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

4.1 Representação gráfica para as trinta palavras mais frequentemente associadas ao tópico pró-governo. 49

4.2 Representação gráfica para as trinta palavras mais frequentemente associadas ao tópico anti-governo. 49

4.3 Representação gráfica para as trinta palavras mais frequentemente associadas ao tópico genérico. 50

LISTA DE ABREVIATURAS E SIGLAS

SUMÁRIO

1 Introdução **8**

1.1 Objetivo 11

1.2 Trabalhos Relacionados 11

2 Técnicas básicas **13**

2.1 Naïve Bayes 13

2.2 *Support Vector Machines* (SVMs) 17

2.3 Métricas para mensurar o desempenho dos classificadores 19

2.4 Validação cruzada de *k* dobras 20

2.5 *Labeled-Latent Dirichlet Allocation* (L-LDA) 21

3 Classificação por perspectiva: revisão e discussão **23**

3.1 Trabalhos Revisados 24

3.1.1 Trabalhos que exploram contagem ou presença de palavras 25

3.1.2 Trabalhos que exploram outras características dos documentos 28

3.1.3 Comparações 31

3.2 Experimentos com L-LDA e Naïve Bayes 31

3.3 Conclusões 36

4 Estudo de caso: Perspectivas sobre o governo brasileiro **39**

4.1 Construindo um corpus para estudo 39

4.2 Identificando perspectivas com um classificador Naïve Bayes 46

4.3 Ilustrando a linguagem por perspectiva 47

4.4 Conclusões e estudos futuros 51

5 Conclusão **52**

5.1 Dificuldades encontradas 52

5.2 Trabalhos futuros 52

Apêndice A – Resultados experimentais **53**

Referências Bibliográficas **54**

1 INTRODUÇÃO

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* 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 revelou 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 (COMSCORE; KELSEY-GROUP, 2007). 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**, termos utilizados como sinônimos (PANG; LEE, 2008) (LIU, 2006).

De acordo com a *survey* de Pang e Lee, referência mais citada para o estudo da área, a Mineração de Opinião envolve o emprego de diversas técnicas computacionais com o intuito de explorar algum dos tópicos abaixo (PANG; LEE, 2008):

1. **Polaridade de sentimento ou graus de polaridade -** Dado um documento opinativo, para o qual se assume que as opiniões se referem basicamente a um único assunto, classifique-o como positivo ou negativo em relação a esse assunto (polaridades opostas) ou localize-o no espectro estabelecido entre duas polaridades opostas;

2. **Detecção de subjetividade e identificação de opinião -** Dado um documento, detecte

se ele é subjetivo ou não, constituindo-se de fatos ou opiniões, ou que partes dele são subjetivas;

3. **Análise de tópico-sentimento -** Dado um documento opinativo, assume-se que suas opiniões podem se referir a tópicos diferentes, e deve-se identificar quais opiniões interagem com quais tópicos;

4. **Pontos de vista ou perspectivas -** Dado um documento opinativo, que apresente uma perspectiva sobre um tema (um ponto de vista, uma orientação ideológica, um posicionamento), em vez de um sentimento polarizado sobre um único assunto, classifique-o de acordo com essa perspectiva;

5. **Outras informações não-factuais -** Dado um documento com caráter emotivo/sentimental, identifique que tipos de humor o permeiam e/ou classifique-o de acordo com as emoções encontradas.

O item **Pontos de vista ou perspectivas**, em particular, é de grande aplicabilidade para as Ciências Humanas. Diferentemente dos outros tópicos, mais aplicáveis à compreensão de opiniões **pontuais** sobre marcas/produtos/pessoas públicas, esse item investiga um fenômeno mais profundo: o posicionamento[1] de indivíduos a respeito de temas mais **abrangentes**. Não se trata, portanto, de investigar opiniões estritamente polarizadas, como *positivo*, *negativo*, *bom* ou *ruim* - mas sim perspectivas e ideologias, como *pró-aborto* ou *anti-pena de morte*. Estas questões motivaram o enfoque dessa monografia para esse item. Em outras palavras, a temática explorada por essa monografia é a **classificação de documentos de acordo com suas perspectivas**.

Como a *survey* de Pang e Lee elenca **apenas três** trabalhos que envolvem classificação de documentos por perspectiva, um dos objetivos principais dessa monografia foi explorar essa revisão, criando um documento que possa servir como referência para estudos futuros nessa linha. Os trabalhos revisados foram escolhidos de acordo com metodologia definida no Capítulo 3, e são apresentados no mesmo. Todos eles classificam documentos baseando-se em como eles usam palavras. A ideia, que remete a estudos de Linguística, é de que indivíduos com posicionamentos diferentes utilizam as palavras de formas distintas (TEUBERT, 2001) - e isso é explorado na classificação.

A revisão executada nesse projeto levou a discussões e experimentos também apresentados no Capítulo 3, que visam à investigação de como o uso das palavras interfere na classificação

[1] Os termos *posicionamento*, *orientação*, *perspectiva*, *ponto de vista* e *ideologia* são utilizados de forma intercambiável nessa monografia, por serem explorados da mesma forma na literatura revisada para este projeto.

de documentos por perspectiva. Esses experimentos ilustram aspectos interessantes dos documentos, ampliando a discussão da classificação por perspectiva. Não foi encontrado **nenhum** outro trabalho que apresente experimentos semelhantes aos apresentados no Capítulo 3 - ou seja, além da revisão em si, tudo indica que a execução desses experimentos é uma contribuição inédita para a área de Mineração de Opinião.

Além da revisão acompanhada de experimentos, esse projeto propõe um estudo de caso envolvendo posicionamentos sobre a política brasileira no Capítulo 4. Além de propor uma classificação por perspectiva, esse capítulo também discute o uso de palavras por cada uma delas. Não foi encontrado **nenhum** trabalho que classifique documentos brasileiros de acordo com seus pontos de vista, o que faz desse estudo, pelo que tudo indica, o primeiro envolvendo uma temática brasileira. Os resultados são animadores, indicando que a classificação de documentos de acordo com seus pontos de vista também pode ser aplicada, de forma bem sucedida, a *datasets*[2] em português.

Nesse trabalho é feita, primeiramente, uma descrição básica dos principais classificadores explorados nessa monografia, no Capítulo 2. Eles são o Naïve Bayes e os *Support Vector Machines* (SVMs), muito populares na área de Aprendizado de Máquina. A aplicação deles compõe a metodologia básica de todos os trabalhos estudados - o que varia, de fato, são as representações dos documentos e a forma como eles são pré-processados. No Capítulo 2 também são apresentadas métricas para se avaliar o desempenho de uma classificação e uma técnica de validação para esse desempenho. Ainda nesse capítulo, por fim, é apresentado um modelo utilizado nos Capítulos 3 e 4, com a finalidade de ilustrar o uso de palavras por documentos escritos sob perspectivas diferentes. O modelo é o *Labeled-Latent Dirichlet Allocation* (L-LDA), que associa documentos a tópicos e relaciona suas palavras a cada um deles. No Capítulo 3, alguns trabalhos são selecionados e revisados de forma comparativa. Este capítulo também discute a relação entre o uso de palavras nos documentos e o desempenho da classificação. No Capítulo 4, é apresentado um estudo de caso envolvendo a política brasileira. O escopo desse capítulo envolve a construção de um corpus[3], a definição dos pontos de vista a serem considerados, a classificação de documentos de acordo com eles, a discussão do uso de palavras e uma série de considerações finais e indicações de estudos futuros. Por fim, no Capítulo 5, são apresentadas - ou ratificadas - conclusões a respeito dos conteúdos explorados pelos Capítulos 3 e 4. Esse capítulo também discute as principais dificuldades encontradas nesse projeto. As próximas seções apresentam, respectivamente, os objetivos desse trabalho e alguns trabalhos relacionados.

