UNIVERSIDADE FEDERAL DA BAHIA

INSTITUTO DE MATEMÁTICA

DEPARTAMENTO DE CIÊNCIA DA COMPUTAÇÃO

Aline Duarte Bessa

Classificação de documentos de acordo com pontos de vista sobre um tema

Aspectos teóricos e práticos

Salvador

2010

Aline Duarte Bessa

Classificação de documentos de acordo com pontos de vista sobre um tema

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

AGRADECIMENTOS

À UFBA, por viabilizar esta monografia.

A meu orientador e amigo Alexandre Passos, por me ajudar a definir o tema, me motivar e participar ativamente de todo o processo.

A meu co-orientador Luciano Porto, pelas sugestões e por acompanhar esta monografia do começo ao fim.

À minha mãe, por ser meu baluarte durante todos esses anos. A meu pai, por me fazer ver tudo de forma mais simples.

A Andrea Bessa e Lucas Cunha, jornalistas que me ajudaram a selecionar bons *sites* de política brasileira, fundamentais para essa monografia. E a Marcelo Pessoa, por gostar tanto de política e analisar esse trabalho sob a ótica de alguém não-envolvido.

A todos os meus amigos, pelos momentos maravilhosos nos fins de semana e compreensão com meus sumiços.

RESUMO

A classificação de documentos de acordo com seus pontos de vista sobre um tema é um tópico de pesquisa que vem recebendo atenção crescente. Dado um conjunto de documentos sobre a política dos Estados Unidos, por exemplo, esse tipo de classificação pode ser empregado para identificar quais deles defendem o ponto de vista *liberal* e quais defendem o ponto de vista *conservador*. Esse tipo de classificação colabora com a análise de documentos argumentativos, especialmente quando não é possível lê-los um por um. Devido ao caráter desses documentos, esse tipo de classificação pode fazer parte de estudos interdisciplinares, envolvendo áreas como Ciência Política e Comunicação.

Diante disso, essa monografia apresenta uma revisão de trabalhos que tratam desse tópico, indicando quais são os principais classificadores utilizados, quais aspectos dos documentos podem ser explorados na identificação de seus pontos de vista, dentre outros aspectos associados a esse tópico de pesquisa. Em seguida, é feito um estudo de caso envolvendo documentos sobre a política brasileira escritos em 2010, ano de eleição presidencial. Eles são classificados de acordo com os pontos de vista *pró-oposição* e *pró-situação*, e as palavras mais enfatizadas por cada ponto de vista também são discutidas.

Palavras-chave: Mineração de opinião, classificação, política brasileira, ponto de vista.

ABSTRACT

Documents classification according to their viewpoints is a research topic that has been receiving a growing attention. Given a set of documents about the United States politics, for example, this kind of classification may be useful to identify those that defend a *liberal* viewpoint and those that defend a *conservative* one. This kind of classification is helpful in analyzing argumentative documents, specially when it is not possible to read them one by one. Due to the content of these documents, this kind of classification may be part of interdisciplinary studies, involving areas such as Political Sciences or Communication.

In the view of this, this monograph presents a review of papers that discuss this topic, indicating which classifiers are the most used, which aspects of the documents may be exploited in the identification of their viewpoints, among other aspects associated to this research topic. After that, a case study is conducted, involving documents about brazilian politics written in 2010, year of brazilian presidential election. They are classified according to the *pro-situation* and *pro-opposition* viewpoints, and the words that are more emphasized by each viewpoint are also discussed.

Keywords: Opinion mining, classification, brazilian politics, viewpoint.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

Antes de assistir a um filme ou comprar um certo produto, é comum ler críticas a seu respeito ou considerar os comentários de outras pessoas. De forma semelhante, para se formular um posicionamento sobre algum tema - como aborto ou pena de morte, por exemplo - é comum levar em conta os pontos de vista de outros indivíduos. Com a disseminação da Web, a busca por textos opinativos ou argumentativos sofreu um aumento expressivo, passando a fazer parte do cotidiano de seus usuários. Naturalmente, isso ampliou os insumos para a formação de opiniões e posicionamentes por parte desses usuários, causando um impacto tanto no processo de consumo quanto na defesa de pontos de vista a respeito dos mais diversos assuntos. Uma pesquisa realizada nos Estados Unidos reforça essa tese (COMSCORE; KELSEYGROUP, 2007), revelando que as opiniões expressas em resenhas *online*, como críticas de restaurantes e albergues, influenciaram fortemente as decisões de 73 a 87% dos consumidores.

Diante disso, estudos com o intuito de extrair opiniões e pontos de vista da Web, e interpretá-los 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** (PANG; LEE, 2008; LIU, 2006). De acordo com a *survey* de Pang e Lee, a Mineração de Opinião é 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 a um único assunto, classifique-o como expressando uma opinião estritamente positiva, estritamente negativa, ou em algum grau bem definido entre esses dois extremos;

2. **Detecção de subjetividade e identificação de opinião -** Dado um documento, detecte se ele é subjetivo ou não, i.e. se consiste de opiniões ou de fatos, ou que partes dele são subjetivas;

3. **Análise de tópico-sentimento -** Dado um documento opinativo, assume-se que suas opi-

niões podem se referir a tópicos diferentes, e deve-se identificar quais opiniões são sobre quais tópicos;

4. **Pontos de vista ou perspectivas -** Dado um documento opinativo, que apresente um ponto de vista[1] sobre um tema, em vez de um sentimento polarizado sobre um único assunto, classifique-o de acordo com esse ponto de vista;

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 (1) tem recebido mais destaque na prática, funcionando como base para ferramentas de monitoramento de conteúdo *online*[2]. Essas ferramentas se popularizaram na área de *Marketing*, sendo utilizadas para avaliar como pessoas públicas, produtos e marcas são comentados na Web. Basicamente, elas fixam um produto, marca ou pessoa pública e classificam as opiniões sobre ele como positivas ou negativas. Dentre os outros itens, o (4) tem recebido atenção crescente como área de pesquisa.

Uma diferença fundamental entre esses dois itens diz respeito à natureza das classes: enquanto, no item (1), elas correspondem normalmente aos pólos positivo ou negativo, no item (4) elas correspondem a diferentes pontos de vista associados a um mesmo tema. Neste último caso, a ideia é classificar os documentos de acordo com esses pontos de vista, que se associam às crenças e atitudes de seus autores em relação a um conjunto de múltiplos assuntos (uma temática) (PANG; LEE, 2008). Não se trata, portanto, de classificar conteúdo como estritamente positivo ou negativo, mas sim como alinhado à defesa de um determinado posicionamento, como *pró-aborto* ou *neoliberal*.

Embora a diferença entre opiniões e pontos de vista seja um tanto subjetiva, no contexto da Mineração de Opinião as opiniões têm sido associadas a eventos pontuais, envolvendo marcas/produtos/pessoas públicas; os pontos de vista, por sua vez, são relacionados a temáticas de cunho social, envolvendo argumentações que revelam ideologias e crenças. O item (4), portanto, pode ser aplicado à compreensão de como as pessoas, na Web, argumentam em defesa dessas ideologias e crenças. Neste sentido, esse item ataca computacionalmente um problema que já era estudado em outras áreas, como Comunicação e Ciências Políticas (GENTZKOW; SHAPIRO, 2006; GROSECLOSE; MILYO, 2005; FADER et al., 2007): a investigação dos

[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.

[2] O TwitterSentiment (http://twittersentiment.appspot.com/) e o Moodviews (http://moodviews.com/) são dois exemplos desse tipo de ferramenta.

pontos de vista contidos em textos. Diante disso, e da crescente atenção que este item vem recebendo, ele foi escolhido como o enfoque dessa monografia.

Essa monografia apresenta uma revisão bibliográfica da área de classificação por ponto de vista, abordada superficialmente na *survey* de Pang e Lee, acompanhada de um estudo de caso. Inicialmente, é feita uma descrição básica dos principais classificadores utilizados na área, no Capítulo 2. Eles são o Naïve Bayes e os *Support vector machines* (SVMs), muito populares na área de Aprendizado de Máquina. Neste capítulo também são apresentadas métricas para se avaliar o desempenho de uma classificação e uma técnica que valida essa avaliação. Ainda nesse capítulo, por fim, é apresentado o modelo *Labeled-latent dirichlet allocation* (L-LDA), que associa documentos a tópicos e relaciona suas palavras a cada um deles. No Capítulo 3, trabalhos sobre classificação por ponto de vista são selecionados, revisados e analisados de forma comparativa. 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 e a discussão do uso de palavras por cada ponto de vista. Por fim, no Capítulo 5, são apresentadas conclusões a respeito dos conteúdos explorados pelos Capítulos 3 e 4. Esse capítulo também discute as principais dificuldades encontradas nessa monografia e trabalhos futuros. As próximas seções apresentam, respectivamente, os objetivos dessa monografia e alguns trabalhos relacionados.

1.1 OBJETIVO

O objetivo dessa monografia é explorar aspectos teóricos e práticos da classificação de documentos de acordo com seus pontos de vista. Para a parte teórica, foi elaborada uma revisão de artigos que tratam do assunto. Para a parte prática, foi proposto um estudo de caso envolvendo um corpus de política brasileira, explorando todos os passos envolvidos na sua classificação e uma análise de como as palavras são enfatizadas por cada ponto de vista.

1.2 TRABALHOS RELACIONADOS

Esta monografia apresenta uma revisão e um estudo de caso sobre classificação de documentos de acordo com seus pontos de vista. Quanto à revisão, o principal trabalho relacionado é a *survey* de Pang e Lee, que propõe a classificação de documentos por ponto de vista como um subproblema da área de Mineração de Opinião (PANG; LEE, 2008). Esse trabalho apresenta

[3]Nesta monografia, os termos *corpus* e *dataset* serão utilizados de forma intercambiável, e se referem a um conjunto de documentos de texto.

uma boa introdução à área de Mineração de Opinião, elencando aplicações para áreas como *Marketing* e Sociologia, dividindo-a em subproblemas e descrevendo os principais desafios associados a cada um deles. Dada a abrangência da *survey*, entretanto, pouca atenção é dedicada à temática dessa monografia. Bing Liu também propõe alguns materiais introdutórios à Mineração de Opinião (LIU, 2006, 2010). O enfoque desses trabalhos, entretanto, é identificar opiniões em documentos e classificá-los como *positivos* ou *negativos* com base nessas opiniões. Por este motivo, os trabalhos de Bing Liu funcionam como um material de apoio a essa monografia, mas não se relacionam diretamente com sua temática.

