo Lula e da sucessão presidencial. A ideia é construir um corpus com alguns desses documentos e investigar suas perspectivas automaticamente, classificando-os de acordo com seus posicionamentos e analisando, de forma subjetiva, as palavras por eles enfocadas.

As próximas seções deste capítulo se estruturam da seguinte forma: na seção 7.1, a construção do corpus é apresentada - desde a seleção dos veículos até o pré-processamento dos artigos; na seção 7.2, experimentos com um classificador Naïve Bayes são conduzidos para, assim como em outros trabalhos revisados para este projeto, se classificar artigos de acordo com suas perspectivas; na seção 7.3, o modelo de tópicos L-LDA é aplicado ao corpus, evidenciando aspectos da linguagem explorada por artigos com posicionamentos diferentes; por fim, na seção 7.4, são apresentadas conclusões sobre o estudo e possíveis extensões para ele.

7.1 CONSTRUINDO UM CORPUS PARA ESTUDO

Os artigos escolhidos para este estudo foram extraídos de colunas, *blogs* e *sites* políticos mantidos por jornalistas de notoriedade nacional. A coleta de *posts* de *blogs* escritos por cidadãos comuns também foi cogitada - entretanto, como eles são pouco conhecidos, comentados

e divulgados, essa opção exigiria um esforço de análise manual dos *posts* que foge ao escopo deste projeto. Além disso, uma vantagem em focar o estudo em material publicado por jornalistas conhecidos é poder correlacionar, posteriormente, os resultados obtidos a investigações sobre a formação de opinião na mídia brasileira *online* - tanto na alternativa quanto na tradicional.

A seleção dos veículos para este estudo de caso resultou do consenso entre a autora desta monografia e dois jornalistas **COMO CITÁ-LOS, DIZER SEUS NOMES?**. O critério básico para as escolhas foi a defesa clara de um ponto de vista sobre o governo Lula e/ou a sucessão presidencial de 2010. Assim como em outros artigos revisados nesta monografia, que dividem os corpora analisados em dois lados antagônicos, assume-se que os artigos do corpus desse estudo de caso dividem-se entre pró e anti governo. O lado pró-governo é composto de artigos veiculados em:

1. **Luis Nassif Online**[1] Este é o *blog* do jornalista Luis Nassif, premiado como Melhor Blog de Política pelo iBest 2008[2]. Nassif, que já trabalhou na TV Cultura e Rede Bandeirantes, mantém o *blog* há cinco anos, enfocando principalmente assuntos relativos à política brasileira. Artigos do *blog* são frequentemente citados, de forma positiva, em veículos de campanha pró-governo, como os *sites* Blog da Dilma[3] e Os Amigos do Presidente Lula[4]. De fato, o Luis Nassif Online adota um posicionamento pró-governo, como comprovam os trechos a seguir:

   > *"Desde o ano passado, estava claro [sic] a falta de competitividade de José Serra, seja por não ter feito um governo brilhante em São Paulo, por não representar o novo e por não conseguir desenvolver um discurso próprio."*
   >
   > Retirado de *"Em Minas, a mãe de todas as batalhas"* - 02/09/2010

   > *"Na entrevista, Bonner se limitou a perguntar da dependência de Dilma em relação à Lula [...] A consequência foi Dilma rebatendo com facilidade cada bobagem dita, reforçando o discurso social, mas sem avançar em uma proposta sequer de programa, explicando a lógica das alianças políticas. E William Bonner interrompendo-a a toda hora, impedindo sequer uma resposta completa. Algo tão desastrado e mal educado que obrigou Fátima Bernardes,*

[1] http://www.advivo.com.br/luisnassif/
[2] http://idgnow.uol.com.br/internet/2008/05/21/ibest-2008-anuncia-vencedores/
[3] http://dilma13.blogspot.com/
[4] http://osamigosdopresidentelula.blogspot.com/

> *do alto de sua elegância, a calá-lo com um sinal, para que parasse de ser inconveniente."*
>
> Retirado de *"O dia em que William Bonner escorregou"* - 10/08/2010

Como o veículo possui muito conteúdo, foram considerados apenas os artigos da categoria "Eleições".