[2] *Datasets* são conjuntos de dados; nesse caso, de documentos.

[3] Nesta monografia, os termos *corpus* e *dataset* serão utilizados de forma intercambiável.

1.1 OBJETIVO

O objetivo desse trabalho é explorar aspectos teóricos e práticos da classificação de documentos de acordo com suas perspectivas. Para a parte teórica, foi elaborada uma revisão de artigos que tratam do assunto; além disso, a fim de aprofundar a compreensão dos trabalhos revisados, discute-se a relação entre o uso de palavras e a classificação, com o apoio de alguns experimentos. Para a parte prática, foi proposto um estudo de caso envolvendo um corpus de política brasileira, explorando todos os passos envolvidos na sua classificação.

1.2 TRABALHOS RELACIONADOS

Classificação de documentos por perspectiva é uma área recente, que vem se popularizando nos principais eventos de Processamento de Linguagem Natural. Apesar disso, ainda há pouco material de apoio para uma introdução ao assunto, sendo necessário ir diretamente aos *sites* desses eventos em busca de artigos. O principal trabalho encontrado nessa direção foi justamente a *survey* de Pang e Lee, que propõe a temática como um subproblema da área de Mineração de Opinião (PANG; LEE, 2008). Esse trabalho apresenta uma boa introdução à área de Mineração de Opinião, apresentando suas principais aplicações, desafios, técnicas e métodos diretamente envolvidos com a identificação, extração e análise de informação opinativa. Dada a abrangência da *survey*, a pouca atenção dedicada à temática dessa monografia era esperada. De todo modo, esse material é aquele que mais se aproxima dos objetivos desse projeto.

No que diz respeito à Mineração de Opinião em geral, Bing Liu propõe materiais introdutórios desde 2006 (LIU, 2006) (LIU, 2010). O enfoque desses trabalhos é apresentar as principais definições da área, seus problemas e técnicas para resolvê-los. Apesar de se aproximar, em diversos pontos, da *survey* de Pang e Lee, seus trabalhos não possuem um caráter de revisão, nem apresentam uma divisão clara da área em subproblemas. Por estes motivos, os trabalhos de Bing Liu funcionam como um bom material de apoio para essa monografia, mas não se relacionam tão intimamente com seus objetivos como a *survey* de Pang e Lee.

Os trabalhos relacionados ao estudo de caso desse projeto, explorado no Capítulo 4, consistem na revisão em si, apresentada no Capítulo 3. São diversos artigos que se propõem a classificar documentos de acordo com suas perspectivas, utilizando uma mesma metodologia básica, com variações significativas na representação dos documentos e na temática dos *datasets*. No que diz respeito à temática explorada, a política brasileira, a ferramenta Eleitorando[4]

[4]http://www.eleitorando.com.br/site/

se aproxima deste projeto. A sua finalidade, entretanto, é monitorar opiniões sobre candidatos às Eleições 2010 nas redes sociais, como Twitter[5] ou YouTube[6]. As informações monitoradas são classificadas como *positivas*, *negativas* ou *neutras*. Não se trata, portanto, do enfoque proposto pela classificação por perspectiva, que busca identificar o ponto de vista defendido em um documento, em vez de opiniões polarizadas. Por fim, o uso de um L-LDA para aprofundar a compreensão dos resultados obtidos com a classificação não foi encontrado em nenhum trabalho estudado para esse projeto.

[5] http://twitter.com/
[6] http://br.youtube.com/

2 TÉCNICAS BÁSICAS

Este capítulo apresenta os principais classificadores explorados nessa monografia, o Naïve Bayes e os *Support Vector Machines*, respectivamente nas seções 2.1 e 2.2. Eles são empregados por todos os trabalhos revisados no Capítulo 3, ainda que alguns deles proponham o uso de outras técnicas. Como elas são muito particulares de cada trabalho, não serão exploradas neste capítulo. O classificador Naïve Bayes também é utilizado em experimentos conduzidos no Capítulo 3 e no estudo de caso sobre a mídia e o governo brasileiro, apresentado no Capítulo 4.

Na seção 2.3, são apresentadas métricas para se avaliar o desempenho de um classificador qualquer. Além de indicarem a qualidade da classificação, elas estabelecem critérios objetivos para a comparação entre diferentes metodologias. Na seção 2.4, a técnica de validação cruzada de k dobras é discutida. Ela evidencia como um método de classificação generaliza para diferentes conjuntos de documentos. Todos os trabalhos apresentados no Capítulo 3, e o estudo de caso do Capítulo 4, fazem uso dessa técnica. Por fim, na seção 2.5, é apresentado o modelo de tópicos *Labeled-Latent Dirichlet Allocation*, ou simplesmente L-LDA (RAMAGE et al., 2009). Esse modelo interpreta documentos como misturas de tópicos, agregando palavras a cada um deles com maior ou menor intensidade. Seu uso facilita a compreensão de como o conteúdo de um corpus qualquer se segmenta, algo explorado nos Capítulos 3 e 4. A ideia, em ambos os capítulos, é identificar que palavras se associam mais frequentemente a cada perspectiva de um corpus, representada como um tópico.

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. 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 apresenta um bom desempenho na classificação por perspectiva,

como pode ser conferido nos trabalhos revisados no Capítulo 3.

O Naïve Bayes é um classificador probabilístico, cuja finalidade é encontrar a classe mais provável para um documento qualquer. Dado um documento d_i pertencente a um corpus D com um vocabulário V, e um conjunto de classes C, a probabilidade da classe c de d_i ser x, $x \in C$, dado d_i ($P(c = x \mid d_i)$) é expressa por uma aplicação do Teorema de Bayes (LEWIS, 1998)

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

onde $P(c = x)$ é a probabilidade de se obter a classe x, independentemente de qualquer documento d_i; $P(d_i \mid c = x)$ é a probabilidade de se obter o documento d_i fixando-se a classe x - ou, em outras palavras, a probabilidade de d_i pertencer à classe x -; e $P(d_i)$ é a probabilidade de se obter o documento d_i, independentemente de qualquer classe. Na prática, como o classificador deve buscar a classe x que maximize a Equação 2.1, e a probabilidade $P(d_i)$ independe de qualquer classe, ela pode ser abstraída.