Os trabalhos relacionados ao estudo de caso consistem na revisão em si, apresentada no Capítulo 3. São diversos artigos que se propõem a classificar documentos de acordo com seus pontos de vista, utilizando uma mesma metodologia básica. No que diz respeito à temática explorada no estudo de caso, a política brasileira, a ferramenta Eleitorando[4] se aproxima desta monografia. A sua finalidade, entretanto, é monitorar opiniões sobre candidatos às Eleições 2010 nas redes sociais, como Twitter[5] ou YouTube[6], e classificá-las como *positivas* ou *negativas*. Não se trata, portanto, do enfoque proposto pela classificação por ponto de vista, que busca identificar o posicionamento defendido em um documento, em vez de opiniões polarizadas.

[4] http://www.eleitorando.com.br/site/
[5] http://twitter.com/
[6] http://br.youtube.com/

2 TÉCNICAS BÁSICAS PARA CLASSIFICAÇÃO E ANÁLISE DOS DOCUMENTOS

Este capítulo apresenta os principais classificadores utilizados nessa monografia, o Naïve Bayes e os *Support vector machines* (SVMs), respectivamente nas seções 2.1 e 2.2. 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 *k-fold*[1] é discutida. Ela evidencia como um método de classificação generaliza para diferentes conjuntos de documentos. 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.

Caso o leitor não entenda perfeitamente o funcionamento interno dos classificadores ou do L-LDA, é importante ter em mente que isso interfere muito pouco na compreensão dos outros capítulos. Isso se deve ao fato de que tanto os classificadores quanto o L-LDA são utilizados como caixas pretas em todos os casos.

2.1 NAÏVE BAYES

O Naïve Bayes é um classificador que se baseia no Teorema de Bayes, apresentado de maneira ilustrativa no Anexo A. Ele assume a independência condicional entre as palavras (LEWIS, 1998), conceito de Estatística que, no contexto do Naïve Bayes, significa que as palavras em qualquer documento ocorrem independentemente umas das outras. Além disso, o

[1] Devido a divergências na tradução do termo *k-fold*, decidiu-se utilizá-lo, ao longo dessa monografia, no original.

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 é capaz de apresentar um bom desempenho na classificação por ponto de vista, como pode ser conferido em alguns trabalhos do Capítulo 3 e no Capítulo 4.

O Naïve Bayes é um classificador probabilístico, cuja finalidade é encontrar a classe mais provável para um documento qualquer. Essa classe, do ponto de vista do classificador, é um número natural que corresponde a uma classe conceitual qualquer, como *contra o aborto*, *preocupado com a Economia* ou *positivo*. Dado um documento d_i pertencente a um corpus D com um vocabulário V[2], e um conjunto de classes C, a probabilidade da classe de d_i ser c_j, $c_j \in C$, dado d_i é representada pela notação $P(c_j \mid d_i)$ e é dada pela seguinte aplicação do Teorema de Bayes (LEWIS, 1998)

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

onde $P(c_j)$ é a probabilidade de se obter a classe c_j, independentemente de qualquer documento d_i; $P(d_i \mid c_j)$ é a probabilidade de se obter o documento d_i fixando-se a classe c_j - ou, em outras palavras, a probabilidade de d_i pertencer à classe c_j -; e $P(d_i)$ é a probabilidade de se obter o documento d_i, independentemente de qualquer classe. Detalhes de notação estão descritos no Anexo A. Na prática, o classificador deve buscar a classe de C que maximize a Equação 2.1. Como a probabilidade $P(d_i)$ independe de qualquer classe, ela pode ser abstraída.

Para efeito de ilustração, pode-se imaginar o seguinte cenário de aplicação do Naïve Bayes: um conjunto grande de artigos sobre aborto, divididos entre as classes pró-vida e pró-escolha. Assume-se que o Naïve Bayes, nesta aplicação, deve classificar todos eles - ou seja, baseando-se em seus conteúdos, ele deve indicar, para cada artigo, se ele é pró-vida ou pró-escolha. Após as estimativas do Naïve Bayes, é comum verificar para quantos documentos ele acertou a classe e para quantos ele errou. Desta forma, mede-se o desempenho do classificador em uma aplicação específica, assunto que será melhor discutido na seção 2.3. O bom desempenho do classificador sugere que ele está generalizando bem os pontos de vista dos artigos. Diversos outros exemplos de aplicação do Naïve Bayes podem ser conferidos no Capítulo 3.

Nesta seção, a modelagem do Naïve Bayes segue os trabalhos de McCallum e Nigam (MCCALLUM; NIGAM, 1998) e Resnik e Hardisty (RESNIK; HARDISTY, 2009). Basicamente, assume-se que há $|C|$ valores possíveis para a classe de d_i, e o Naïve Bayes assume que eles são

[2] Assume-se que esse vocabulário é composto das palavras dos documentos, mas nada impede que ele contenha outros elementos dos textos, como um símbolo correspondendo à cada classe sintática (sujeito, objeto direto etc.).

distribuídos de acordo com uma distribuição $Binomial(|C|, \pi)$. O parâmetro π, por sua vez, é um valor real no intervalo $(0,1)$. A sua escolha está associada à distribuição $Beta(\alpha, \beta)$. Os parâmetros α e β, por sua vez, são fixados antes de se iniciar o processo de classificação. Diante disso, a probabilidade de se obter exatamente o valor π é dada por

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

A função B é aplicada aos valores α e β para garantir que a distribuição de probabilidade *Beta*, quando integrada, some um (EVANS; HASTINGS; PEACOCK, 2000)[3]. Considerando-se o lado direito da Equação 2.1, e o fato de que os valores para a classe de d_i são distribuídos de acordo com $Binomial(|C|, \pi)$, tem-se que a probabilidade de se obter c_j, na prática, depende de $|C|$ e π. Por esse motivo, em vez de $P(c_j)$, o que se busca, na prática, é $P(c_j \mid \pi,\ |C|)$. Esse valor é dado por

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

Ainda considerando o lado direito da Equação 2.1, assume-se sem perda de generalidade que a classe c_j foi amostrada estatisticamente, de acordo com a distribuição de probabilidade $Binomial(|C|, \pi)$[4]. Sendo assim, deve-se estimar $P(d_i \mid c_j)$. 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. Isso é consequência de que a distribuição de acordo com a qual d_i é amostrado depende de um vetor, em vez de um escalar. Como c_j é um escalar entre 0 e $|C|-1$, associa-se um vetor de |V| entradas a ele: θ_j.

O Naïve Bayes assume que cada entrada de θ_j é um valor real no intervalo $(0,1)$, e o vetor é amostrado 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

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

A função B, 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-

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

[4] $\binom{|C|}{c_j}$ é a combinação das $|C|$ classes tomadas de c_j em c_j.

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 $P(d_i \mid c_j)$, pode ser reescrita portanto como

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

$N(w_k, d_i)$, por sua vez, é o número de vezes que a k-ésima palavra do vocabulário V, w_k, ocorre no documento d_i. A esse número, dá-se o nome de **contagem** de w_k em d_i[5]. 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. A essa representação, dá-se o nome de **ausência/presença** de w_k em d_i[6]. As duas representações foram as mais exploradas nos trabalhos revisados no Capítulo 3, especialmente a primeira. Dado um documento que contenha apenas a sentença *Há uma cadeira e uma mesa*, por exemplo, a contagem da palavra *uma* é dois; a ausência/presença, um. Tanto a contagem quanto a ausência/presença de *elefante* são zero.

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. Além disso, eles costumam ser escolhidos arbitrariamente antes da execução do modelo, em vez de serem amostrados de acordo com alguma distribuição de probabilidade.

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 algum algoritmo de inferência, como **Gibbs Sampling** ou **Expectation-Maximization**. Por questões de escopo, eles não serão apresentados nessa monografia, podendo ser consultados no livro de Aprendizado de Máquina de Christopher Bishop (BISHOP, 2006). Na prática, no decorrer das iterações, tem-se que o θ que maximiza a Equação 2.5 está associado à classe que mais se assemelha, *proporcionalmente*, a d_i (RESNIK; HARDISTY, 2009). Essa semelhança é quantificada através de valores associados aos documentos, como suas contagens de palavras ou ausência/presença de palavras. Se o Naïve Bayes utiliza as contagens de palavras, portanto, tem-se que, intuitivamente, quão mais diferentes forem essas contagens considerando-se documentos de classes distintas, 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 Es-

[5] Do inglês *word count*.

[6] Do inglês *absence or presence of a word*.

tados Unidos (MULLEN; MALOUF, 2006), inclusive, associa o mau desempenho obtido na classificação de documentos a proporções muito parecidas de contagens por classe.

Na prática, também para melhorar o desempenho da classificação, pode-se criar 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). A adoção dessa estratégia faz com que o classificador aprenda melhor a caracterizar cada classe, aumentando sua probabilidade de estimar corretamente a classe de um documento qualquer. O Naïve Bayes não precisa reamostrar as classes dos documentos do conjunto de treinamento, 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 nos parágrafos anteriores, 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. A Figura 2.1 indica como o Naïve Bayes, aplicado iterativamente e utilizando conjuntos de treinamento e teste, pode ser utilizado para estimar as classes de um conjunto de documentos.