2. **Conversa Afiada**[5] O *site* se define como um portal de jornalismo independente, contendo principalmente artigos produzidos por Paulo Henrique Amorim. O jornalista, que já trabalhou para as Redes Globo e Bandeirantes e para a revista Carta Capital, mantém o *site* desde 2006. Enfocando a política brasileira, o Conversa Afiada apóia, dentre outras iniciativas do governo federal, a candidatura da ex-ministra Dilma Rousseff[6]. Os trechos abaixo justificam a escolha do *site* como representante da mídia *online* pró-governo:

> *"O Governo Lula é um sucesso e a popularidade dele, recordista desde o primeiro dia de Governo. Promoveu a inclusão social, ampliou a classe média e assistiu os pobres. Fez uma política externa que não tirou o sapato para os Estados Unidos. A Dilma é a sua legítima sucessora: foi a CEO do Governo Lula. O Serra é um nada."*
>
> Retirado de *"A Dilma não é um tsunami. Dilma é o rio que segue para o mar"* - 27/08/2010

> *"Segundo a tevê DEMO-Tucana da Bahia, a afiliada da Globo, Jacques Wagner está na frente de Paulo Souto por 46% a 19%. Paulo Souto é o aliado de Serra na Bahia. A TV Bahia, também."*
>
> Retirado de *"Sumiram com o dinheiro do Serra. Serra é barrado em procissão"* - 07/08/2010

Também por possuir muito conteúdo, apenas os artigos pertencentes à categoria "Política"foram considerados.

3. **Escrevinhador**[7] O *blog*, mantido pelo *site* da revista Caros Amigos, é escrito pelo jornalista Rodrigo Vianna, que também é repórter da Rede Record. Ele está no ar desde 2008, enfocando acontecimentos da vida política do Brasil e do Mundo. No que diz respeito

[5] http://www.conversaafiada.com.br/

[6] http://www.conversaafiada.com.br/brasil/2010/07/02/mino-explica-por-que-apoia-a-dilma-porque-ela-e-melhor-que-o-serra/

[7] http://www.rodrigovianna.com.br/

ao Brasil, o conteúdo do *blog* assume uma perspectiva pró-governo, como ilustram os trechos abaixo:

> *"Abandonado pelos aliados do DEM e do PSDB, em queda nas pesquisas, Serra refugia-se na mídia. O candidato do PSDB virou isso: porta-voz dos interesses da velha mídia. Faz sentido. É quem, em última instância, sustenta a candidatura."*
>
> Retirado de *"Serra, porta-voz da velha mídia; é Zé ou Mané?"* - 19/08/2010

> *"O programa da Dilma foi um show. [...] Foi um programa em que Lula não apareceu mais que Dilma, e nem sumiu – porque seria falso, ela é a candidata dele. Foi um programa em que Lula passou o bastão a Dilma. De forma eficiente, corajosa e, ao mesmo tempo, emocionante."*
>
> Retirado de *"Dilma acerta a mão; Serra quer virar 'Zé'"* - 18/08/2010

Como o *blog* também trata de outros assuntos, apenas as categorias "Plenos Poderes"e "Palavra Minha", mais direcionadas à política, foram consideradas para extração de artigos.

4. **Brasília, eu vi**[8] O *blog*, escrito pelo jornalista Leandro Fortes, que também trabalha para a revista Carta Capital, agrega alguns de seus artigos para a revista e outros textos sobre política. Estes artigos têm boa recepção em *sites* de campanha pró-governo, como o Blog da Dilma[9]. De fato, eles assumem uma perspectiva de defesa da situação, como justificam os trechos abaixo:

> *"Assim, enquanto a imprensa mundial se dedica a decodificar as engrenagens e circunstâncias que fizeram de Lula o mais importante líder mundial desse final de década, a imprensa brasileira se debate em como destituí-lo de toda glória, de reduzí-lo a um analfabeto funcional premiado pela sorte, a um manipulador de massas movido por programas de bolsas e incentivos [...]."*
>
> Retirado de *"Não verás Lula nenhum"* - 18/05/2010

> *"Ao acusar o presidente Luiz Inácio Lula da Silva de ter transformado o Brasil em uma "república sindicalista", José Serra optou por agregar a seu modelito eleitoral, definitivamente, o discurso udenista de origem, de forma literal, da*

[8] http://brasiliaeuvi.wordpress.com/
[9] http://dilma13.blogspot.com/2010/08/caso-lunus-verdade-dos-fatos.html

> *maneira como foi concebido pelas elites brasileiras antes do golpe militar de 1964."*
>
> Retirado de *"Serra precisa de mais amigos"* - 15/07/2010

O lado anti-governo, por sua vez, é composto de artigos veiculados em:

