UNIVERSIDADE FEDERAL DA BAHIA

INSTITUTO DE MATEMÁTICA

DEPARTAMENTO DE CIÊNCIA DA COMPUTAÇÃO

Aline Duarte Bessa

PROVISORIO: Um estudo sobre *Opinion Mining*

PROVISORIO: Aspectos teóricos e práticos

Salvador

2010

Aline Duarte Bessa

PROVISORIO: Um estudo sobre *Opinion Mining*

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

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

Salvador

2010

RESUMO

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

Palavras-chave: monografia, graduação, projeto final.

ABSTRACT

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

Keywords: monograph, graduation, final project.

LISTA DE FIGURAS

2.1 Exemplo de modelo gráfico. 11

2.2 Modelo gráfico Naïve Bayes. 13

2.3 Modelo gráfico LDA. 15

LISTA DE ABREVIATURAS E SIGLAS

SUMÁRIO

1 Introdução **8**

1.1 Motivação 8

1.2 Proposta 9

1.3 Estrutura da Monografia 9

2 Técnicas básicas e ferramentas utilizadas **10**

2.1 Modelos Gráficos 10

2.1.1 Naïve Bayes 11

2.1.2 LDA 13

2.2 Classificadores 16

2.2.1 Naïve Bayes 17

2.2.2 SVMs 18

3 Principais trabalhos e *datasets* estudados **20**

4 Métodos baseados em frequências de palavras **21**

4.1 Introdução 21

4.2 Trabalhos Analisados 22

4.3 Experimentos com L-LDA e Naïve Bayes 22

4.4 Conclusão 25

5 Metodologias que usam informação extra-documento **27**

5.1 Concordância e discordância entre documentos 27

5.2 Meta-informações sobre os autores 27

6 Metodologias que usam relações intra-documento 28

7 Estudo de caso: Eleições 2010 29

7.1 Introdução 29

7.2 Seleção e pré-processamento do corpus 29

7.3 Identificando perspectivas com um classificador Naïve Bayes 29

7.4 Ilustrando a linguagem por perspectiva 29

7.5 Conclusões 30

8 Trabalhos relacionados 31

9 Conclusão 32

9.1 Dificuldades encontradas 32

9.2 Trabalhos futuros 32

Apêndice A – Resultados experimentais 33

Referências Bibliográficas 34

1 INTRODUÇÃO

1.1 MOTIVAÇÃO

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

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

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

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

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

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

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

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

1.2 PROPOSTA

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

1.3 ESTRUTURA DA MONOGRAFIA

Falar da metodologia de busca dos artigos

2 TÉCNICAS BÁSICAS E FERRAMENTAS UTILIZADAS

Introduzir o capítulo quando tudo já estiver escrito.

2.1 MODELOS GRÁFICOS

Modelos gráficos consistem na representação, através de um grafo, das relações entre um conjunto finito de variáveis aleatórias, provendo uma maneira simples de se representar distribuições de probabilidade (BISHOP, 2006). Cada vértice do grafo corresponde a uma variável aleatória (ou a um conjunto de variáveis aleatórias) ou a um parâmetro do modelo, e cada aresta reflete a relação entre dois vértices. Modelos gráficos são categorizados como dirigidos ou não-dirigidos. Por questões de escopo, apenas modelos dirigidos (também conhecidos como Redes Bayesianas) serão discutidos nesta seção.

Em um modelo gráfico dirigido, tem-se um grafo direcionado acíclico que representa a distribuição de probabilidade conjunta[1] para suas variáveis aleatórias. Cada aresta corresponde a uma distribuição de probabilidade condicional, incidindo no vértice cuja distribuição de probabilidade está condicionada ao valor do vértice de onde ela parte. Quando mais de uma variável tem distribuição de probabilidade condicionada aos mesmos vértices, é possível sintetizar a notação representando todas elas com um único vértice. Neste caso, o vértice fica dentro de um retângulo rotulado com o número de variáveis que ele representa.