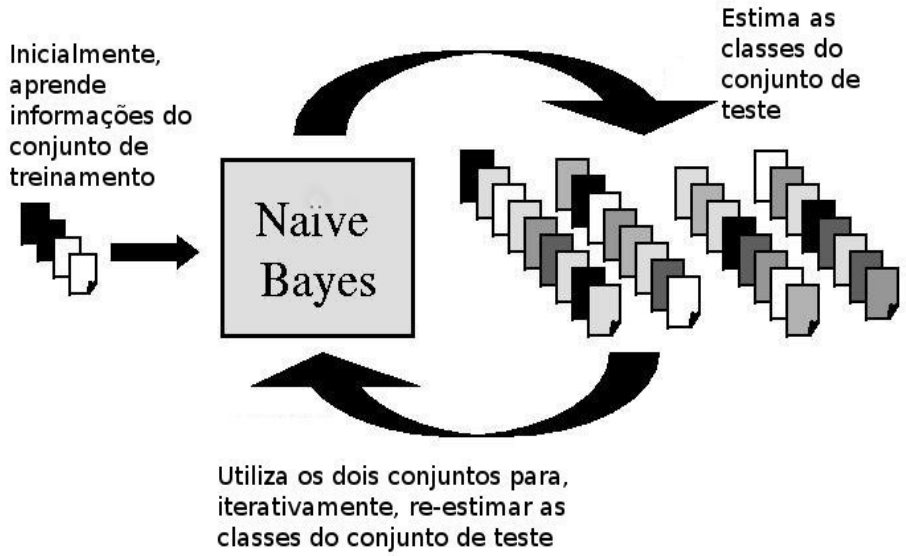

Figura 2.1: Naïve Bayes aplicado iterativamente e utilizando conjuntos de treinamento e teste (NIGAM, 2001).

Para esse projeto, a implementação de Naïve Bayes desenvolvida[7] aplica o algoritmo 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

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

para estabilizar a classificação em todos os experimentos conduzidos, apresentados no Capítulo 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, que são 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[8] 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. Greene e Resnik, em seu trabalho sobre pena de morte, não utilizam palavras para representar os documentos, mas sim *tuplas sintáticas* (GREENE; RESNIK, 2009). A geração dessas tuplas é descrita no Capítulo 3. 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 2.6 e 2.7 (OGURI, 2006)

$$\theta_1 \equiv \mathbf{x} \cdot \mathbf{w} + b = 1 \tag{2.6}$$

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

As representações de documentos que pertencem a θ_1 ou θ_{-1} recebem o nome de *vetores suporte*[9] (OGURI, 2006). O objetivo inicial de um classificador SVM é escolher os parâmetros $\mathbf{w}$, um vetor, e b, um escalar, que maximizem a distância entre esses hiperplanos. A solução para hiperplanos de duas dimensões (planos) está ilustrada na Figura 2.2. Para determinar

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

[9] Do inglês *support vectors*.

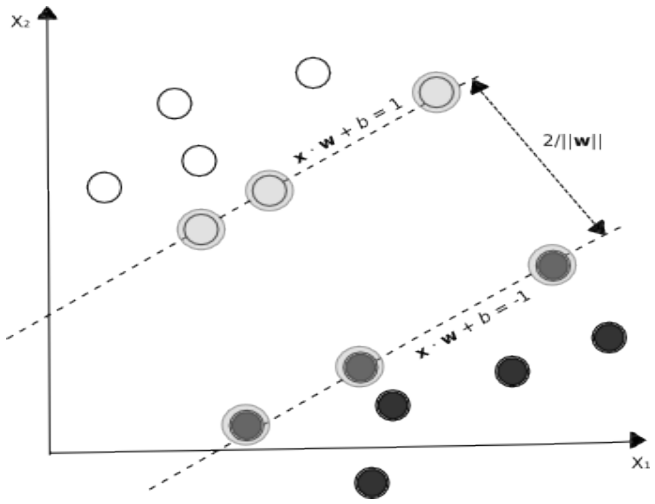

Figura 2.2: Solução para um cenário bidimensional. Os círculos escuros representam documentos da classe -1; os claros, aqueles da classe 1. As representações que incidem nos hiperplanos são os vetores de suporte. $||\mathbf{w}||$ é a norma euclidiana do vetor $\mathbf{w}$.

esses parâmetros, o problema pode ser remodelado com n multiplicadores de Lagrange $\{\alpha_i\}$, onde n é a cardinalidade do conjunto de treinamento (OGURI, 2006). Discussões sobre esses multiplicadores fogem do escopo dessa monografia, podendo ser consultadas na dissertação de Pedro Oguri (OGURI, 2006).

Em seguida, 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 de (OGURI, 2006)

$$y(x_j) = sinal\left(\sum_{i=1}^{n} \alpha_i y_i (x_i \cdot x_j) + b\right) \tag{2.8}$$

Para efeito de ilustração, pode-se imaginar o seguinte cenário de aplicação de um SVM: um conjunto de documentos sobre os preceitos cristãos, divididos entre os pontos de vista protestante (1) e católico (-1). O SVM, após a construção dos hiperplanos θ_1 e θ_{-1}, deve discriminar a classe de cada um desses documentos de acordo com a Equação 2.8. Assim como ocorre com o Naïve Bayes, é comum medir o desempenho do SVM após essa discriminação, para saber se ele generalizou bem o aprendizado dos dois pontos de vista. Diversos exemplos do uso de SVMs estão elencados no Capítulo 3.

Representações muito parecidas para documentos pertencentes a classes diferentes comprometem 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.

Embora o Naïve Bayes possa ser adaptado para explorar outros elementos dos documentos, além de palavras, os trabalhos revisados no Capítulo 3 indicam que o SVM é mais utilizado nesses cenários. Ambos os classificadores são aplicados pela maioria dos trabalhos revisados nesse capítulo e, pelo que tudo indica, apresentam desempenho semelhante. Para classificação por ponto de vista, ora um funcionou melhor, ora outro.

2.3 MÉTRICAS PARA MEDIR O DESEMPENHO DOS CLASSIFICADORES

A classificação de documentos, de acordo com seus pontos de vista, é o principal objetivo 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, recuperação ou métrica F1. A **taxa de acerto**[10] é definida por (MANNING; RAGHAVAN; SCHUTZE, 2008)

$$\frac{\#documentos\ classificados\ corretamente}{\#total\ de\ documentos} \tag{2.9}$$

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)

$$\frac{\#documentos\ classificados\ corretamente\ como\ c}{\#total\ de\ documentos\ classificados\ como\ c} \tag{2.10}$$

A **recuperação**[11], medida também para uma classe $c \in C$ qualquer, é definida como (MANNING; RAGHAVAN; SCHUTZE, 2008)

$$\frac{\#documentos\ classificados\ corretamente\ como\ c}{\#total\ de\ documentos\ pertencentes\ a\ c} \tag{2.11}$$

A recuperação também evidencia o quanto o classificador está *acertando* - mas por classe. A **métrica F1**[12], também medida para uma classe $c \in C$ qualquer, é dada por (MANNING;

[10] Do inglês *accuracy*.
[11] Do inglês *recall*.
[12] Do inglês *F1-measure*.

RAGHAVAN; SCHUTZE, 2008)

$$2 \times \frac{precisao \times recuperao}{precisao + recuperao} \tag{2.12}$$

A métrica F1 pondera os valores obtidos para a precisão e para a recuperação de uma classe c qualquer. Essas métricas revelam aspectos diferentes do desempenho de um classificador. 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 *K-FOLD*

A validação cruzada *k-fold* 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, com aproximadamente a mesma cardinalidade (KOHAVI, 1995). O conjunto de teste corresponde a um dos k subconjuntos e os $k-1$ restantes, unidos, compõem o conjunto de treinamento.

A classificação deve ser executada k vezes, com um subconjunto diferente como conjunto de teste por vez. As métricas associadas ao desempenho de cada classificação podem ser consideradas conjuntamente, através de uma média aritmética, ou em separado (REFAEILZADEH; TANG; LIU, 2009). Nesta monografia, toda vez que se afirmar que um determinado valor, para uma métrica, foi obtido via validação cruzada *k-fold*, está se fazendo alusão à média aritmética. 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 5 e 10.

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

O modelo *Labeled-latent dirichlet allocation*, ou simplesmente L-LDA, é aplicado para se entender melhor como o conteúdo de documentos se segmenta em tópicos. Esses tópicos, para efeito prático, podem ser assuntos, ideias ou pontos de vista. A finalidade desse modelo, por-

tanto, não é classificar documentos, mas sim evidenciar como suas palavras se relacionam com esses tópicos. O L-LDA fundamenta-se na ideia de que um documento pode tratar de múltiplos tópicos, refletidos nas palavras que o compõem (GRIFFITHS; STEYVERS, 2004). 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. 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 for a probabilidade de se obter uma palavra w dado um tópico t, mais forte será a relação entre eles. De forma semelhante, quão maior for a probabilidade de se amostrar um tópico t associado a d, dada uma distribuição θ_d, mais importante será o tópico no contexto desse documento (GRIFFITHS; STEYVERS, 2004).

O L-LDA foi proposto em 2009, em um artigo no qual tópicos correspondem a *tags* de *blogs*, como *books* ou *religion* (RAMAGE et al., 2009). Um conjunto de documentos marcados

com essas *tags* foi analisado, e palavras associadas a cada *tag* estão elencadas na Tabela 2.1. Elas indicam que o modelo é útil para agrupar termos semanticamente relacionados, algo explorado no Capítulo 4. Nesse capítulo, tópicos representam pontos de vista sobre um tema. O L-LDA, utilizado dessa forma, evidencia quais palavras melhor definem, semanticamente, cada ponto de vista do estudo de caso. É válido ressaltar que esse uso do L-LDA não foi encontrado em nenhum trabalho revisado para esta monografia.