Há $|C|$ valores possíveis para a classe c de d_i, e o Naïve Bayes assume que eles são distribuídos de acordo com uma distribuição $Binomial(|C|, \pi)$. O parâmetro π, por sua vez, é assumido como um dos valores possíveis para uma variável aleatória ϕ. Assim como o classificador associa uma distribuição binomial para os valores possíveis da classe c, ele assume que os valores de ϕ estão distribuídos de acordo com uma distribuição $Beta(\alpha, \beta)$. Os parâmetros α e β são fixados antes de se iniciar o processo de classificação. Diante disso, a probabilidade de se obter o valor π é dada por (RESNIK; HARDISTY, 2009)

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

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)[1]. Considerando-se o lado direito da Equação 2.1, e o fato de que os valores para c são distribuídos de acordo com $Binomial(|C|, \pi)$, tem-se que a probabilidade de se obter $c = c_j$, $P(c = cj)$, é dada na prática por (RESNIK; HARDISTY, 2009)

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

Ainda considerando o lado direito da Equação 2.1, assume-se sem perda de generalidade

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

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. O Naïve Bayes assume que θ_j é um dos valores que uma variável ε pode assumir, e eles são distribuídos de acordo com uma distribuição $Dirichlet(\gamma_j)$. O parâmetro γ_j também deve ser fixado antes de se iniciar o processo de classificação. Sendo assim, a probabilidade de se obter o valor θ_j é dada por (RESNIK; HARDISTY, 2009)

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

A função B, aplicada a γ_j, 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, definida na Equação 2.1 como $P(d_i \mid c = c_j)$, pode ser reescrita portanto como (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, ocorre no documento d_i (RESNIK; HARDISTY, 2009). A esse número, dá-se o nome de **contagem** de w_k em d_i (NIGAM, 2001). Observa-se, portanto, que as contagens de palavras em d_i estão intrinsecamente relacionadas com a Equação 2.1, pilar da classificação. Alternativamente, é possível utilizar, no lugar de $N(w_k, d_i)$, um *bit* representando a ausência (0) ou presença (1) da k-ésima palavra de V em d_i. De todo modo, essa mudança não interfere no uso da Equação 2.1.

Os parâmetros α, β e γ_j, $j \in \{0, ..., |C| - 1\}$, recebem o nome de **hiperparâmetros**. Isso se deve ao fato de que eles são diferentes de parâmetros como π e θ_j, $j \in \{0, ..., |C| - 1\}$, pois modelam distribuições de probabilidade *antes* de serem feitas observações sobre as classes e palavras de D. Essas distribuições - no caso, $Beta$ e $Dirichlet$ - recebem o nome de distribuições ***a priori*** (BISHOP, 2006). No caso dos hiperparâmetros α e β, uma estratégia comum é escolher o mesmo valor para ambos, favorecendo uma distribuição uniforme para a variável aleatória ϕ (NIGAM, 2001). Estratégia semelhante pode ser adotada na escolha das $|V|$ entradas de um γ_j qualquer.

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 parâmetros e classes 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). Em ambas as técnicas, valores iniciais devem ser atribuídos às classes dos documentos, de acordo com as distribuições apresentadas nos parágrafos anteriores. Em seguida, eles devem ser reamostrados considerando-se as **contagens** de todas as palavras do corpus, separadas por classe. A contagem de uma palavra w em uma classe é a soma das contagens de w em todos os seus documentos (RESNIK; HARDISTY, 2009). Na prática, no decorrer das iterações, tem-se que o θ que maximiza a Equação 2.5 está associado à classe cujas contagens de palavras, **proporcionalmente**, mais se assemelham às contagens de d_i (RESNIK; HARDISTY, 2009). Intuitivamente, portanto, tem-se que quão mais diferentes forem essas proporções por classe, mais difícil é, para o Naïve Bayes, *se enganar* na escolha da classe de d_i. O artigo de Mullen e Malouf sobre política dos Estados Unidos (MULLEN; MALOUF, 2006), inclusive, associa o mau desempenho obtido na classificação de documentos a proporções muito parecidas de contagens por classe. Essa questão é discutida mais detalhadamente no Capítulo 3.

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 anterior, só se aplica a esse segundo conjunto, que pode corresponder a D ou a um subconjunto dele, caso parte de seus documentos constitua um conjunto de treinamento. Todos as aplicações de Naïve Bayes discutidas nos Capítulos 3 e 4 fazem uso de conjuntos de treinamento e teste.

Apesar de simples, o Naïve Bayes apresenta um bom desempenho na classificação de documentos por perspectiva, como evidenciam os trabalhos revisados no Capítulo 3. Exemplos de

sua aplicação podem ser encontrados nesse capítulo e também no Capítulo 4. Para esse projeto, a implementação desenvolvida[2] aplica a técnica de Gibbs Sampling, amostrando valores para classes e parâmetros 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 3 e 4.

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[3] 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)

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

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

As representações de documentos que pertencem a θ_1 ou θ_{-1} recebem o nome de *vetores suporte*[4] (OGURI, 2006). O objetivo inicial de um classificador SVM é escolher os parâmetros

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

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

[4] Do inglês *support vectors*.

w e b que maximizem a distância entre esses hiperplanos. Para realizar essa otimização, o problema pode ser remodelado com multiplicadores de Lagrange $\{\alpha_i\}$, $1 \leq i \leq n$, levando à Equação 2.9. Busca-se, então, a minimização desta equação com relação a **w** e b e maximização com relação a $\{\alpha_i\}$, com todo $\alpha_i \geq 0$ (OGURI, 2006)

$$L(\alpha, b, \mathbf{w}) = \frac{1}{2}||\mathbf{w}||^2 - \sum_{i=1}^{n} \alpha_i \left[y_i(x_i \cdot \mathbf{w} + b) - 1 \right] \tag{2.9}$$

onde n é a cardinalidade do conjunto de treinamento e $||\mathbf{w}||$ é a norma euclidiana do vetor **w**. Após a obtenção dos valores de $\{\alpha_i\}$ que maximizam a Equação 2.9, 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)

$$y(x_j) = sinal\left(\sum_{i=1}^{n} \alpha_i y_i (x_i \cdot x_j) + b \right) \tag{2.10}$$

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)

$$y_i(x_i \cdot \mathbf{w} + b - 1) \geq 0, \; i = 1, ..., n \tag{2.11}$$

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 de folga[5] ε_i, uma para cada ponto (x_i, y_i). ε_i é igual a zero se $y(x_i) = y_i$ e igual a $|y_i - y(x_i)|$ em caso contrário (BISHOP, 2006). O SVM deve, neste caso, buscar o valor de **w** que minimize (OGURI, 2006)

$$C \sum_{i=1}^{n} \varepsilon_i + \frac{1}{2}||\mathbf{w}||^2 \tag{2.12}$$

onde C é um parâmetro responsável por controlar o compromisso entre a penalidade das variáveis de folga e a distância máxima entre os hiperplanos θ_1 e θ_{-1} (OGURI, 2006). Na modelagem com multiplicadores de Lagrange, a Equação 2.9 deve ser otimizada de tal forma que todo α_i deve ser maximizado obedecendo à restrição $0 \leq \alpha_i \leq C$. Desta forma, também se obtém a Equação 2.10 para determinação da classe de um novo documento.

Representações muito parecidas para documentos pertencentes a classes diferentes compro-

[5] Do inglês *slack variables*.

metem a determinação de hiperplanos θ_1 e θ_{-1} que conduzam a uma boa classificação. Quão mais diferentes forem as representações por classe, menor a probabilidade de que um SVM *se engane* na determinação da classe de um novo documento. Neste sentido, portanto, o SVM se aproxima bastante do Naïve Bayes. De fato, esse princípio permeia a noção de classificação em qualquer nível: dois elementos quaisquer devem ser suficientemente diferentes, em algum aspecto, para pertencerem a classes distintas.