1. **Reinaldo Azevedo**[10] O *blog*, escrito pelo jornalista homônimo, é mantido pela revista Veja. Autor da frase *"Tudo que é bom para o PT é ruim para o Brasil"* (AZEVEDO, 2008), Reinaldo Azevedo, que já foi editor da Folha de S. Paulo, alimenta seu *blog* com críticas ao governo atual, como evidenciam os trechos abaixo:

   > *"O problema dos petistas é que eles são viciados no aulicismo, na cortesania. Ao conviver com pessoas que sempre têm um preço, ficam chocados e tomam como ofensa pessoal a descoberta de que nem todos se comportam com essa moral anã."*
   >
   > Retirado de *"Presidente do PT repete ladainha autoritária do programa 'Rubriquei, mas não traguei'. Ou: 'Ai que vontade de censurar a Veja!!!' Contenha a coceira, companheiro!* - 15/07/2010

   > *"Cinco centrais sindicais assinaram um vergonhoso manifesto contra a candidatura do tucano José Serra à Presidência. Antes de mais nada, e a despeito da mentira essencial que está contida no texto — já falo a respeito —, cumpre destacar: trata-se de um manifesto ilegal, de mais um crime eleitoral escancarado."*
   >
   > Retirado de *"Acuado pelo 'Rubriquei, mas não traguei', PT mobiliza centrais sindicais. E elas assinam um documento ilegal e mentiroso."* - 12/07/2010

2. **Coluna do Augusto Nunes**[11] A coluna, parte da revista Veja, é escrita pelo jornalista Augusto Nunes, que também apresenta o programa Roda Viva na TV Cultura. Seus artigos têm má recepção em alguns veículos que defendem o atual governo, como o Luis Nassif Online[12] e o Blog da Dilma[13], justamente por assumirem uma posição anti-governo. Os trechos abaixo justificam esta perspectiva:

[10] http://veja.abril.com.br/blog/reinaldo/
[11] http://veja.abril.com.br/blog/augusto-nunes/
[12] http://www.advivo.com.br/blog/luisnassif/serra-e-fhc-uma-relacao-delicada
[13] http://dilma13.blogspot.com/2010/01/mais-uma-do-tucano-augusto-nunes.html

> *"Como todo sinal de alarme, o som de um neurônio em ebulição é perturbador, mas muito útil. Quem tem juízo entenderá que Dilma Rousseff não é uma candidata em campanha. É uma ameaça a caminho."*
>
> Retirado de *"O som perturbador do neurônio em ebulição"* - 20/07/2010

> *"O eleitor merece saber se Lula recebeu uma herança maldita e reconstruiu o país, como repete há pelo menos seis anos, ou se resolveu valer-se de mentiras e fantasias para desqualificar o legado do antecessor que acabou com a inflação, consolidou a democracia constitucional e fixou diretrizes econômicas que, em sua essência, vigoram até hoje."*
>
> Retirado de *"FHC aceita o convite para o duelo que Lula não pode recusar."* - 11/02/2010

Todos os artigos extraídos dessa coluna pertencem à categoria "Direto ao Ponto", por ela tratar especificamente da política brasileira atual.

3. **Coluna do Diogo Mainardi**[14] A coluna, escrita desde 2002, é a mais lida da revista Veja segundo ela mesma, reunindo críticas à política e à economia brasileiras. O jornalista Diogo Mainardi se opõe aos governos petistas, tendo inclusive publicado, em 2007, o livro Lula é Minha Anta (MAINARDI, 2007), no qual agrupa diversos artigos escritos para sua coluna na Veja. Os trechos abaixo ilustram a posição de Mainardi como um grande crítico do governo do PT e de sua candidata Dilma Rousseff:

   > *"Dilma Rousseff teve uma loja de produtos importados. O empreendimento durou menos de um ano e meio. Se Dilma Rousseff mostrar como presidente da República o mesmo talento que mostrou como empresária, o Brasil já pode ir fechando as portas."*
   >
   > Retirado de *"Dilma 1,99 Rousseff"* - 04/09/2010

   > *"No futuro, quando alguém quiser relatar os fatos deste período, terá de recorrer necessariamente aos processos judiciais, que detalharam o modo lulista de se organizar, de se acumpliciar, de se infiltrar e de fazer negócios."*
   >
   > Retirado de *"A história em inquéritos"* - 20/03/2010

4. **Portal de Carlos Alberto Sardenberg**[15] O portal contém artigos do jornalista para suas colunas nos jornais O Globo e O Estado de S. Paulo, além de outros textos de análise política e econômica. Além destas ocupações, Sardenberg também é comentarista da

[14] http://veja.abril.com.br/blog/mainardi/
[15] http://www.sardenberg.com.br/site/index.php