Tag	**Palavras**
Books	book, image, pdf, review, library, posted, read, copyright, books, title
Science	works, water, map, human, life, work, science, time, world, years, sleep
Religion	comment, god, jesus, people, gospel, bible, reply, lord, religion, written
Computer	windows, file, version, linux, computer, free, system, software, mac

Tabela 2.1: Palavras mais frequentemente associadas a algumas *tags* de *blogs*, de acordo com um L-LDA em artigo de Ramage et al. (RAMAGE et al., 2009).

A saída da execução de um L-LDA é um conjunto de listas que indicam 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 alto, 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 no Capítulo 4, fixa-se um mesmo k para todos os tópicos: 15.

No L-LDA, assim como no Naïve Bayes, algum algoritmo de inferência deve ser utilizado para, iterativamente, aproximar as distribuições de probabilidade apresentadas nessa seção o máximo possível daquelas presentes no corpus. A implementação do L-LDA utilizada no Capítulo 4 também utiliza Gibbs Sampling e está disponível no repositório *online* de Alexandre Passos[13]. O número de iterações para o modelo foi fixado em 100, valor suficiente para estabilizar as amostragens de tópicos e palavras.

[13] http://github.com/alextp/pylda/blob/master/llda.py

3 CLASSIFICAÇÃO DE ACORDO COM PONTOS DE VISTA: REVISÃO E DISCUSSÃO

A classificação de documentos de acordo com seus pontos de vista 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 ponto de vista visa a separar os documentos de um corpus de acordo com seus diferentes posicionamentos sobre um tema, como *pró-Israel* ou *pró-Palestina*. 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).

A *survey* de Pang e Lee, uma das principais referências para este projeto, apresenta superficialmente alguns trabalhos que fazem classificação de acordo com pontos de vista, na seção em que discute o tema. Todos eles foram publicados entre os anos de 2004 e 2006. Assim como nessa *survey*, esta monografia também explora trabalhos recentes: todos eles foram publicados entre 2006 e 2010. A metodologia aplicada na busca por trabalhos a serem revisados foi a seguinte:

1. Consideração dos 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*[1] pelas palavras *perspective*, *viewpoint*, *politics*, *political*, *ideo-*

[1] A *Association of Computational Linguistics*, *ACL*, mantém o maior arquivo digital envolvendo trabalhos de

logy e *ideological*. Estas palavras foram escolhidas por (1) funcionarem como sinônimos nos artigos coletados anteriormente ou (2) pela relevância em suas temáticas;

5. Busca nos *sites* dos eventos *EMNLP*[2], *NAACL*[3], *AAAI*[4] e *CoNNL*[5], 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 de acordo com pontos de vista. Alguns, como a tese de Alice Oh, se propõem a 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 onze trabalhos a serem revisados.

Todos esses trabalhos apresentam uma metodologia em comum: eles organizam um corpus de documentos, definem em que pontos de vista ele se divide, possivelmente pré-processam os documentos e, em seguida, classificam-nos utilizando alguma técnica - na maioria dos casos, SVMs ou Naïve Bayes[6]. Independentemente da forma de classificação, tem-se que elas sempre se baseiam em características dos documentos para determinar suas classes. Quase todos os trabalhos revisados nesse capítulo exploram, como características, contagens de palavras ou alguma variação delas[7]. Boa parte deles não faz uso de nenhuma outra característica e alguns trabalhos as comparam com outras, que normalmente são seu enfoque. Essas outras características são escolhidas de forma muito particular, variando de um trabalho para outro. Em vez de apenas considerarem ocorrências de palavras em textos, esses trabalhos exploram suas relações sintáticas, suas propriedades semânticas e também valores associados às interações entre dois ou mais documentos.

Dado que essas últimas características variam bastante, e considerando também a popularidade das primeiras nos trabalhos estudados, decidiu-se dividir a revisão dos trabalhos nas seguintes seções: a seção 3.1 apresenta trabalhos que classificam documentos baseando-se exclusivamente em contagens de palavras ou variações - ou seja, eles desconsideram *quaisquer* aspectos gramaticais dos documentos. A seção 3.2 apresenta trabalhos que exploram outras características dos documentos na classificação. É importante informar que alguns trabalhos

Linguística Computacional, o que inclui Mineração de Opinião (http://aclweb.org/anthology-new/).

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

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

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

[5] *Computational Natural Language Learning Conference.*

[6] Esses classificadores são apresentados no Capítulo 2.

[7] O conceito de contagens de palavras é apresentado no Capítulo 2.

da segunda seção, por questões comparativas, também classificam documentos exclusivamente de acordo com contagens de palavras ou variações - mas este não é o foco deles. Por fim, na seção 3.3, é apresentada uma pequena análise comparativa envolvendo os trabalhos revisados nas seções anteriores.

3.1 TRABALHOS QUE EXPLORAM CONTAGENS DE PALAVRAS OU VARIAÇÕES

Esta seção apresenta seis trabalhos que classificam documentos baseando-se apenas na contagem de suas palavras ou em pequenas variações, como ausência/presença de palavras. Eles assumem, ainda que implicitamente, que os pontos de vista dos documentos já são suficientemente discrimináveis no nível das escolhas de palavras. A ênfase em palavras diferentes, ou até mesmo o uso de palavras específicas, é um elemento chave para a transmissão de posicionamentos distintos sobre um determinado assunto. Essa ideia, é válido ressaltar, encontra respaldo na área de Linguística: indivíduos com posicionamentos diferentes utilizam palavras distintas, ou as enfatizam de modos diversos, para identificar mais facilmente quem pensa da mesma forma e quem se opõe ideologicamente (TEUBERT, 2001).

O trabalho de **Lin et al.** classifica artigos do *site* Bitterlemons[8] 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 vêm de diferentes divisões entre os conjuntos de treinamento e teste. Em alguns casos, foi utilizada a validação cruzada *10-fold*[9]; em outros, os conjuntos foram divididos de acordo com os autores dos artigos.

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 sentenças. 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.

[8] http://bitterlemons.org/

[9] Essa técnica de validação é descrita no Capítulo 2.

Para essas divisões, as taxas de acerto obtidas com o Naïve Bayes foram de 84.85% e 93.46% respectivamente. O SVM não foi comparado diretamente com o LSPM. Apesar do desempenho superior ao do Naïve Bayes, nenhum outro trabalho revisado faz uso desse classificador. O tutorial de Resnik e Hardisty sobre Naïve Bayes e LSPM sugere algumas equações para a implementação deste último (RESNIK; HARDISTY, 2009), mas a autora dessa monografia não foi capaz de reproduzir esses resultados. Isso foi corroborado por (HARDISTY, 2010). O LSPM, que utiliza hiperparâmetros[10] assim como o Naïve Bayes, aparenta ser muito sensível aos valores fixados.

Mullen e Malouf propõem dois trabalhos que analisam um conjunto de *posts* do fórum de discussão política Politics[11]. 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 foi de 60.37%, via validação cruzada *10-fold*. 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 de 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

[10] Ver definição no Capítulo 2, seção 2.1.
[11] http://politics.com

mais de 500 palavras de fala no fórum, a taxa de acerto relatada é de 73.00%, via validação cruzada *5-fold*. 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 ausência/presença de palavras nos documentos (uma variação das contagens de palavras). 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, 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[12], 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. Todos os resultados apresentados foram obtidos via validação cruzada *10-fold*. 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

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

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 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, por exemplo, ou situação e oposição.

O trabalho de **Klebanov, Beigman e Diermeier** avalia o desempenho dos classificadores Naïve Bayes e SVM, utilizando para isso algumas variações de contagens de palavras (KLEBANOV; BEIGMAN; DIERMEIER, 2010). Para o Naïve Bayes, os autores comparam a escolha entre (1) ausência/presença 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. 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. (LIN et al., 2006); 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%. 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 (DURANT; SMITH, 2006), 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. Todos os resultados apresentados foram obtidos via validação cruzada *10-fold*.

3.2 TRABALHOS QUE EXPLORAM OUTRAS CARACTERÍSTICAS DOS DOCUMENTOS

Esta seção apresenta cinco trabalhos que quantificam outras características dos documentos, a fim de identificar seus pontos de vista. Elas abrangem escolhas sintáticas ou semânticas dos autores dos documentos, como nos trabalhos de Greene e Resnik (RESNIK; HARDISTY, 2009) ou Jiang e Argamon (JIANG; ARGAMON, 2008), ou a forma como dois ou mais documentos interagem, como no trabalho de Efron (EFRON, 2004). Os trabalhos dessa seção assumem que as escolhas gramaticais nos documentos, bem como a forma com que eles interagem, contribuem de forma determinante na identificação de seus diferentes pontos de vista. Greene e Resnik inclusive citam as sentenças *Man suffocates 24-year old woman* e *Suffocation kills 24-year-old woman*, que não se contradizem, para evidenciar como escolhas sintáticas mudam suas conotações (GREENE; RESNIK, 2009) - algo intimamente relacionado com expressão de pontos de vista.

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 (KLEBANOV; BEIGMAN; DIERMEIER, 2010), apresentado na seção anterior. O objetivo do trabalho é classificá-lo de acordo com os pontos de vista 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 *sintática* 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 os termos a papéis sintáticos, como *sujeito*, *objeto direto*, *verbo transitivo*, dentre outros. Na tupla, ou o termo corresponde sintaticamente ao papel associado ou se subordina a algum termo que possui esse papel. Dada a frase *Aline matou uma mosca*, por exemplo, o método primeiro gera as relações de dependência (*Aline, matou*) e (*matou, uma mosca*) para, em seguida, destrinchá-las nas tuplas (*Aline, sujeito*), (*sujeito, matou*), (*matou, verbo transitivo*), (*verbo transitivo, uma mosca*) e (*uma mosca, objeto direto*). A determinação dessas tuplas foi feita com o apoio do programa *Stanford Parser*[13], adaptado à língua inglesa.