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 3 e 4 envolvem apenas o Naïve Bayes. Apesar disso, SVMs são aplicados em boa parte dos trabalhos revisados no Capítulo 3 - por esse motivo, foi considerado 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**[6] é definida por (MANNING; RAGHAVAN; SCHUTZE, 2008)

$$\frac{\#documentos\ classificados\ corretamente}{\#total\ de\ documentos} \tag{2.13}$$

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)

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

A **rechamada**[7], medida também para uma classe $c \in C$ qualquer, é definida como (MANNING; RAGHAVAN; SCHUTZE, 2008)

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

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

A rechamada também evidencia o quanto o classificador está *acertando* - mas por classe. A **métrica F1**[8], também medida para uma classe $c \in C$ qualquer, é dada por (MANNING; RAGHAVAN; SCHUTZE, 2008)

$$2 \times \frac{precisao \times rechamada}{precisao + rechamada} \tag{2.16}$$

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**[9] (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**.

[8] Do inglês *F1-measure*.

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

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

O modelo *Labeled-Latent Dirichlet Allocation*, ou simplesmente **L-LDA**, fundamenta-se na ideia de que um documento pode tratar de múltiplos tópicos, refletidos nas palavras que o compõem (GRIFFITHS; STEYVERS, 2004). Esses tópicos, para efeito prático, podem ser assuntos, ideias ou pontos de vista. A finalidade desse modelo não é classificar documentos, mas sim evidenciar como as palavras se relacionam com seus tópicos. O L-LDA associa as palavras dos documentos a tópicos diferentes, com maior ou menor probabilidade, criando agrupamentos que compartilham uma semelhança semântica/temática.

Dado um conjunto de documentos D e um conjunto de tópicos T, O L-LDA interpreta um documento $d \in D$ como uma lista de palavras $\langle w_{d_1}, ..., w_{d_{|d|}} \rangle$, e o associa a uma lista binária representando a presença/ausência de cada tópico de T, $\langle l_1, ..., l_{|T|} \rangle$. Cada palavra w pertence ao vocabulário de D, V, e cada l funciona como um *bit*, assumindo os valores 0 ou 1 (RAMAGE et al., 2009). Antes de se executar o modelo, portanto, já se sabe quais dos tópicos de T se associam a cada documento.

O L-LDA representa cada documento d como uma mistura de seus tópicos, associando-o a um parâmetro θ_d. Esse parâmetro é amostrado de acordo com uma distribuição de probabilidade $Dirichlet(\alpha)$, onde α é um hiperparâmetro escolhido antes da execução do modelo. De forma análoga, cada tópico $t \in T$ é interpretado como uma mistura de palavras e se associa a um parâmetro ϕ_t. Esse parâmetro é amostrado de acordo com uma distribuição $Dirichlet(\beta)$, onde β é um hiperparâmetro também escolhido antes da execução do modelo (GRIFFITHS; STEYVERS, 2004). Em seguida, para cada palavra de d, um dos tópicos associados ao documento é amostrado de acordo com a distribuição de probabilidade $Multinomial(T_d, \theta_d)$. T_d é o subconjunto de T que corresponde exatamente aos tópicos relacionados a d. Por fim, considerando-se que o tópico escolhido foi um t_i, uma palavra $w \in V$ é amostrada de acordo com a distribuição $Multinomial(V, \phi_{t_i})$ (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 naquele contexto. As probabilidades associadas às escolhas de tópicos e palavras, neste cenário, são semelhantes àquelas apresentadas na seção 2.1 e, portanto, não serão exploradas na presente seção. Para um aprofundamento sobre o assunto, portanto, sugere-se consulta ao artigo de Ramage et al., no qual o L-LDA é proposto (RAMAGE et al., 2009). 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 amostrar um tópico t associado a d, dada uma distribuição θ_d, mais importante é o tópico no contexto desse documento (GRIFFITHS; STEYVERS, 2004).

A saída da execução de um L-LDA é uma tabela que indica quantas vezes cada palavra de D foi amostrada para cada tópico de T. Essa informação também pode ser obtida separadamente por documento, indicando como os tópicos segmentam seus conteúdos. Como o número de palavras associadas a cada tópico costuma ser bastante extenso, arbitra-se um valor k para cada um deles e, na prática, visualiza-se apenas as k palavras mais frequentemente associadas. No trabalho de Ramage et al., onde o L-LDA é proposto, os valores de k variam entre 9 e 12 (RAMAGE et al., 2009). Para os experimentos conduzidos nos Capítulos 3 e 4, fixa-se um mesmo k para todos os tópicos: 10.

O L-LDA foi proposto em 2009, mas até o momento já foi citado por pelo menos quinze outros trabalhos científicos[10]. Trata-se, portanto, de um modelo novo que está se popularizando. Analisando o objetivo desses trabalhos, concluiu-se que todos eles associam tópicos a assuntos. O artigo de Ramage et al., inclusive, associa tópicos a *tags* de *blogs*, como *books* ou *religion*. Nos Capítulos 3 e 4, tópicos representam pontos de vista, de modo que cada documento é rotulado com apenas um deles. Há também um outro tópico, neutro, associado a todos os documentos. Como ele se relaciona com todos os elementos de D, a tendência é que as palavras menos marcantes, por perspectiva, se associem mais frequentemente a ele. É válido ressaltar que esse uso do L-LDA não foi encontrado em **nenhum** trabalho revisado para esta monografia.

No L-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 L-LDA utilizada nos Capítulos 3 e 4 também utiliza Gibbs Sampling e está disponível no repositório *online* de Alexandre Passos[11]. O número de iterações para o modelo foi fixado em 100, valor suficiente para estabilizar as amostragens de tópicos e palavras.

[10] Essa verificação foi feita em 11 de Outubro de 2010, com o apoio do *Google Scholar*.
[11] http://github.com/alextp

3 CLASSIFICAÇÃO POR PERSPECTIVA: REVISÃO E DISCUSSÃO

A classificação de documentos de acordo com suas perspectivas é um tópico de pesquisa relativamente novo. Ele foi proposto como subproblema da área de Mineração de Opinião em 2008, pelas pesquisadoras Bo Pang e Lillian Lee (PANG; LEE, 2008). De acordo com elas, esse subproblema se diferencia da classificação por opinião por não configurar as classes como pólos (*positivo*, *negativo* etc.). De fato, a classificação por perspectiva visa a separar os documentos de um corpus de acordo com pontos de vista diferentes, como *pró-Israel* ou *pró-Palestina*. Embora, em certo nível, isso também polarize o corpus, já que normalmente os pontos de vista são antagônicos, dividi-lo de acordo com suas perspectivas é uma tarefa mais subjetiva do que aquela envolvendo opiniões. Um bom exemplo da diferença entre um ponto de vista e uma opinião, no contexto dos trabalhos revisados e de acordo com a *survey* de Pang e Lee, é o seguinte: *'Matrix' é um filme excelente* (opinião sobre um objeto específico, mais pontual) e *a paz mundial dificilmente será alcançada* (ponto de vista sobre um tema; algo que pode implicar em uma série de opiniões alinhadas com este pensamento). Apesar de ter sido proposto formalmente apenas em 2008, uma conferência importante para a área de Mineração de Opinião, a *AAAI*[1] *Conference*, lançou, na edição de 2010, uma trilha específica para trabalhos nesta direção: *Sentiment and Perspective* (**??**). Isso indica que o subproblema deve se fortalecer como área de pesquisa nos próximos anos.