TV Globo e âncora da Rádio CBN, tecendo comentários sobre a economia mundial e brasileira. Os trechos abaixo transparecem seu posicionamento anti-governo:

> *"O governo Lula não quer fazer concessões à iniciativa privada porque está num ímpeto estatizante, em ano eleitoral. Só que o Estado não tem os recursos para fazer nada de substancial. Fica por isso mesmo."*
>
> Retirado de *"As tarefas de Lula"* - 22/03/2010

> *"É verdade que o país está de novo em um bom momento. Mas não é verdadeira a conclusão que o 'lulismo' tira disso: que isso tudo só está acontecendo porque Lula é o presidente."*
>
> Retirado de *"A salvação?"* - 01/04/2010

É válido ressaltar que os autores dos artigos muitas vezes colocam trechos de notícias, ou mesmo textos opinativos de outros autores, em seus escritos, colaborando para a riqueza da linguagem no corpus.

Outros veículos foram cogitados, como o Blog do Noblat[16], o *blog* de Miriam Leitão para o jornal O Globo[17], a coluna de Cristiana Lôbo para o portal G1[18] e o *blog* de Celso Ming para o jornal O Estado de S. Paulo[19]. Os posicionamentos contidos nestes veículos, entretanto, não foram considerados claros o suficiente para os propósitos deste estudo.

Todos os artigos contidos nas colunas, *sites* e *blogs* selecionados foram publicados entre 01/01/2010 e 06/09/2010. O período fixado, por fazer parte de um ano eleitoral, encerra uma quantidade significativa de artigos pró e anti-governo - muitos deles focados na sucessão presidencial. Por este motivo, e também para manter o escopo do estudo atrelado às eleições 2010, artigos de anos anteriores não foram coletados. A extração dos documentos foi feita de forma automatizada com *scripts* escritos nas linguagens de programação Python e **UNIX Shell script**[20]. Como os jornalistas eventualmente publicam sobre política mundial ou outros assuntos, foi feita uma filtragem nos artigos, de modo a restarem apenas aqueles que contêm pelo menos uma das seguintes palavras-chave: "Lula", "FHC", "Dilma", "Serra", "Marina", "PT", "PV", "PSDB". Todos os documentos foram, por fim, anonimizados, para que os nomes de seus autores não interferissem nos estudos.

[16] http://oglobo.globo.com/pais/noblat/
[17] oglobo.globo.com/economia/miriam/
[18] http://g1.globo.com/platb/cristianalobo/
[19] blogs.estadao.com.br/celso-ming/
[20] Todos eles estão disponíveis no repositório *online* de Aline Bessa (http://github.com/alibezz)

Veículo	Coleta	Filtragem/Anonimização
Reinaldo Azevedo	2490	2377*
Augusto Nunes	579	450
Diogo Mainardi	40	32
Carlos Sardenberg	59	33
Conversa Afiada	375	337
Luis Nassif Online	994	525
Escrevinhador	222	179
Brasília, eu vi	34	24

Tabela 7.1: Quantidades de artigos disponíveis em cada etapa da construção do corpus. *Apenas 550, amostrados aleatoriamente, foram aproveitados.

Após filtragem e anonimização, restaram 1065 artigos pró-governo e 2747 anti-governo. Para os estudos feitos com o corpus, envolvendo o classificador Naïve Bayes e o modelo de tópicos L-LDA, reduziu-se a quantidade de documentos anti-governo para 1065, utilizando-se apenas 550 dos 2377 artigos extraídos do *blog* de Reinaldo Azevedo, amostrados aleatoriamente. Essa estratégia foi adotada porque o desempenho do Naïve Bayes se mostrou sensível a quantidades muito discrepantes de palavras por perspectiva. Como o uso do L-LDA estende as análises feitas com o classificador, decidiu-se manter o corpus idêntico para ambos os estudos.

O número de artigos coletados em cada veículo varia bastante, como pode ser observado na Tabela 7.1. No corpus **Bitterlemons**[21], estudado por Lin et al., este comportamento também é observado, e os resultados obtidos são de alta qualidade (LIN et al., 2006). Isto reforça a ideia de que essa variação não interfere significativamente na qualidade dos experimentos feitos com o corpus deste estudo de caso.