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

Greene e Resnik então aplicam um SVM nos documentos reduzidos a essas representações. Conforme apresentado no Capítulo 2, o SVM representa o documento numericamente, como um vetor em um espaço multi-dimensional. Embora os autores não indiquem, os valores numéricos provavelmente correspondem às contagens dessas tuplas nos documentos, ou a algo similar. As taxas de acerto obtidas por Greene e Resnik foram de 82.09% e 88.10%, a depender da escolha de verbos. 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, o SVM provavelmente explora as contagens desses bigramas nos documentos, ou algo similar, e obtém taxas de acerto de 68.37% e 71.96%, também a depender da escolha dos verbos. Todos os resultados apresentados pelos autores foram obtidos via validação cruzada *4-fold*. O uso das tuplas, portanto, trouxe uma melhoria muito significativa para a classificação desse corpus, ainda que esteja intimamente relacionada a uma língua específica.

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 ao ponto de vista liberal ou conservador, 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 ausência/presença 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 recuperação 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*[14]. 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, recuperação 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 ausência/presença de palavras e, adicionalmente, a ausência/presença das expressões opinativas separadas para cada lado. A taxa de acerto neste cenário foi de 84.96%. As precisão, recuperação e métrica F1 médias foram de, respectivamente, 83.24%, 83.48% e 83.27%. Todos os resultados apresentados por Jiang e Argamon foram obtidos via validação cruzada *10-fold*. 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

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

General Inquirer online.

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 os pontos de vista direita e esquerda e o segundo, de acordo com as orientações pró-alternativa ou pró-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 em palavras*. 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 o posicionamento de cada documento de acordo com a probabilidade deles serem co-citados com documentos fixados como referência para cada ponto de vista. Essa probabilidade é estimada a partir dos números de documentos retornados pelo buscador AltaVista[15] 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 os posicionamentos 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 ausência/presença 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,

[15] http://altavista.com/

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, 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, no novo esquema, 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 (THOMAS; PANG; LEE, 2006), 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, Cardie 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.

3.3 ANÁLISE COMPARATIVA

Os trabalhos revisados neste capítulo propõem, em quase todos os casos, *datasets* relacionados com questões políticas. Neste sentido, o trabalho de **Efron**, revisado na seção 3.2, destaca-se por também abordar pontos de vista sobre música (pró-alternativa e pró-popular) (EFRON, 2004). De fato, não é simples definir o que é um ponto de vista ou perspectiva, mas, considerando-se a discussão proposta na *survey* de Pang e Lee (PANG; LEE, 2008), essas terminologias não abrangem apenas temáticas políticas - mesmo que seja mais fácil visualizá-las nesses contextos. Adicionalmente, Alice Oh, em seu trabalho sobre modelagem de perspectiva, aborda uma temática esportiva (jogos de *baseball*) (OH, 2008). Embora ela não lide com classificação, seu estudo envolve as perspectivas sobre os jogos, evidenciando que é possível explorar temáticas não-políticas. Seria interessante, portanto, ampliar os estudos de classificação por ponto de vista para outros domínios sócio-culturais. Por fim, é importante escolher esses pontos de vista com certo cuidado, de modo que eles correspondam efetivamente ao conteúdo do corpus. Neste sentido, o trabalho de **Hirst, Riabinin e Graham**, apresentado na seção 3.1, apresenta uma boa discussão (HIRST; RIABININ; GRAHAM, 2010).

Quanto à definição dos pontos de vista, todos os trabalhos revisados dividiram seus corpora em *dois* deles, associando cada um a uma classe. Talvez a exploração de temáticas políticas tenha favorecido esse recorte, mas é importante ressaltar que, em tese, nada impede estudos em *datasets* com mais pontos de vista. No que diz respeito ao pré-processamento dos documentos, destaca-se o segundo trabalho de **Mullen e Malouf**, apresentado na seção 3.1 (MULLEN; MALOUF, 2008). Apesar da classificação em si ser simples, baseando-se apenas em contagens de palavras, a geração dos documentos envolve a formulação de um grafo e a análise de padrões de citação semelhantes. É válido reforçar que essa estratégia só é válida em cenários onde há muitas citações e padrões razoavelmente diferentes. Se todos os documentos se citam de forma semelhante, a geração de agrupamentos significativos pode ser prejudicada.

No que diz respeito aos classificadores, foi observado que não há consenso quanto à escolha pelo Naïve Bayes ou por um SVM, principais técnicas abordadas. Aparentemente, o desempenho de ambos não difere muito; o que realmente apresenta um impacto na classificação é a escolha das características dos documentos, exploradas por algum dos classificadores. De todo modo, **Lin et al.** destaca, em seu trabalho revisado na seção 3.1, o desenvolvimento de um novo classificador, o LSPM (LIN et al., 2006). Esse classificador apresenta um desempenho apenas um pouco superior ao do Naïve Bayes, além de não ter sido explorado por nenhum outro trabalho pesquisado. Por outro lado, os trabalhos de Efron, **Thomas, Pang e Lee** e **Bansal, Cardie e Lee** apresentam metodologias de classificação completamente diferentes do Naïve

Bayes e dos SVMs, e obtêm melhorias significativas em seus resultados (EFRON, 2004; THOMAS; PANG; LEE, 2006; BANSAL; CARDIE; LEE, 2008). Essas metodologias, entretanto, exploram valores associados às interações entre documentos: quando elas não são significativas, essas metodologias podem não ser úteis.

Os trabalhos revisados exploram características diferentes dos documentos e, até quando exploram as mesmas, podem escolher apenas um subconjunto delas ou investir no pré-processamento dos documentos. Tudo isso contribui para a riqueza metodológica desses trabalhos, de modo que ainda não há consenso sobre qual é a melhor forma de classificar documentos de acordo com seus pontos de vista. Aparentemente, a resposta para essa questão depende do cenário. Em alguns casos, a simples aplicação de um Naïve Bayes ou de um SVM, explorando-se apenas contagens de palavras ou alguma variação, já conduz a bons resultados. É o caso dos trabalhos de Lin et al., Hirst, Riabinin e Graham e **Klebanov, Beigman e Diermeier**, revisados na seção 3.1, e de **Jiang e Argamon**, revisado na seção 3.2 (LIN et al., 2006; HIRST; RIABININ; GRAHAM, 2010; KLEBANOV; BEIGMAN; DIERMEIER, 2010; JIANG; ARGAMON, 2008). Em outros casos, é possível melhorar o desempenho utilizando-se *ainda* apenas as contagens de palavras ou alguma variação. Nestes casos, a estratégia pode envolver mudar o pré-processamento dos documentos, como fazem Mullen e Malouf em seu segundo trabalho, ou escolher um subconjunto ótimo de palavras, como no trabalho de **Durant e Smith** (MULLEN; MALOUF, 2008; DURANT; SMITH, 2006).

Quando essas mudanças não são suficientes, os autores dos documentos analisados não estão consolidando seus pontos de vista, de forma suficiente, no nível das palavras. Ou seja, em um nível no qual se despreza qualquer aspecto semântico ou sintático dos documentos. É provável que, nestes cenários, as palavras estejam ocorrendo nos textos de forma muito semelhante, de modo que suas contagens (ou alguma variação das contagens), no contexto dos classificadores, não são suficientes para uma discriminação adequada dos pontos de vista. Nestes casos, sugere-se a investigação de outras características dos documentos, como indicam os trabalhos de **Greene e Resnik**, Efron e Thomas, Pang e Lee, revisados na seção 3.2 (GREENE; RESNIK, 2009; EFRON, 2004; THOMAS; PANG; LEE, 2006).

O uso de outras características pode envolver a aplicação de *softwares* de apoio ou consultas a dicionários semânticos adaptados a *línguas específicas*, como evidenciam os trabalhos de Greene e Resnik (GREENE; RESNIK, 2009) e Jiang e Argamon (JIANG; ARGAMON, 2008), e nem sempre apresentam uma melhoria significativa em relação a metodologias mais simples, como nesse último caso (aproximadamente 3% de aumento em relaçao a um SVM baseado em ausência/presença de palavras). A exploração da forma como documentos interagem ou são

co-citados também pode ser desaconselhável em cenários onde as co-citações, ou interações, não são expressivas. Essa exploração pode criar ruídos na classificação, em vez de efetivamente indicar semelhanças úteis para a identificação correta das classes dos documentos.

4 ESTUDO DE CASO: PONTOS DE VISTA 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 (HIRST; RIABININ; GRAHAM, 2010), quer sejam artigos escritos por cidadãos ou especialistas[1] (MULLEN; MALOUF, 2006, 2008). Considerando essa tendência, e o fato de que 2010 foi ano de eleições para presidente no Brasil, decidiu-se realizar um estudo de caso que aproveita a abundância de artigos, disponíveis na *Web*, sobre o governo Lula e a sucessão presidencial. A ideia é construir um corpus com alguns desses documentos e investigar seus pontos de vista automaticamente, classificando-os de acordo com eles e analisando, de forma subjetiva, as palavras por eles enfocadas. É válido ressaltar que não foi encontrado nenhum outro trabalho, em português, que classifique documentos de acordo com seus pontos de vista.

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 a definição dos pontos de vista explorados; na seção 4.2, experimentos com um classificador Naïve Bayes[2] são conduzidos para, assim como em outros trabalhos revisados neste projeto, se classificar artigos de acordo com seus pontos de vista; na seção 4.3, o modelo de tópicos L-LDA[3] é aplicado ao corpus, evidenciando o uso de palavras por artigos com posicionamentos diferentes.

[1] Todos os trabalhos mencionados nesta sentença foram revisados no Capítulo 3.
[2] Esse classificador é apresentado no Capítulo 2.
[3] Esse modelo é apresentado no Capítulo 2.

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 os jornalistas Lucas Cunha Almeida e Andrea Duarte Bessa[4]. 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 *dataset* desse estudo de caso dividem-se entre pró-situação e pró-oposição. Neste sentido, há autores que apóiam candidatos de ambos os lados, muitas vezes criticando o governo Lula e suas personalidades. Os trechos de *posts* apresentados nessa seção reforçam a ideia de que essa divisão é adequada. O lado pró-situação é composto de artigos veiculados em:

1. **Luis Nassif Online**[5] Este é o *blog* do jornalista Luis Nassif, premiado como Melhor Blog de Política pelo iBest 2008[6]. 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-situação, como os *sites* Blog da Dilma[7] e Os Amigos do Presidente Lula[8]. De fato, o Luis Nassif Online adota um posicionamento pró-situação, 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

[4] Lucas Cunha, atualmente, é repórter do jornal baiano A Tarde; Andrea Bessa é gerente de comunicações da empresa Cetrel.