Por fim, vértices representados com círculos brancos correspondem a variáveis latentes - ou seja, cujos valores não são observáveis diretamente no conjunto de dados ao qual o modelo é aplicado; círculos cinzas, por sua vez, correspondem a variáveis observáveis, cujos valores estão explícitos no conjunto de dados. Variáveis latentes permitem que distribuições de probabilidade muito complexas, envolvendo variáveis observáveis, sejam construídas a partir de distribuições

[1] Nesta monografia, os termos "distribuição de probabilidade conjunta"e "distribuição conjunta"serão utilizados de forma intercambiável.

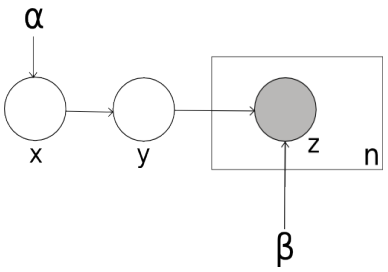

Figura 2.1: Exemplo de modelo gráfico.

condicionais mais simples (BISHOP, 2006). A Figura 2.1 corresponde a um modelo gráfico com a seguinte distribuição conjunta

$$P(x|\alpha)P(y|x)\prod_{i=1}^{n}P(z_i|y,\beta) \tag{2.1}$$

α e β são parâmetros de distribuições de probabilidade, x e y são variáveis latentes e $z_1,...,z_n$, sintetizadas na Figura 2.1 através do vértice z, são variáveis observáveis.

Uma categoria de modelos gráficos explorada neste projeto são os **modelos generativos**. Eles associam distribuições de probabilidade a todas as variáveis aleatórias envolvidas, permitindo a geração - i.e. simulação - de seus valores. Modelos generativos são úteis para expressar os processos pelos quais dados observáveis são obtidos. Em um modelo deste tipo, os valores das variáveis aleatórias podem ser obtidos através de técnicas de amostragem aplicadas à distribuição de probabilidade conjunta. Além do Naïve Bayes, discutido na seção 2.1.1, os modelos gráficos generativos *Latent Dirichlet Allocation* (LDA) e *Labeled Latent Dirichlet Allocation* (L-LDA) foram empregados em experimentos ao longo de todo o projeto. Uma discussão sobre eles pode ser encontrada na seção 2.1.2.

2.1.1 NAÏVE BAYES

O Naïve Bayes é um modelo gráfico generativo que assume a independência condicional das características $F_1,...,F_k$ presentes em um conjunto de documentos D. Se D é composto de documentos de texto, estas características normalmente correspondem a todas as suas palavras distintas. Isto equivale a assumir, portanto, que a presença de uma palavra em um documento qualquer não é informativa sobre a presença de nenhuma outra.

A finalidade básica do Naïve Bayes é estimar a probabilidade de um documento d pertencer a uma certa classe c. Para isto, d é representado de forma simplificada, através de um vetor v_d em que cada posição corresponde a uma de suas n palavras. Com esta representação, a

probabilidade de d pertencer a uma classe c pode ser calculada via Teorema de Bayes como

$$P(c|v_{d1},...,v_{dn}) = \frac{p(c) \times p(v_{d1},...,v_{dn}|c)}{p(F_1,...,F_k)} \tag{2.2}$$

Como o Naïve Bayes assume que as palavras dos documentos são condicionalmente independentes, a equação 2.2 pode ser reescrita como

$$P(c|v_{d1},...,v_{dn}) = \frac{p(c) \times p(v_{d1}|c)p(v_{d2}|c)...p(v_{dn-1}|c)p(v_{dn}|c)}{p(F_1,...,F_k)} = \frac{p(c) \times \prod_{i=1}^{n} p(v_{di}|c)}{p(F_1,...,F_k)} \tag{2.3}$$