7.2 IDENTIFICANDO PERSPECTIVAS COM UM CLASSIFICADOR NAÏVE BAYES

O primeiro estudo conduzido com esse corpus consiste na classificação dos artigos de acordo com suas perspectivas - problema fundamental na área de Mineração de Perspectiva (PANG; LEE, 2008). O classificador Naïve Bayes, escolhido para o estudo por sua simplicidade, se mostrou adequado para o problema: a taxa de acerto obtida foi de 89.43%; a precisão, de 89.68%; a métrica F1, de 89.42%. Assim como em outros artigos revisados para esta monografia (LIN et al., 2006) (MULLEN; MALOUF, 2006) (KLEBANOV; BEIGMAN; DIERMEIER, 2010), esses valores foram obtidos via validação cruzada de dez dobras. O bom desempenho do método indica que a simples análise das palavras utilizadas nos artigos - descon-

[21] A descrição deste corpus encontra-se na seção 3.1 desta monografia.

siderando, portanto, aspectos sintáticos e semânticos dos mesmos - já evidencia suas diferentes perspectivas.

Os valores obtidos com o classificador Naïve Bayes são comparáveis àqueles apresentados por Durant e Smith em seu trabalho sobre o posicionamento de *blogs* frente às atitudes de George W. Bush na guerra do Iraque (DURANT; SMITH, 2006): 89.77%. É válido ressaltar que, diferentemente da metodologia adotada por Durant e Smith, nenhuma palavra foi descartada no processamento dos textos para este estudo de caso. A alta taxa de acerto obtida encoraja estudos semelhantes ao desenvolvido por Durant e Smith, envolvendo *blogs* e *sites* políticos brasileiros escritos por cidadãos comuns.

Os artigos escolhidos para este corpus são compostos, muitas vezes, de textos de outros autores. Isto reforça o fato de que o classificador Naïve Bayes está efetivamente aprendendo as perspectivas dos documentos, em vez de estilos de escrita. De todo modo, assim como no estudo de Lin et al. com o corpus **Bitterlemons** (LIN et al., 2006), foi conduzido um experimento em que os artigos pertencentes aos conjuntos de treinamento e teste são escritos por autores diferentes. Se o que está sendo aprendido são de fato as perspectivas dos documentos, a performance do classificador não deve ser muito diferente da obtida na validação cruzada de dez dobras. Testando com artigos da coluna de Augusto Nunes e do *site* Conversa Afiada, e treinando com os demais, a taxa de acerto obtida foi de 92.79%, acompanhada de precisão de 93.32% e métrica F1 de 91.86%. Este experimento, portanto, ratifica os outros resultados, evidenciando que o classificador Naïve Bayes cumpre bem a tarefa de identificar as perspectivas pró e anti governo.

7.3 ILUSTRANDO A LINGUAGEM POR PERSPECTIVA

O bom desempenho do classificador Naïve Bayes indica que o simples processamento das palavras contidas nos artigos já é suficiente para a compreensão automática das perspectivas pró e anti governo. Para aprofundar o estudo sobre a linguagem de cada perspectiva, o modelo generativo L-LDA foi aplicado ao corpus. Cada artigo foi associado a dois tópicos: um genérico, igual para todos eles, e um referente à sua perspectiva (pró ou anti governo). Há, portanto, três tópicos diferentes nesta aplicação. O modelo relaciona as palavras contidas nos documentos a seus tópicos, de modo que aquelas mais comuns se associam mais frequentemente ao tópico genérico, enquanto outras, mais particulares de cada perspectiva, aos outros dois tópicos.

A Tabela 7.2 indica que as palavras utilizadas por autores com posicionamentos diferentes muitas vezes são as mesmas, diferindo apenas na forma como são enfatizadas. Os artigos anti-

Tópico	Palavras
Genérico	governo, brasil, serra, estado, lula, poder, presidente, nacional, vez, campanha, federal, história, pt, forma, pessoas, psdb, vida, brasileira, dinheiro, programa, texto, lei, ministro, nome, direito, brasileiro, momento, eleitoral, passado, ministério
Pró-Governo	serra, dilma, lula, psdb, presidente, pt, candidato, folha, tucano, eleições, partido, jornal, campanha, pesquisa, fhc, brasil, henrique, eleitoral, candidata, rousseff, globo, mundo, governador, entrevista, imprensa, presidência, petista, dem, candidatura, turno
Anti-Governo	dilma, lula, presidente, brasil, rousseff, pt, gente, candidata, mundo, candidato, petista, entrevista, partido, josé, eleições, chefe, tucano, presidência, sarney, eleitoral, petistas, fernando, casa, ministro, companheiro, amigo, planalto, brasileiros, senador, saber