[5] http://www.advivo.com.br/luisnassif/

[6] http://idgnow.uol.com.br/internet/2008/05/21/ibest-2008-anuncia-vencedores/

[7] http://dilma13.blogspot.com/

[8] http://osamigosdopresidentelula.blogspot.com/

> *"Na entrevista, Bonner se limitou a perguntar da dependência de Dilma em relação à [sic] 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, 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**[9] 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[10]. Os trechos abaixo justificam a escolha do *site* como representante da mídia *online* pró-situação:

   > *"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.

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

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

3. **Escrevinhador**[11] 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 ao Brasil, o conteúdo do *blog* assume uma perspectiva pró-situação, 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**[12] 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-situação, como o Blog da Dilma[13]. 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

[11] http://www.rodrigovianna.com.br/
[12] http://brasiliaeuvi.wordpress.com/
[13] http://dilma13.blogspot.com/2010/08/caso-lunus-verdade-dos-fatos.html

> *"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 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 pró-oposição, por sua vez, é composto de artigos veiculados em:

1. **Reinaldo Azevedo**[14] 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**[15] 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[16] e o Blog da Dilma[17], justamente por assumirem uma posição pró-oposição. Os trechos abaixo justificam esta perspectiva:

[14] http://veja.abril.com.br/blog/reinaldo/
[15] http://veja.abril.com.br/blog/augusto-nunes/
[16] http://www.advivo.com.br/blog/luisnassif/serra-e-fhc-uma-relacao-delicada
[17] 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**[18] 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**[19] 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

[18] http://veja.abril.com.br/blog/mainardi/
[19] 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 pró-oposição:

> *"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 críticos de outros autores, em seus escritos, colaborando para a riqueza da linguagem no corpus.

Outros veículos foram cogitados, como o Blog do Noblat[20], o *blog* de Miriam Leitão para o jornal O Globo[21], a coluna de Cristiana Lôbo para o portal G1[22] e o *blog* de Celso Ming para o jornal O Estado de S. Paulo[23]. 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-situação e pró-oposição. 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* desenvolvidos nas linguagens de programação Python[24] e UNIX ShellScript[25]. Todos eles estão disponíveis no repositório *online* de Aline Bessa[26]. 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.

[20] http://oglobo.globo.com/pais/noblat/
[21] oglobo.globo.com/economia/miriam/
[22] http://g1.globo.com/platb/cristianalobo/
[23] blogs.estadao.com.br/celso-ming/
[24] http://python.org/
[25] http://www.gnu.org/software/bash/manual/bashref.html
[26] 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-situação e 2747 pró-oposição. 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 pró-oposição para 1065, utilizando-se apenas 550 dos 2377 artigos extraídos do *blog* de Reinaldo Azevedo, amostrados aleatoriamente. Essa estratégia foi adotada porque o classificador se mostrou sensível a quantidades muito discrepantes de palavras por ponto de vista. 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 4.1. No corpus sobre o conflito Israel-Palestina, estudado por Lin et al. em trabalho revisado no Capítulo 3, este comportamento também foi observado. Os documentos desse *dataset* foram escritos por diversos autores, e alguns deles contribuíram com muito mais textos do que outros. Ainda assim, como pode ser verificado na revisão apresentada no Capítulo 3, 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 CLASSIFICANDO DOCUMENTOS COM UM NAÏVE BAYES

O primeiro estudo conduzido com esse corpus consiste na classificação dos artigos de acordo com os pontos de vista pró-oposição e pró-situação. Esse capítulo não se propõe a comparar classificadores diferentes, de modo que o Naïve Bayes será a única técnica explorada em seu escopo. O classificador Naïve Bayes foi escolhido porque, de acordo com a revisão do Capítulo 3, não há um consenso sobre qual dos dois - ele ou um SVM - é mais indicado para classificação de acordo com pontos de vista. Em alguns trabalhos o Naïve Bayes conduz a

melhores resultados; em outros, o SVM. Essas questões, e a consideração de que o Naïve Bayes é mais simples de implementar, justificam sua escolha.

O classificador explora apenas as contagens de palavras dos documentos para discriminar suas classes. Os motivos para essa escolha foram: popularidade do uso nos trabalhos revisados no Capítulo 3, o que indica essa escolha como a mais natural; simplicidade de computação, dado que essa escolha envolve apenas a soma das ocorrências de palavras; fácil adaptação a qualquer língua, em particular a portuguesa.

Essas escolhas se mostraram adequadas para o problema: a taxa de acerto obtida na classificação foi de 89.43%. Assim como em outros artigos revisados no Capítulo 3 (LIN et al., 2006; MULLEN; MALOUF, 2006; KLEBANOV; BEIGMAN; DIERMEIER, 2010), esse valor foi obtido via validação cruzada *10-fold*[27]. Ou seja, ele corresponde à média aritmética das taxas de acerto obtidas via essa forma de validação. De modo semelhante, a precisão, recuperação e métrica F1[28] associadas a cada classe (ponto de vista) foram obtidas, e constam na Tabela 4.2.

Classe	**Precisão**	**Recuperação**	**Métrica F1**
Pró-situação	91.60%	87.08%	89.18%
Pró-oposição	87.78%	91.79%	89.65%

Tabela 4.2: Métricas de desempenho associadas a cada classe.

A primeira coluna da Tabela 4.2 indica que o Naïve Bayes classificou menos documentos erroneamente como pró-situação do que como pró-oposição. Em outras palavras, a classificação como pró-situação foi mais precisa. A segunda coluna indica que, por outro lado, o Naïve Bayes classificou mais documentos pró-oposição corretamente. Neste sentido, a recuperação funciona como a taxa de acerto. Por fim, a terceira coluna indica que, ponderando as colunas anteriores via métrica F1, os resultados obtidos são praticamente idênticos. Isso significa, na prática, que o Naïve Bayes apresenta um desempenho bastante equilibrado considerando-se ambas as classes.

O bom desempenho obtido com o Naïve Bayes indica que o uso de metodologias mais complexas, envolvendo aspectos sintáticos dos documentos, por exemplo, não é necessário. Além disso, é válido ressaltar que a adoção de características sintáticas/semânticas, para a classificação de um corpus em português, não é tão simples como para um corpus em inglês. Isso advém da carência de *softwares* e dicionários semânticos para a língua portuguesa - em particular, a variação utilizada no Brasil. Esses *softwares* e dicionários colaboram com a determinação de diversos aspectos gramaticais/linguísticos dos textos, minimizando o esforço manual requerido quando eles não podem ser utilizados.

[27] Essa técnica de validação é apresentada no Capítulo 2.
[28] Essas métricas são descritas no Capítulo 2.

Os valores obtidos com o classificador Naïve Bayes são comparáveis àqueles obtidos por **Durant e Smith** em seu trabalho sobre George W. Bush e a Guerra do Iraque[29] (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.

Os artigos escolhidos para este corpus são compostos, muitas vezes, de textos de outros autores. Isto reforça a ideia de que o classificador Naïve Bayes está efetivamente aprendendo os pontos de vista dos documentos, em vez de estilos de escrita. De todo modo, assim como no estudo de **Lin et al.** sobre o conflito Israel-Palestina[30] (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 os pontos de vista dos documentos, o desempenho do classificador não deve ser muito diferente do obtido na validação cruzada *10-fold*. 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 média de 93.32% e métrica F1 média de 91.86%. Este experimento, portanto, ratifica os outros resultados, evidenciando que o classificador Naïve Bayes cumpre bem a tarefa de identificar os pontos de vista pró-oposição e pró-situação.

Esses experimentos indicam que o Naïve Bayes aprendeu a generalizar bem os pontos de vista contidos nos documentos, o que sugere que seu uso, estendido à classificação de outros documentos dentro da mesma temática, pode ser adequado. Apenas a caráter de ilustração, segue um trecho de um dos documentos que o Naïve Bayes classificou corretamente:

> *"A frase do Guimarães Rosa, ali no alto do blog, lembra: 'Toda saudade é uma espécie de velhice'. Hoje me senti mais velho. Bateu uma saudade danada de outros tempos, ao assistir esse debate da 'Band' entre os presidenciáveis. Com um olho na careca do Boechat e outro na do Rogério Ceni (fiquei secando o São Paulo no outro canal), acabei perdendo alguma coisa. Mas o que vi foi triste, desmaiado, sem força nem brilho."*
>
> Retirado de "Um debate pálido na "Band": saudade de 89!", do *blog* de Rodrigo Vianna -06/08/2010

Também a caráter de ilustração, segue um trecho de um dos documentos que o Naïve Bayes classificou incorretamente:

[29] Esse trabalho foi revisado no Capítulo 3.
[30] Esse trabalho foi revisado no Capítulo 3.

"Na última terça-feira, José Dirceu, defendendo-se da denúncia de estar fazendo loby [sic] para um empresário, no caso Telebrás, disse que tudo se explicava pela 'oposição política e ideológica' ao plano do governo Lula de recriar uma grande tele estatal. Solicitado a especificar essa oposição, disse: 'Evidentemente existe interesse das telefônicas, das TVs abertas, porque do que estamos falando? De um mercado de bilhões e bilhões de reais. Vamos supor que se crie a Telebrás. Se as empresas do governo passam a trabalhar com a Telebrás, [isso] sai das empresas de telefonia. É disso que se trata a discussão'."