A *survey* de Pang e Lee, uma das principais referências para este projeto, elenca três trabalhos que fazem classificação por perspectiva, publicados entre os anos de 2004 e 2006. Assim como nesse *survey*, esta monografia enfoca trabalhos recentes, publicados entre 2000 e 2010. A metodologia aplicada na busca por trabalhos a serem revisados pode ser resumida nos seguintes passos:

[1] *Association for the Advancement of Artificial Intelligence*.

1. Consideração dos três trabalhos indicados na *survey* de Pang e Lee;

2. Busca, no *Google Scholar*, por trabalhos citados nesses artigos;

3. Busca, no *Google Scholar*, por trabalhos que citam os artigos coletados nos itens anteriores;

4. Busca, na *ACL Anthology*[2] pelas palavras *perspective*, *viewpoint*, *politics*, *political*, *ideology* e *ideological*. Estas palavras foram escolhidas por (1) funcionarem como sinônimos nos artigos coletados anteriormente ou (2) pela relevância em suas temáticas, escritos por especialistas;

5. Busca nos *sites* dos eventos *EMNLP*[3], *NAACL*[4], *AAAI*[5] e *CoNNL*[6], realizados entre 2000 e 2010, pelas mesmas palavras do item anterior. Esses eventos foram selecionados pela relevância na área de Mineração de Opinião.

Nem todos os trabalhos coletados, de acordo com essa metodologia, tinham necessariamente a ver com classificação por perspectiva. Alguns, como a tese de Alice Oh, se propõem modelar computacionalmente o conceito de perspectiva (OH, 2008); outros, como o artigo de Laver et al., se propõem a criar uma escala ideológica e comparar documentos de acordo com ela (LAVER; BENOIT; COLLEGE, 2003). Por este motivo, foi feita uma filtragem que resultou em 14 trabalhos a serem revisados. Na seção 3.1, essa revisão é apresentada. Quase todos os trabalhos consideram, ainda que não exclusivamente, as contagens de palavras dos documentos na classificação. Além disso, eles divergem bastante quanto à escolha de outras características dos documentos. Por estes motivos, a seção 3.2 discute a relação entre a classificação baseada em contagens de palavras e o emprego delas por diferentes perspectivas. Por fim, na seção 3.3, apresentam-se conclusões sobre os conteúdos expostos nas seções 3.1 e 3.2.

3.1 TRABALHOS REVISADOS

Todos os trabalhos revisados nesta seção apresentam uma metodologia em comum: eles organizam um corpus de documentos, definem em que perspectivas ele se divide, decidem como os documentos serão pré-processados e, em seguida, classificam-nos utilizando alguma

[2] A *Association of Computational Linguistics*, *ACL*, mantém o maior arquivo digital envolvendo trabalhos de Linguística Computacional, o que inclui Mineração de Opinião (http://aclweb.org/anthology-new/).

[3] *Empirical Methods in Natural Language Processing Conference.*

[4] *North American Chapter of the Association for Computational Linguistics Conference.*

[5] *Association for the Advancement of Artificial Intelligence Conference.*

[6] *Computational Natural Language Learning Conference.*

técnica - na maioria dos casos, SVMs ou Naïve Bayes. Esses classificadores sempre se baseiam em características dos documentos para determinar suas classes. Na maioria das vezes, essas características são contagens de palavras ou a ausência/presença de cada uma delas. Em alguns outros casos, elas consistem na semântica das palavras ou em relações sintáticas presentes nos documentos. O uso de características semânticas ou sintáticas pode dificultar a adaptação de um trabalho para *datasets* em línguas diferentes, enquanto a simples consideração de contagens ou ausência/presença de palavras independe da língua do corpus. Essa divisão é particularmente interessante para essa monografia, que conduz um estudo de caso com um corpus em português. Por estes motivos, essa revisão divide-se em trabalhos que exploram contagens ou presença de palavras, na subseção 3.1.1, e trabalhos que exploram características semânticas ou sintáticas, na subseção 3.1.2. Por fim, na subseção 3.1.3, é apresentada uma tabela que sintetiza todos os trabalhos, evitando uma visão fragmentada de cada um deles.

3.1.1 TRABALHOS QUE EXPLORAM CONTAGEM OU PRESENÇA DE PALAVRAS

O trabalho de **Lin et al.** classifica artigos do *site* Bitterlemons[7] como pró-Palestina ou pró-Israel (LIN et al., 2006). Inicialmente, os documentos são representados como listas de palavras reduzidas a seus radicais. Isso significa, por exemplo, que as palavras *political* e *politics* são representadas através de um mesmo termo, o radical *politic*. Os classificadores Naïve Bayes e SVM são então aplicados, explorando as contagens desses radicais em cada documento. A taxa de acerto obtida com o SVM variou entre 81.48% e 97.24%; com o Naïve Bayes, ela variou entre 84.85% e 99.09%. Essas variações advêm de diferentes divisões entre os conjuntos de treinamento e teste. Por fim, o trabalho de Lin et al. propõe o classificador *Latent Sentence Perspective Model*, ou simplesmente LSPM. Ele consiste em uma variação do Naïve Bayes que, em vez de considerar documentos como listas de palavras, os representa como listas de senteças. A hipótese assumida por ele é a seguinte: nem todas as sentenças carregam um ponto de vista, de modo que o classificador, antes de definir a classe de um documento, deve selecionar quais delas carregam palavras relevantes. A taxa de acerto obtida com o LSPM variou entre 86.99% e 94.93%, também como consequência de diferentes divisões entre os conjuntos de treinamento e teste. Para essas divisões, as taxas de acerto obtidas com o Naïve Bayes foram de 84.85% e 93.46%. O SVM não foi comparado diretamente com o LSPM. Apesar do desempenho superior a do Naïve Bayes, nenhum outro trabalho revisado faz uso desse classificador.

Mullen e Malouf propõem dois trabalhos que analisam um conjunto de *posts* do fórum

[7] http://bitterlemons.org/

de discussão política Politics[8]. No **primeiro trabalho**, eles tratam cada documento como a união de todos os *posts* de um mesmo usuário (MULLEN; MALOUF, 2006). Cada documento é representado como uma lista de palavras, e um Naïve Bayes é aplicado considerando suas contagens em cada documento. A taxa de acerto obtida, via validação cruzada de dez dobras, foi de 60.37%. O tamanho do *dataset* (apenas 185 documentos) e o uso parecido de palavras por ambas as ideologias foram apontados como alguns dos principais motivos para a obtenção desse resultado. Mullen e Malouf também sugerem que a presença de documentos menores, correspondentes a usuários do fórum que raramente postam, pode contribuir negativamente para o desempenho do Naïve Bayes. Uma última observação desse trabalho 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*. 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.