Tabela 7.2: As trinta palavras mais frequentemente associadas aos tópicos Pró-Governo, Anti-Governo e Genérico, em ordem e excluindo-se artigos, preposições, conjunções, advérbios e pronomes pessoais.

governo, por exemplo, dão muito destaque às palavras *lula* e *dilma*; os pró-governo, por sua vez, também enfatizam estas palavras, mas dão um destaque maior a *serra*, candidato à presidência pelo PSDB. A associação de palavras semelhantes, ainda que em intensidades diferentes, aos tópicos anti e pró governo advém do fato de que os artigos compartilham um tema geral - o governo brasileiro - e, consequentemente, o mesmo vocabulário básico. É diferente do que acontece quando os tópicos correspondem a temas diferentes em vez de perspectivas, como pode ser visto no trabalho de Ramage et al. sobre L-LDA e *tags* de *blogs* (RAMAGE et al., 2009).

As palavras na Tabela 7.2 estão ordenadas de acordo com o número de vezes que se associam aos tópicos, mas isto não é suficiente para compreender o quanto cada perspectiva realmente as enfatiza. Para compreender melhor o uso das palavras pelos diferentes pontos de vista do corpus, elas foram processadas pelo *software* wordle[22], resultando nas figuras 7.1, 7.2 e 7.3. O tamanho das palavras nas imagens corresponde ao quanto elas se associam a cada tópico[23]. As imagens 7.1 e 7.2 evidenciam o destaque dado aos políticos Lula, Dilma Rousseff e José Serra nos artigos analisados. A imagem 7.3 dá certo destaque a Lula e José Serra, mas também enfatiza outros termos, como *governo* e *brasil*, relacionados mais genericamente ao tema geral dos artigos: a política brasileira.

É válido ressaltar, por fim, que, apesar dos textos terem caráter opinativo, as palavras elen-

[22] http://wordle.net

[23] Os valores correspondentes a cada uma das palavras, por tópico, se encontra no **ANEXO BLA**.

Figura 7.1: Representação gráfica para as trinta palavras mais frequentemente associadas ao tópico pró-governo.

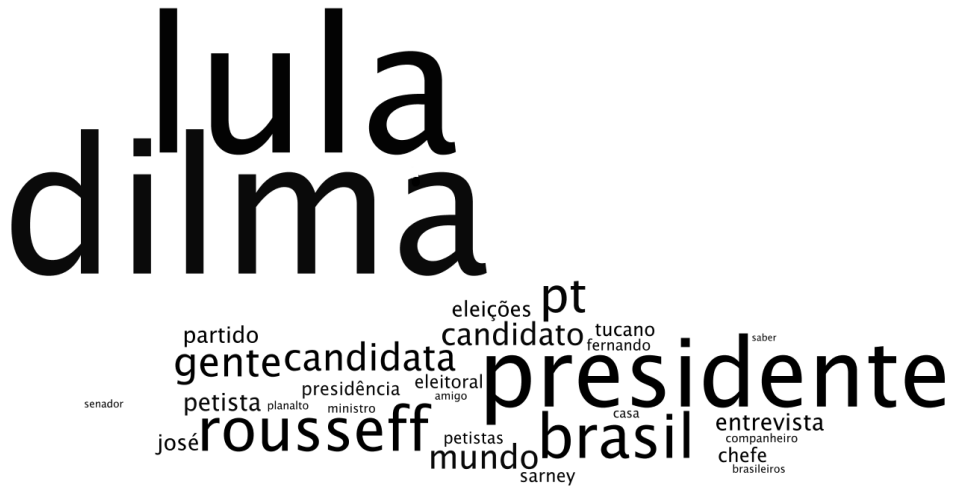

Figura 7.2: Representação gráfica para as trinta palavras mais frequentemente associadas ao tópico anti-governo.

cadas na Tabela 7.2 não carregam uma polaridade natural, como no caso dos adjetivos "bom"ou "ruim". Para aprofundar o entendimento da relação que elas estabelecem com as perspectivas dos artigos, portanto, recomenda-se ler um número razoável de passagens de texto que as contenham. Alguns trechos foram selecionados abaixo, em caráter ilustrativo:

> *"O que parece estarrecedor para quem nunca ouviu **Dilma** antes - e tenho colegas jornalistas que nunca a viram discursando ou dando **entrevista** - é absolutamente familiar para os frequentadores desta coluna. Que há nove meses têm acesso a veementes indícios, há muito transformados em provas documentais, de que **Dilma** é uma afronta imposta ao **Brasil** por **Lula**, num [sic] crime lesa-pátria sem perdão."*
>
> Retirado de "O som perturbador do neurônio em ebulição", da coluna de Augusto Nunes -

Figura 7.3: Representação gráfica para as trinta palavras mais frequentemente associadas ao tópico genérico.