Retirado de "Novos cotistas", do portal de Carlos Sardenberg -25/02/2010

4.3 ILUSTRANDO O USO DE PALAVRAS POR PONTO DE VISTA

A análise das principais palavras enfocadas por cada ponto de vista amplia a compreensão dos resultados obtidos na seção anterior. A alta taxa de acerto obtida sugere que os jornalistas se expressam de forma bastante diferente, de modo que o Naïve Bayes não *se engana* com muita frequência. Em outras palavras, como essas expressões foram quantificadas via contagens de palavras, a alta taxa de acerto indica que essas características foram suficientes para discrimar documentos de classes diferentes. Essa seção se propõe a ilustrar quais palavras foram mais destacadas por eles, colaborando para a transmissão de pontos de vista diferentes.

Para a investigação sobre essas palavras, foi feita uma aplicação do modelo L-LDA. Cada documento foi associado a dois tópicos: um neutro, igual para todos eles, e um referente a seu ponto de vista (pró-oposição ou pró-situação). Há, portanto, três tópicos diferentes nesta aplicação. O uso de um tópico neutro, associado a todos os documentos, ajuda a filtrar palavras muito comuns nos *datasets*, independentemente de ponto de vista. Dois exemplos são as palavras *vez* e *campanha*. Essa é a diferença fundamental entre essa aplicação do L-LDA e a simples contagem de palavras em documentos, dividida entre os dois pontos de vista. Esse tipo de contagem não diferencia quais palavras são *mais destacadas* em documentos escritos sob uma certa perspectiva e quais são *muito utilizadas* por todos eles. Essas últimas possivelmente não os discriminam tanto quanto as primeiras.

A Tabela 4.3 elenca as quinze palavras mais frequentemente associadas a cada ponto de vista. O tópico neutro não será analisado, haja vista que foi utilizado apenas para auxiliar na filtragem das palavras mais comuns do corpus. Em alguns casos, essas palavras coincidem nos dois pontos de vista. Isso significa que ambos deram muito destaque a elas.

Tópico	Palavras
Pró-situação	serra, dilma, lula, psdb, presidente, pt, candidato, folha, tucano, rio, eleições, partido, jornal, campanha, pesquisa
Pró-oposição	dilma, lula, presidente, brasil, rousseff, rio, pt, gente, candidata, mundo, candidato, petista, entrevista, partido, josé

Tabela 4.3: As quinze palavras mais frequentemente associadas aos tópicos pró-situação e pró-oposição, em ordem e excluindo-se artigos, preposições, conjunções, advérbios e pronomes pessoais.

A Tabela 4.3 indica que as palavras enfocadas por pontos de vista diferentes são as mesmas em alguns casos, diferindo apenas na forma como são enfatizadas. Em outras palavras, a ordem das palavras nessa tabela corresponde, diretamente, ao quanto elas se associaram a cada tópico. Os artigos pró-oposição, por exemplo, dão muito destaque às palavras *lula* e *dilma*; os pró-situação, por sua vez, também enfatizam estas palavras, mas dão um destaque maior a *serra*, candidato à presidência pelo PSDB. A associação de palavras semelhantes aos tópicos pró-oposição e pró-situação 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 pontos de vista, como pode ser visto no trabalho de Ramage et al. sobre *tags* de *blogs* (RAMAGE et al., 2009).

Para visualizar melhor a ênfase dada às palavras na Tabela 4.3, elas foram processadas pelo *software* wordle[31], resultando nas figuras 4.1 e 4.2. O tamanho das palavras nas imagens é proporcional ao número de vezes que elas se associam a cada tópico. As imagens evidenciam o destaque dado aos políticos Lula, Dilma Rousseff e José Serra nos artigos analisados.

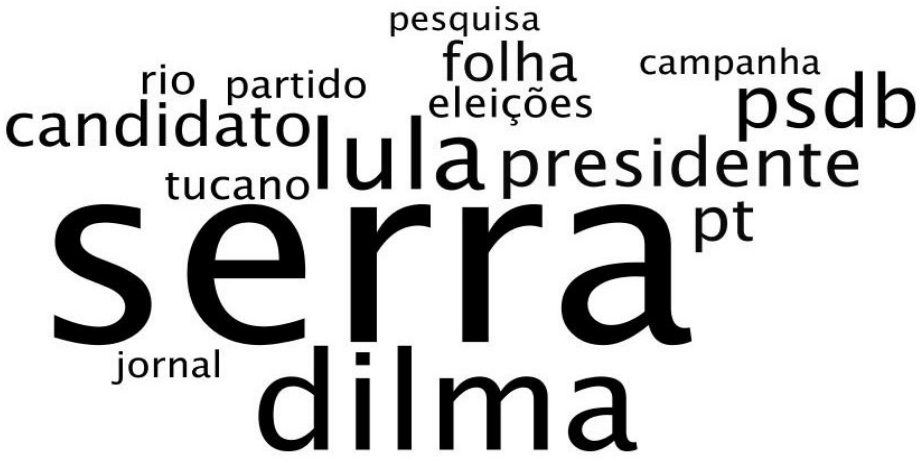

Figura 4.1: Representação gráfica para as quinze palavras associadas ao tópico pró-situação na Tabela 4.3.

[31] http://wordle.net

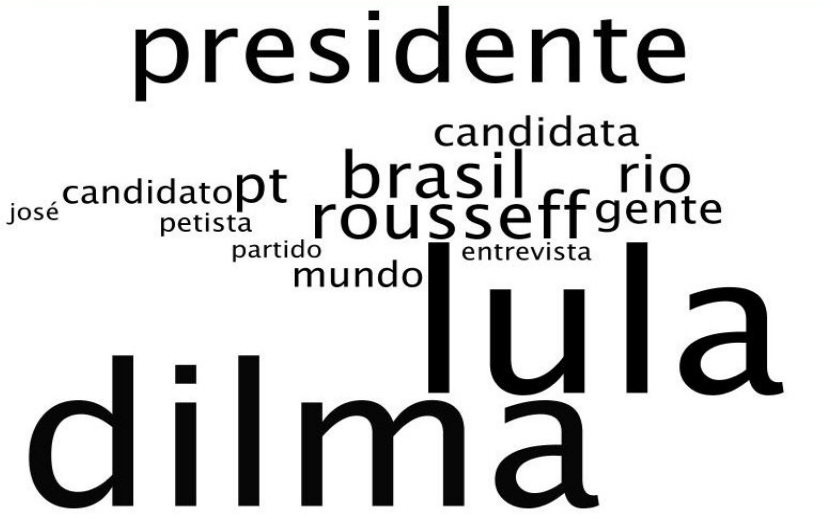

Figura 4.2: Representação gráfica para as quinze palavras associadas ao tópico pró-oposição na Tabela 4.3.

Essas figuras também sugerem que os artigos pró-situação dão mais enfoque a personalidades relacionadas à situação, como Lula e Dilma Rousseff, do que os pró-oposição a personalidades da oposição, como José Serra ou Marina Silva. Essa última, inclusive, candidata à presidência pelo PV, não foi mencionada nas palavras listadas na Tabela 4.3. Essas duas observações evidenciam 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.

Os pontos de vista desse corpus são construídos através de argumentações complexas, como indicam os trechos de artigos apresentados na seção 4.1. Para compreender melhor a relação que as palavras da Tabela 4.3 estabelecem com eles, recomenda-se, por fim, a leitura de um número razoável de passagens de texto que as contenham. Alguns trechos foram selecionados abaixo, a 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 - 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

5 CONCLUSÃO

Com a disseminação da Web, a elaboração e disseminação de textos carregados com um ponto de vista se popularizou. Não se tratam de documentos que trazem opiniões pontuais a respeito de um único objeto, como um filme ou um livro, mas sim a exposição de argumentos e ideias que, unidos, transmitem a defesa de uma posição a respeito de um certo tema. A leitura da *survey* de Pang e Lee e de alguns artigos, como o de Pang, Lee e Vaithyanathan sobre filmes no IMDb[1] ou o de Dave, Lawrence e Pennock sobre produtos vendidos na Amazon[2], sugere que a classificação de acordo com opiniões tem sido mais explorada em *datasets* que envolvem produtos ou serviços (PANG; LEE, 2008; PANG; LEE; VAITHYANATHAN, 2002; DAVE; LAWRENCE; PENNOCK, 2003). Os artigos revisados no Capítulo 3, por sua vez, indicam que a classificação de acordo com pontos de vista é mais aplicada a *datasets* que envolvem posicionamentos políticos e temáticas polêmicas, como pena de morte.

A revisão apresentada no Capítulo 3, além de indicar preferências temáticas na área, também aponta algumas predileções metodológicas. A maioria dos trabalhos classifica os documentos com um Naïve Bayes ou um SVM[3]. Em boa parte dos casos, essa classificação explora apenas as diferentes contagens de palavras nos documentos. Alguns outros artigos exploram propriedades sintáticas ou semânticas dos *datasets* analisados. Outros, por fim, consideram interações ou co-citações entre documentos na determinação de seus pontos de vista. A contagem de palavras, simples de ser computada em qualquer língua, foi a característica mais estudada pelos trabalhos revisados - ainda que não tenha recebido destaque em todos eles. Em boa parte dos casos, seu uso exclusivo já é suficiente para uma boa classificação. Em alguns outros, o uso de outras estratégias, mais complexas, foi fundamental para a melhoria na classificação. Diante disso, recomenda-se, inicialmente, para o problema de classificação de acordo com pontos de vista, o uso *exclusivo* de contagens de palavras. Se os resultados obtidos não forem satisfatórios, recomenda-se a escolha de um subconjunto ótimo de palavras, como indica o trabalho de Durant e Smith (DURANT; SMITH, 2006). Apenas se não houver melhorias significati-

[1] http://imdb.com/
[2] http://amazon.com/
[3] Esses classificadores são apresentados no Capítulo 2.

vas, recomenda-se o uso de outras características, um pré-processamento de documentos mais refinado ou a investigação das interações (ou co-citações) entre os documentos. Essas recomendações, portanto, partem das ideias mais simples para as mais complexas, evitando esforços desnecessários.