Essa forma de interação entre os participantes do fórum é explorada pelo **segundo trabalho de Mullen e Malouf** (MULLEN; MALOUF, 2008), com o intuito de melhorar a classificação. Para isto, cria-se um grafo de co-citação em que cada vértice representa um participante e cada citação em uma fala a outra é indicada por uma aresta entre seus autores. Os participantes são agrupados de acordo com seus padrões de citação e, em seguida, as falas de cada grupo obtido são tratadas como um único documento. Aplica-se um Naïve Bayes a esta nova coleção de documentos, também explorando apenas suas contagens de palavras, e os resultados obtidos são propagados para todos os participantes de cada grupo. Essa metodologia atinge resultados significativamente melhores do que aqueles obtidos no trabalho anterior: para participantes com mais de 500 palavras de fala no fórum, a taxa de acerto relatada é de 73%. É válido salientar que, muito provavelmente, o sucesso dessa técnica é consequência do fato de que usuários alinhados ideologicamente não se citam muito. Se, além de citar seus oponentes, eles também se citassem bastante, os padrões de citação seriam sempre muito parecidos, inviabilizando os agrupamentos adequados no grafo.

O trabalho de **Durant e Smith** constrói um corpus sobre os posicionamentos do ex-Presidente George W. Bush quanto à Guerra do Iraque (DURANT; SMITH, 2006). O objetivo desse trabalho é classificar *posts* de diversos *blogs* políticos de acordo com os pontos de vista pró Guerra do Iraque e anti Guerra do Iraque. Os classificadores Naïve Bayes e SVM foram aplicados, considerando apenas a presença/ausência de palavras nos documentos. A taxa de acerto obtida com um SVM foi de 75.47%. Com o Naïve Bayes, ela foi de 78.06%. Os autores, em seguida,

[8] http://politics.com

investigam se a seleção de apenas parte das palavras pode melhorar a classificação. Eles aplicam uma técnica denominada *CfsSubsetEval*, disponível na ferramenta WEKA 3.4[9], que busca um subconjunto de palavras que maximize a taxa de acerto obtida. As palavras devem ter alta correlação com as classes escolhidas, mas baixa correlação entre si. Aplicando os classificadores SVM e Naïve Bayes, e considerando apenas os subconjuntos selecionados pela técnica, as taxas de acerto obtidas foram de 87.66% e 89.77%, respectivamente. É válido ressaltar que a técnica *CfsSubsetEval* pode trazer uma melhoria de desempenho significativa para *datasets* escritos em qualquer língua.

O trabalho de **Hirst, Riabinin e Graham** constrói dois *datasets*, um em inglês e outro em francês, que consistem de discursos de congressistas canadenses nas reuniões parlamentares (HIRST; RIABININ; GRAHAM, 2010). Além de classificar esses discursos como liberais ou conservadores, esse trabalho investiga se as classes correspondem de fato a ideologias diferentes ou, simplesmente, a expressões de ataque e defesa. Inicialmente, os autores analisam o 36^o parlamento canadense, período em que um partido liberal estava com a maioria no poder. Excluindo apenas as palavras menos frequentes do corpus, e aplicando um SVM baseado apenas nas contagens de palavras dos documentos, as taxas de acerto obtidas foram de 83.8% e 75.5% para os corpora inglês e francês, respectivamente. Em seguida, os autores selecionaram o 39^o Parlamento, período em que os partidos liberais eram oposição, para verificar o que de fato estava determinando a classificação. Treinando com documentos do 36^o Parlamento e testando com aqueles do 39^o, a classificação entre liberal e conservador é insatisfatória: as taxas de acerto foram de 44.9% e 45.7% para os corpora inglês e francês respectivamente. Treinando com o 39^o e testando com o 36^o, as taxas de acerto foram ainda piores: 36.8% e 35.2% para os corpora inglês e francês respectivamente. Diante disso, os autores concluem que que as ideologias liberal e conservadora não são exploradas adequadamente pelos discursos. Se o fossem, e considerando que elas não mudaram significativamente de um parlamento para o outro, estes últimos experimentos apresentariam resultados melhores. Eles indicam que os discursos envolvem muitas expressões de ataque e defesa, que mudam de partido conforme eles se alternam no poder. Por este motivo, os conjuntos de treinamento e teste apresentam padrões de ataque e defesa invertidos, implicando no mau desempenho da classificação. Esse trabalho alerta, portanto, para a definição adequada de que perspectivas estão contidas no corpus. Neste caso, em vez de liberal e conservadora, seria mais adequado utilizar pró-governo e anti-governo, ou situação e oposição.

O trabalho de **Klebanov, Beigman e Diermeier** avalia o desempenho dos classificadores Naïve Bayes e SVM, utilizando para isso características diferentes dos documentos. Para o

[9] http://www.cs.waikato.ac.nz/ml/weka/

Naïve Bayes, os autores comparam a escolha entre (1) presença/ausência de palavras e (2) contagens de palavras. Para o SVM, as comparações envolvem a escolha entre (1), (3) suas contagens normalizadas em relação ao documento (frequências) e (4) suas frequências ponderadas em relação aos outros documentos do corpus. É válido ressaltar que todas essas características são apenas variações de contagens de palavras. Os autores selecionaram quatro *datasets* para comparar essas escolhas: o primeiro envolve debates sobre o aborto, dividido entre pró-escolha e pró-vida; o segundo consiste em artigos sobre a pena de morte, dividido entre pró pena de morte e anti pena de morte; o terceiro é composto de artigos sobre o conflito Israel-Palestina, escritos por convidados do *site* Bitterlemons e estudados por Lin et al. ; o quarto também é composto de documentos desse *site*, cada um escrito por um especialista diferente. Esses dois últimos, assim como no trabalho de Lin et al., são divididos entre pró-Palestina e pró-Israel. No caso do Naïve Bayes, o uso de (1) se mostrou melhor em alguns *datasets* e o de (2), em outros. A maior diferença de desempenho observada está relacionada com o corpus de pena de morte: com (1), a taxa de acerto obtida foi de 88%; com (2), 93%. Os autores indicam, diante disso, que essa escolha não é tão relevante para problemas de classificação por perspectiva. Quanto ao SVM, a escolha por (1) se mostrou superior ou equivalente às outras em todos os casos. A maior diferença de desempenho observada também está associada ao corpus de pena de morte: enquanto (1) conduz a uma taxa de acerto de 83%, (3) resulta em 82% e (4) em 73%. Esse trabalho, assim como o de Durant e Smith, também investiga o uso de um subconjunto de palavras. Com o apoio do *toolkit* WEKA, os autores selecionam subconjuntos pequenos, contendo entre 100 e 500 palavras, e indicam que a classificação dos *datasets*, nestes casos, conserva seu desempenho. As taxas de acerto não aumentam, como no trabalho de Durant e Smith, mas também não diminuem.