20/07/2010

*"Que o **tucano José** Serra se saiu muito melhor no **Jornal** Nacional e que a eleição é, sim, de continuidade — no sentido de que não cabe mais falar em ruptura. E fiz uma crítica ou outra ao governo **Lula**."*

Retirado de "A cabeça dos brasileiros... autoritários", do *blog* de Reinaldo Azevedo - 15/08/2010

*"A última bala na agulha do **Serra** é a baixaria. Só que, na era da internet, a baixaria – Lunus (para desconstruir Roseana Sarney) e aloprados do **PT** (para mandar as ambulâncias superfaturadas para o Inferno) – não tem o mesmo efeito do passado.É o caso dos aloprados do tal dossiê que ele vai ter que explicar na Justiça."*

Retirado de "Serra só tem uma saída: pendurar FHC no pescoço", do *site* Conversa Afiada - 07/06/2010

*"A entrevista de **Dilma** ao JN foi didática: **Dilma** conseguiu colar sua **candidatura** como continuidade das políticas do governo **Lula**. Ponto pra ela.Por outro lado, o casal número um do JN da **Globo** escorregou e mostrou claramente contra quem trabalham em 2010 e a favor de quem se esforçam para mudar tudo o que está aí."*

Retirado de "O povo não é (mais) bobo...", do *blog* Luis Nassif Online - 10/08/2010

7.4 CONCLUSÕES E ESTUDOS FUTUROS

Este estudo de caso, inicialmente, apresentou todos os passos envolvidos na criação de um corpus sobre a atual política brasileira, dividido entre as perspectivas pró e anti governo. É válido ressaltar que não foi encontrado nenhum outro corpus brasileiro desenvolvido para um estudo de Mineração de Perspectiva. A alta taxa de acerto obtida com um classificador Naïve Bayes, na identificação das perspectivas dos artigos, evidencia que a escolha de palavras feita por seus autores já reflete suficientemente seus pontos de vista, claramente antagônicos. Resultados semelhantes foram obtidos em outros corpora revisados neste projeto, conforme abordado no capítulo 4.

O experimento com o modelo de tópicos L-LDA, por sua vez, proporciona uma análise subjetiva da linguagem dos artigos, evidenciando os diferentes enfoques dados por cada perspectiva. As figuras 7.1 e 7.2, referentes, respectivamente, às perspectivas pró e anti governo, apresentam uma característica em comum: ambas enfatizam termos que têm a ver com o lado a que se opõem. No primeiro caso, a palavra *serra*, que corresponde ao candidato à presidência da oposição José Serra, é rapidamente visualizável. De forma análoga, no segundo caso, as palavras *lula* e *dilma*, que correspondem ao atual presidente e sua candidata, recebem mais destaque. Essas figuras também indicam que os artigos pró-governo dão mais enfoque a personalidades relacionadas à situação, como Lula e Dilma Rousseff, do que os anti-governo a personalidades da oposição, como José Serra ou Marina Silva. Essa última, inclusive, candidata à presidência pelo PV, não é mencionada nas palavras listadas na Tabela 7.2, o que indica que os veículos, no período analisado, concentraram seus antagonismos em personalidades políticas dos partidos PT, como Lula e Dilma Rousseff, e PSDB, como José Serra. Por fim, é importante frisar que não foi encontrado nenhum outro estudo de Mineração de Perspectiva que tenha feito uso do modelo de tópicos L-LDA para analisar o emprego de palavras por diferentes perspectivas.

Futuramente, pretende-se estender este estudo de caso a textos políticos escritos por cidadãos comuns em seus *blogs*, o que pode contribuir para a compreensão de como o brasileiro se posiciona politicamente na Internet. Além disso, o estudo também deve ser ampliado para identificar as perspectivas contidas nos comentários feitos aos artigos do corpus, a fim de se avaliar como eles refletem o posicionamento dos leitores em relação àquilo que leram. Este tipo de análise pode ajudar a compreender o impacto destes artigos em seus leitores e a formação de opinião na mídia brasileira *online*.

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

AZEVEDO, R. *O pais dos petralhas*. [S.l.]: Record, 2008. ISBN 978-85-01-08232-9.

BISHOP, C. *Pattern Recognition and Machine Learning*. [S.l.]: Springer, 2006. ISBN 0387310738.