Aproveitando as revisões e análises apresentadas no Capítulo 3, decidiu-se fazer um estudo de caso envolvendo um *dataset* brasileiro. Não foi encontrado nenhum outro trabalho que aplique as técnicas dessa sub-área de Mineração de Opinião a documentos escritos em português - em particular, envolvendo conteúdo brasileiro. Considerando que 2010 foi ano de eleições presidenciais no Brasil, a abundância de artigos que carregam pontos de vista típicos da oposição e da situação foi explorada. Um corpus sobre o atual governo e as eleições foi construído, composto de material coletado em colunas, *sites* e *blogs* mantidos por jornalistas de notoriedade nacional. Em seguida, esse corpus foi dividido entre os pontos de vista pró-situação e pró-oposição, e foi classificado com um Naïve Bayes. As características dos documentos exploradas por esse classificador foram suas contagens de palavras. Os resultados obtidos com essa metodologia foram muito positivos: a taxa de acerto, em particular, foi de 89.43%. Isso significa que esses jornalistas consolidam suas perspectivas sobre a política brasileira já nas palavras que destacam. De fato, alguns trechos elencados no Capítulo 4, no qual o estudo é apresentado, sugerem que seus pontos de vista são defendidos com muita veemência.

Esse estudo de caso também investiga quais palavras receberam mais destaque por cada ponto de vista. A análise com um L-LDA[4], associando cada ponto de vista a um tópico e um tópico neutro a todos os documentos, indica que os jornalistas pró-situação dão mais ênfase ao candidato de oposição José Serra do que os pró-oposição. Esses últimos, por sua vez, dão maior destaque ao presidente Lula e à sua candidata, Dilma Rousseff - ambos correspondem à situação. Isso sugere que os jornalistas enfatizam o ataque aos candidatos que se opõem, ideologicamente, aos pontos de vista que eles defendem. Isso reforça a ideia, empiricamente observável, de que a defesa de um posicionamento, muitas vezes, compreende o ataque a posicionamentos opostos, como sugerem Somasundaran e Wiebe em seu trabalho sobre debates *online* (SOMASUNDARAN; WIEBE, 2009).

5.1 DIFICULDADES ENCONTRADAS

A classificação de documentos de acordo com seus pontos de vista sobre um tema é um problema relativamente novo. De fato, ele só foi estabelecido como sub-área da Mineração de

[4]O L-LDA é apresentado no Capítulo 2.

Opinião em 2008, na *survey* de Pang e Lee. Por este motivo, para entender melhor o problema, foi necessário buscar artigos em diversas conferências que envolviam a área de Mineração de Opinião. A filtragem de quais resultados realmente tinham a ver com o tema dessa monografia não foi exatamente difícil, mas envolveu um trabalho manual considerável. Para o estudo de caso, a extração, filtragem e padronização dos documentos que compõem o corpus também envolveu algum trabalho manual. Embora essas tarefas tenham sido realizadas de forma automatizada, a construção dos *scripts* dependia da compreensão de como o conteúdo estava disposto em cada veículo.

5.2 TRABALHOS FUTUROS

Futuramente, pretende-se estender o estudo de caso do Capítulo 4 a textos políticos escritos por cidadãos comuns em seus *blogs*, o que pode contribuir positivamente para a compreensão de como o brasileiro se posiciona politicamente na Web. Além disso, o estudo também deve ser ampliado para identificar os pontos de vista contidos nos comentários feitos aos artigos do corpus, em seus *blogs*, *sites* e colunas, 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 posicionamentos na mídia brasileira *online*. Caso essas tarefas de classificação não sejam bem resolvidas com o uso exclusivo de contagens de palavras, esforços direcionados na busca de características mais complexas, envolvendo aspectos sintáticos/semânticos dos documentos, devem ser feitos. É válido ressaltar que isso provavelmente implicaria no desenvolvimento de ferramentas gramaticais de apoio, voltadas para a língua portuguesa.

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 (CoLing)*. Manchester, United Kingdom: [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. Ultimo acesso em 09 de novembro de 2010. Disponível em: <http://www.comscore.com/press/release.asp?press=1928>.

DAVE, K.; LAWRENCE, S.; PENNOCK, D. M. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In: *Proceedings of the World Wide Web Conference (WWW)*. Budapest, Hungary: [s.n.], 2003. p. 519–528.

DURANT, K. T.; SMITH, M. D. Predicting the political sentiment of web log posts using supervised machine learning techniques coupled with feature selection. In: *Proceedings of the Knowledge Discovery on the Web International Conference on Advances in Web Mining and Web Usage Analysis*. Philadelphia, United States: [s.n.], 2006. p. 187–206.

EFRON, M. Cultural orientation: Classifying subjective documents by cociation analysis. In: *Proceedings of the Association for the Advancement of Artificial Intelligence Fall Symposium*. Arlington, United States: [s.n.], 2004. p. 41–48.

EVANS, M.; HASTINGS, N.; PEACOCK, B. *Statistical Distributions*. New York, United States: Wiley-Interscience, 2000. ISBN 0471371246.

FADER, A. et al. Mavenrank: Identifying influential members of the u.s. senate using lexical centrality. In: *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*. Prague, Czech Republic: [s.n.], 2007. p. 658–666.

GENTZKOW, M.; SHAPIRO, J. M. Media bias and reputation. p. 280–316, 2006.

GREENE, S.; RESNIK, P. More than words: Syntactic packaging and implicit sentiment. In: *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*. Boulder, United States: [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.

GROSECLOSE, T.; MILYO, J. A measure of media bias. p. 1191–1237, 2005.

HARDISTY, E. *Publicação online (mensagem pessoal)*. 2010. Mensagem recebida por alibezz@gmail.com (Aline Bessa) em 29 de Junho de 2010.

HIRST, G.; RIABININ, Y.; GRAHAM, J. Party status as a confound in the automatic classification of political speech by ideology. In: *Proceedings of the International Conference on Statistical Analysis of Textual Data*. Rome, Italy: [s.n.], 2010. p. 173–182.

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

KLEBANOV, B. B.; BEIGMAN, E.; DIERMEIER, D. Vocabulary choice as an indicator of perspective. In: *Proceedings of the Association for Computational Linguistics Conference*. Uppsala, Sweden: [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 International Joint Conference on Artificial intelligence*. Montreal, Canada: [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 the European Conference on Machine Learning (ECML)*. [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. In: *Proceedings of the Conference on Natural Language Learning (CoNLL)*. New York, United States: [s.n.], 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.

MCCALLUM, A.; NIGAM, K. P. A comparison of event models for naive bayes text classification. In: *Proceedings of the Association for the Advancement of Artificial Intelligence Workshop on Learning for Text Categorization*. Madison, United States: AAAI Press, 1998. p. 41–48.

MULLEN, T.; MALOUF, R. A preliminary investigation into sentiment analysis of informal political discourse. In: *Proceedings of the Association for the Advancement of Artificial Intelligence Spring Symposium*. Palo Alto, United States: [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.

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, 2008.

PANG, B.; LEE, L.; VAITHYANATHAN, S. Thumbs up? sentiment classification using machine learning techniques. In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Philadelphia, United States: [s.n.], 2002. p. 76–86.

RAMAGE, D. et al. Labeled lda: A supervised topic model for credit attribution in multi-labeled corpora. In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Downtown Core, Singapore: [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. Ultimo acesso em 09 de novembro de 2010. Disponível em: <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.156.2875>.

SOMASUNDARAN, S.; WIEBE, J. Recognizing stances in online debates. In: *Proceedings of the International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing (ACL-IJCNLP)*. Downtown Core, Singapore: [s.n.], 2009. p. 226–234.

SPIEGELHALTER, D.; RICE, K. Bayesian statistics. *Scholarpedia*, v. 4, n. 8, p. 5230, 2009.

TEUBERT, W. *A province of a federal superstate, ruled by an unelected bureaucracy - keywords of the euro-sceptic discourse in Britain.* [S.l.]: Ashgate, 2001. 45–86 p. ISBN 978-0-7546-1431-9.

THOMAS, M.; PANG, B.; LEE, L. Get out the vote: Determining support or opposition from Congressional floor-debate transcripts. In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Sydney, Australia: [s.n.], 2006. p. 327–335.

APÊNDICE A – TEOREMA DE BAYES E NOTAÇÕES

A Estatística se alia à Teoria da Probabilidade para estimar e analisar a ocorrência de eventos, como *Ganho de cem reais em sorteio* ou *Chuva ao meio-dia*. Dados dois eventos A e B, temos que a probabilidade *a priori* de A acontecer **ignora** a ocorrência do evento B. Essa probabilidade é representada pela notação $P(A)$ (SPIEGELHALTER; RICE, 2009). Analogamente, para o evento B, tem-se $P(B)$. Na prática sabe-se, intuitivamente, que a ocorrência de determinados eventos interfere no acontecimento de outros. Se às onze e meia da manhã observa-se o evento *Céu nublado*, a probabilidade do evento *Chuva ao meio-dia* ocorrer pode diferir daquela que ignora esse primeiro evento. Neste sentido, a probabilidade de um evento A ocorrer *dado* que B ocorreu recebe o nome de probabilidade condicional de A dado B, e é denotada por $P(A \mid B)$ (SPIEGELHALTER; RICE, 2009). Analogamente, tem-se $P(B \mid A)$. O Teorema de Bayes correlaciona essas probabilidades da seguinte forma (SPIEGELHALTER; RICE, 2009)

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)} \tag{A.1}$$

No contexto da Equação A.1, $P(A \mid B)$ recebe o nome de probabilidade *a posteriori* de A e $P(B \mid A)$ recebe o nome de *likelihood* (SPIEGELHALTER; RICE, 2009). O classificador Naïve Bayes se baseia em uma aplicação direta do Teorema de Bayes. Dados os eventos *Obtenção de um documento x* e *Obtenção de uma classe y*, o classificador deve estimar a ocorrência do segundo evento assumindo que o primeiro já ocorreu.