3.1.2 TRABALHOS QUE EXPLORAM OUTRAS CARACTERÍSTICAS DOS DOCUMENTOS

O trabalho de **Jiang e Argamon** constrói um corpus de documentos extraídos de *blogs* sobre política escritos em inglês (JIANG; ARGAMON, 2008). Cada *blog* é associado à perspectiva liberal ou conservadora, divisão explorada na classificação. Os documentos que constituem o corpus são apenas as páginas iniciais de cada um deles. Inicialmente, Jiang e Argamon aplicam um SVM às páginas, considerando apenas a presença/ausência de palavras nos documentos. A taxa de acerto obtida foi de 81.92% e, considerando ambas as classes, tem-se que a precisão média foi de 81.76%, a rechamada média foi de 81.93% e a métrica F1 média foi de 81.79%. Em seguida, as páginas foram reduzidas a suas sentenças subjetivas, através de uma

seleção baseada no dicionário semântico *General Inquirer*[10]. Para uma sentença ser escolhida, ela deveria conter pelo menos duas palavras categorizadas, segundo o dicionário, como pertencentes às categorias *Dor*, *Hostilidade*, *Prazer* ou *Virtude*, dentre outras. A taxa de acerto, neste cenário, subiu para 83.28%. As precisão, rechamada e métrica F1 médias foram de, respectivamente, 83.26%, 83.38% e 83.24%. Esse trabalho, por fim, busca extrair as expressões opinativas que mais se relacionam aos pontos de vista liberal e conservador, baseando-se nas sentenças subjetivas e em consultas ao dicionário *General Inquirer*. O SVM é aplicado considerando a presença/ausência de palavras e, adicionalmente, a presença/ausência das expressões opinativas separadas para cada lado. A taxa de acerto neste cenário foi de 84.96%. As precisão, rechamada e métrica F1 médias foram de, respectivamente, 83.24%, 83.48% e 83.27%. As melhorias em relação ao uso exclusivo de ausência/presença de palavras, portanto, são pequenas. Além disso, essas metodologias não são facilmente adaptáveis a línguas que não dispõem de dicionários como o *General Inquirer online*.

O trabalho de **Greene e Resnik** enfoca no estudo de um corpus sobre a pena de morte (GREENE; RESNIK, 2009). Esse corpus é posteriormente analisado no trabalho de Klebanov, Beigman e Diermeier, apresentado na seção anterior. O objetivo do trabalho é classificá-lo de acordo com as perspectivas pró pena de morte e anti pena de morte, e a metodologia desenvolvida baseia-se na hipótese de que existe uma conexão entre a estrutura de uma sentença e seu ponto de vista. Neste sentido, os autores selecionam um conjunto de verbos relevantes no corpus, como *kill* e *murder*, e criam representações para todos os termos que se relacionam sintaticamente com eles. Essas representações consistem em tuplas que associam cada termo a seu papel sintático na frase. Esses papéis, por sua vez, podem ser *sujeito*, *objeto direto*, *verbo transitivo*, dentre outros. A determinação desses papéis foi feita com o apoio do programa *Stanford Parser*[11], adaptado à língua inglesa. Greene e Resnik então aplicam um SVM nos documentos reduzidos a essas representações, atingindo taxas de acerto de 82.09% e 88.10%, a depender da escolha de verbos. Em vez de uma lista de palavras, portanto, o SVM é aplicado a uma lista de tuplas. Para comparar sua metodologia, os autores aplicam o SVM a documentos representados como sequências de duas palavras (bigramas), todas reduzidas a seus radicais. Nesse caso, as taxas de acerto obtidas foram de 68.37% e 71.96%, também a depender da escolha dos verbos. Essa metodologia, portanto, traz uma melhoria muito significativa para a classificação desse corpus. Apesar disso, aplicando a mesma metodologia ao corpus sobre o conflito Israel-Palestina, estudado por Lin et al., Greene e Resnik não obtêm nenhuma melhoria significativa. Em alguns cenários, os resultados obtidos por Lin et al. são ligeiramente melhores; em outros,

[10] http://www.wjh.harvard.edu/ inquirer/

[11] http://nlp.stanford.edu/software/lex-parser.shtml

são aqueles apresentados por estes autores.

O trabalho de **Efron** constrói dois *datasets*: um envolvendo artigos sobre a política dos Estados Unidos e outro composto de textos sobre artistas musicais (EFRON, 2004). O primeiro corpus é classificado de acordo com as perspectivas direita e esquerda e o segundo, de acordo com as orientações alternativa ou popular. Inicialmente, Efron classifica os dois corpora com um Naïve Bayes e um SVM. O autor não especifica se são utilizadas contagens de palavras ou alguma outra característica, mas afirma que os dois métodos se baseiam apenas nas palavras contidas nos textos. Com o Naïve Bayes, as taxas de acerto obtidas nestes corpora foram de, respectivamente, 64.71% e 50.1%. Para o primeiro corpus, a taxa de acerto obtida com um SVM foi de 72.96%; quanto ao segundo, não foram realizados experimentos com o SVM por carência de recursos computacionais. A fim de melhorar as taxas de acerto, Efron desenvolve uma classificação baseada em citações que não envolve contagens de palavras nem citações de um documento a outro, mas sim suas semelhanças temáticas. Na prática, Efron determina a perspectiva de cada documento de acordo com a probabilidade deles serem co-citados com documentos fixados como referência para cada perspectiva. Essa probabilidade é estimada a partir dos números de documentos retornados pelo buscador AltaVista[12] quando dois documentos são buscados, indicando co-citação entre eles. Essa metodologia de classificação, aplicada ao primeiro *dataset*, resultou em uma taxa de acerto de 94.1%. Quanto ao segundo, as taxas de acerto obtidas foram de 82.18% e 88.84%. No primeiro caso, todos os documentos foram considerados; no segundo, os menos co-citados foram descartados. Apesar de simples e aparentemente eficiente, essa metodologia apresenta uma séria limitação: se o corpus for composto de documentos pouco citados na Web, as classes podem ser estimadas erroneamente com uma alta frequência, não trazendo nenhuma melhoria significativa à classificação.

O trabalho de **Thomas, Pang e Lee** se propõe a determinar os pontos de vista de congressistas dos Estados Unidos quanto a novas leis (THOMAS; PANG; LEE, 2006). O corpus é dividido entre as perspectivas suporte e oposição em relação a novas leis, e cada documento corresponde, a princípio, a uma fala de um congressista. Inicialmente, cada texto é classificado de forma isolada, através da aplicação de um SVM que considera a presença/ausência de palavras em cada um deles. As taxas de acerto obtidas foram de 66.05% e 70.04%, a depender do subconjunto de documentos classificado. Tratando cada documento como a concatenação de todas as falas de um mesmo congressista, as taxas obtidas com o SVM, considerando as mesmas características, foram de 70.00% e 71.60%. A fim de melhorar essas taxas de acerto, os autores comparam trechos de documentos e determinam valores positivos que representam o quanto eles concordam. Quando esses valores são menores do que um determinado θ fixado,

[12] http://altavista.com/

considera-se que não há indícios suficientes de que os documentos concordam; eles são, então, reduzidos a zero. Adicionalmente, os autores aproveitam os experimentos com o SVM para determinar o grau de preferência para classificar cada documento como suporte ou oposição. A classe escolhida para cada um deles deve ser aquela que favorece o seguinte cenário: (1) ela não deve, idealmente, ser a rejeitada pelo SVM e (2) documentos que concordam muito não devem ser associados a classes diferentes. As taxas de acerto obtidas, considerando cada documento como uma fala, variaram entre 70.81% e 89.11%, a depender do subconjunto classificado e do valor de θ. Considerando cada documento como a concatenação de todas as falas de um mesmo congressista, as taxas de acerto obtidas variaram entre 71.28% e 88.72%, também a depender do subconjunto classificado e do valor de θ. A determinação errônea de concordâncias entre documentos pode prejudicar significativamente a classificação. Além disso, de acordo com os próprios autores, a adoção dessa metodologia só traz benefícios quando há artigos difíceis de classificar individualmente.