BLEI, D.; NG, A. Y.; JORDAN, M. I. Latent dirichlet allocation. *Journal of Machine Learning Research*, v. 3, p. 993–1022, 2003.

DURANT, K. T.; SMITH, M. D. Predicting the political sentiment of web log posts using supervised machine learning techniques coupled with feature selection. In: . [S.l.: s.n.], 2006. p. 187–206.

EFRON, M. Cultural orientation: Classifying subjective documents by cociation analysis. *Proceedings of the AAAI Fall Symposium on Style and Meaning in Language, Art, Music, and Design*, p. 41–48, 2004.

EVANS, M.; HASTINGS, N.; PEACOCK, B. *Statistical Distributions*. [S.l.]: Wiley-Interscience, 2000. ISBN 0471371246.

GRIFFITHS, T. L.; STEYVERS, M. Finding scientific topics. *Proceedings of the National Academy of Sciences*, v. 101, p. 5228–5235, Abril 2004.

HIRST, G.; RIABININ, Y.; GRAHAM, J. Party status as a confound in the automatic classification of political speech by ideology. In: *Proceedings of JADT 2010*. [S.l.: s.n.], 2010. p. 173–182.

KLEBANOV, B. B.; BEIGMAN, E.; DIERMEIER, D. Vocabulary choice as an indicator of perspective. In: . [S.l.: s.n.], 2010. p. 253–257.

KOHAVI, R. A study of cross-validation and bootstrap for accuracy estimation and model selection. In: *Proceedings of the 14th international joint conference on Artificial intelligence*. [S.l.: s.n.], 1995. p. 1137–1143.

LEWIS, D. D. Naive (bayes) at forty: The independence assumption in information retrieval. In: *Proceedings of ECML-98, 10th European Conference on Machine Learning*. [S.l.]: Springer Verlag, 1998. p. 4–15.

LIN, W.-H. et al. Which side are you on? identifying perspectives at the document and sentence levels. *CoNLL'06: Proceedings of the Conference on Natural Language Learning (CoNLL)*, 2006.

MAINARDI, D. *Lula e minha anta*. [S.l.]: Record, 2007. ISBN 8501080705.

MANNING, C.; RAGHAVAN, P.; SCHUTZE, H. *An introduction to Information Retrieval*. [S.l.]: Cambridge university Press, 2008. ISBN 9780521865715.

MULLEN, T.; MALOUF, R. A preliminary investigation into sentiment analysis of informal political discourse. In: . [S.l.: s.n.], 2006. p. 159–162.

MULLEN, T.; MALOUF, R. Taking sides: User classification for informal online political discourse. *Internet Research*, v. 18, p. 177–190, 2008.

NG, A.; JORDAN, M. On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. In: *NIPS '02*. [S.l.: s.n.], 2002. v. 15.

NIGAM, K. P. *Using unlabeled data to improve text classification*. Dissertação (Mestrado) — Carnegie Mellon University, 2001.

OGURI, P. *Aprendizado de Maquina para o Problema de Sentiment Classification*. Dissertação (Mestrado) — Pontificia Universidade Catolica do Rio de Janeiro, Rio de Janeiro, 2006.

PANG, B.; LEE, L. *Opinion Mining and Sentiment Analysis*. [S.l.]: Foundations and Trends in Information Retrieval series. Now publishers, 2008.

POLANYI, L.; ZAENEN, A. Contextual valence shifters. In: *Proceedings of the AAAI Spring Symposium on Exploring Attitude and Affect in Text: Theories and Applications*. [S.l.: s.n.], 2004.

RAMAGE, D. et al. Labeled lda: A supervised topic model for credit attribution in multi-labeled corpora. In: . [S.l.: s.n.], 2009. p. 248–256.

REFAEILZADEH, P.; TANG, L.; LIU, H. *Cross Validation*. [S.l.]: Springer, 2009.

RESNIK, P.; HARDISTY, E. *Gibbs Sampling for the Uninitiated*. [S.l.], 2009. Disponível em: <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.156.2875>.

TEUBERT, W. A province of a federal superstate, ruled by an unelected bureaucracy - keywords of the euro-sceptic discourse in britain. In: *Attitudes towards Europe: Language in the unification process*. [S.l.: s.n.], 2001. p. 45–86.

THOMAS, M.; PANG, B.; LEE, L. Get out the vote: Determining support or opposition from Congressional floor-debate transcripts. In: *Proceedings of EMNLP*. [S.l.: s.n.], 2006. p. 327–335.