O trabalho de **Bansal, Cardie e Lee** propõe uma extensão ao trabalho de Thomas, Pang e Lee, considerando também a *discordância* entre documentos (BANSAL; CARDIE; LEE, 2008). A hipótese assumida em ambos os trabalhos é a mesma: se é difícil determinar a classe de um documento *x*, mas sabe-se que ele concorda *ou discorda* de um documento *y*, fácil de classificar, a determinação da classe de *x* também se torna mais fácil. O trabalho apresenta algumas estratégias para determinação dos valores de discordância e, em seguida, compara seus usos com a metodologia proposta por Thomas, Pang e Lee. Bansal, Cardi e Lee também lidam com valores negativos, associados à discordância. Na prática, é preciso mapear esses valores negativos em positivos, pois eles dificultam o problema de otimização envolvido na determinação das classes dos documentos. Esses mapeamentos são justamente o que diferencia uma estratégia de outra. Utilizando o mesmo *dataset* analisado por Thomas, Pang e Lee, os autores deste trabalho obtêm resultados superiores ou equivalentes àqueles que consideram apenas a concordância entre documentos, para a maioria das estratégias testadas. As taxas de acerto, entretanto, não são informadas explicitamente neste trabalho, sendo representadas através de um gráfico comparativo.

3.1.3 COMPARAÇÕES

3.2 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 **??**.

O primeiro *dataset* estudado é o mesmo discutido na seção **??**, 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 **??**, 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[13], é um dos três que não utilizam apenas contagens de palavras para a classificação, sendo detalhado no Capítulo **??**. 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%[14].

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

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

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

ajuda a identificar palavras muito comuns nos *datasets*, independentemente de perspectiva. Essa é a diferença fundamental entre essa aplicação do L-LDA[15] 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 3.2. 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 3.1: 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

> *"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

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

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 3.2: 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 3.2 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 3.2 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 3.2 foram representadas graficamente com o apoio do *software* wordle[16]. 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.

3.3 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 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

[16] http://wordle.net/

taxa de acerto, considerando apenas contagens de palavras. Outros três trabalhos, apresentados no capítulo **??**, 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[17].

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 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 docu-

[17] Verificar **ANEXO BLA**.

mentos 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.

4 *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 4.1, a construção do corpus é apresentada - desde a seleção dos veículos até o pré-processamento dos artigos; na seção 4.2, 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 4.3, o modelo de tópicos L-LDA é aplicado ao corpus, evidenciando aspectos da linguagem explorada por artigos com posicionamentos diferentes; por fim, na seção 4.4, são apresentadas conclusões sobre o estudo e possíveis extensões para ele.

4.1 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 4.1: 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.

4.2 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 **??** 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.

4.3 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 4.2: 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 4.1, 4.2 e 4.3. O tamanho das palavras nas imagens corresponde ao quanto elas se associam a cada tópico[23]. As imagens 4.1 e 4.2 evidenciam o destaque dado aos políticos Lula, Dilma Rousseff e José Serra nos artigos analisados. A imagem 4.3 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 4.1: Representação gráfica para as trinta palavras mais frequentemente associadas ao tópico pró-governo.

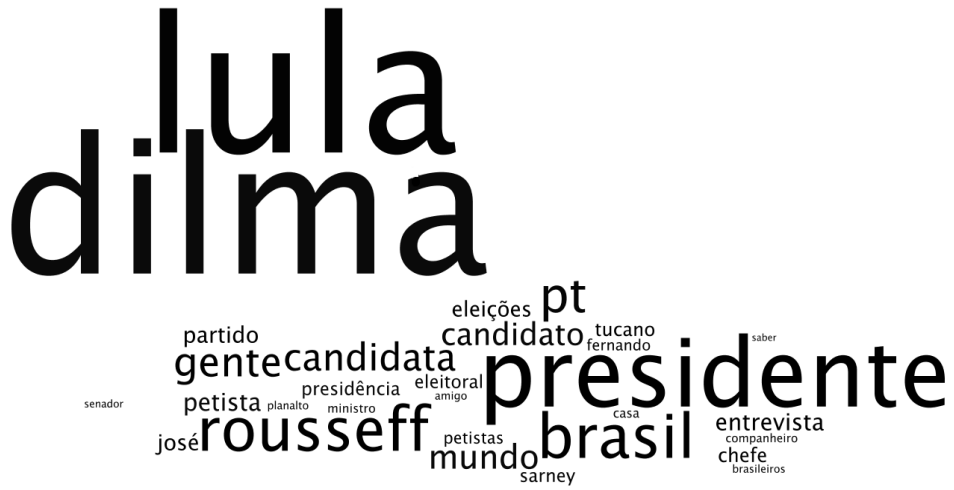

Figura 4.2: Representação gráfica para as trinta palavras mais frequentemente associadas ao tópico anti-governo.

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 4.3: 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

4.4 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 3.

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 4.1 e 4.2, 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*.

5 CONCLUSÃO

5.1 DIFICULDADES ENCONTRADAS

5.2 TRABALHOS FUTUROS

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.

BANSAL, M.; CARDIE, C.; LEE, L. The power of negative thinking: Exploiting label disagreement in the min-cut classification framework. In: *Proceedings of the International Conference on Computational Linguistics (Poster paper)*. [S.l.: s.n.], 2008.

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.

COMSCORE; KELSEYGROUP. *Online consumer-generated reviews have significant impact on offline purchase behavior*. Press Release, 2007. Disponível em: <http://www.comscore.com/press/release.asp?press=1928>.

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.

GREENE, S.; RESNIK, P. More than words: Syntactic packaging and implicit sentiment. In: *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. [S.l.: s.n.], 2009. p. 503–511.

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.

JIANG, M.; ARGAMON, S. Political leaning categorization by exploring subjectivities in political blogs. In: *Proceedings of the 4th International Conference on Data Mining (DMIN 2008)*. [S.l.: s.n.], 2008. p. 647–653.

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.

LAVER, M.; BENOIT, K.; COLLEGE, T. Extracting policy positions from political texts using words as data. *American Political Science Review*, p. 311–331, 2003.

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.

LIU, B. Opinion mining. In: LIU, B. (Ed.). *Web Data Mining: Exploring Hyperlinks, Contents and Usage Data*. [S.l.]: Springer, 2006. ISBN 3540378812.

LIU, B. Sentiment analysis and subjectivity. In: INDURKHYA, N.; DAMERAU, F. J. (Ed.). *Handbook of Natural Language Processing, Second Edition*. [S.l.]: CRC Press, Taylor and Francis Group, 2010. ISBN 978-1420085921.

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.

OH, A. *Generating multiple summaries based on computational model of perspective*. Tese (Doutorado) — Massachusetts Institute of Technology (MIT), 2008.

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

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.