Como o Naïve Bayes é um modelo generativo, ele permite que se simule a criação de um documento d, pertencente a uma classe c, através da amostragem de suas variáveis aleatórias. Sem perda de generalidade, assume-se que c é uma variável aleatória que pode assumir m valores naturais distintos, variando de 0 a $m-1$. Cada valor corresponde a uma classe diferente, sendo escolhido de acordo com

$$c \sim Binomial(m-1, \pi) \tag{2.4}$$

Antes de iniciar o processo de geração de documentos, define-se um parâmetro π para a distribuição binomial em 2.4 de acordo com

$$\pi \sim Beta(\alpha, \beta) \tag{2.5}$$

α e β são denominados *hiperparâmetros*, pois são parâmetros de uma distribuição através da qual se escolhe um dos parâmetros do modelo - no caso, π (RESNIK; HARDISTY, 2010). Após uma classe ter sido fixada de acordo com 2.4, seleciona-se uma palavra para cada posição j do vetor v_d, de acordo com uma distribuição de probabilidade sobre $F_1,...,F_k$

$$v_{dj} \sim Multinomial(F_1,...,F_k, \theta_c) \tag{2.6}$$

A distribuição utilizada depende do valor de c amostrado anteriormente, de modo que há m parâmetros θ_c. Cada θ_c é escolhido antes do processo de geração dos documentos, de acordo com

$$\theta_c \sim Dirichlet(\gamma_c) \tag{2.7}$$

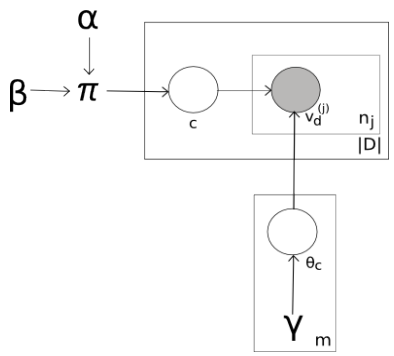

Figura 2.2: Modelo gráfico Naïve Bayes.

γ_c, assim como α e β, é um hiperparâmetro.

A distribuição conjunta para este modelo generativo é dada por

$$P(\pi|\alpha,\beta)\prod_{i=0}^{m-1}P(\theta_{c=i}|\gamma_{c=i})\prod_{j=1}^{|D|}P(c_j|\pi)P(v_d^{(j)}|\theta_{c_j},c_j) \tag{2.8}$$

em que c_j é a classe selecionada para o j-ésimo documento de D e $v_d^{(j)}$ é seu vetor de palavras. A Figura 2.2 representa esta distribuição conjunta de forma gráfica, com vértices para variáveis aleatórias e relações de probabilidade condicional evidenciadas pelas arestas.

2.1.2 LDA

O modelo LDA parte da ideia de que um documento pode tratar de múltiplos tópicos, refletidos nas palavras que o compõem (GRIFFITHS; STEYVERS, 2004). Assim como no modelo Naïve Bayes, palavras podem ser geradas de acordo com distribuições multinomiais específicas. A diferença é que, no LDA, cada palavra é gerada a partir de uma mistura de tópicos; no Naïve Bayes, elas são geradas a partir de um só tópico (classe) (CARPENTER, 2010).

O LDA associa as palavras dos documentos a tópicos diferentes, com maior ou menor probabilidade, criando agrupamentos que se relacionam semanticamente. Os tópicos em um LDA são variáveis latentes, cujos significados requerem uma interpretação posterior ao processamento. Esta interpretação baseia-se nas relações semânticas entre as palavras que se associaram mais fortemente a cada um deles.

Para ilustrar como as palavras evidenciam o significado de um tópico, um experimento envolvendo receitas culinárias extraídas do *site* **allrecipes.com** foi executado. Apenas os ingredientes de cada receita foram considerados. Na Tabela 2.1, constam as cinco palavras mais fortemente associadas a quatro tópicos, de acordo com o LDA.

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

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

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

O modelo LDA trata cada documento pertencente a um conjunto de documentos D como uma mistura de tópicos, representada por uma distribuição de probabilidade sobre um conjunto de tópicos T. Cada tópico, por sua vez, é visto como uma mistura de palavras, representada por uma distribuição sobre todas as palavras distintas de D. Para cada documento $d \in D$, é fixada uma distribuição de probabilidade sobre tópicos θ_d, condicionada a um hiperparâmetro α

$$\theta_d \sim Dirichlet(\alpha) \tag{2.9}$$

Para cada tópico $t \in T$, é fixada uma distribuição de probabilidade ϕ_t sobre palavras, condicionada a um hiperparâmatro β

$$\phi_t \sim Dirichlet(\beta) \tag{2.10}$$

Em seguida, para cada uma das n palavras de d, um tópico t é escolhido, de acordo com θ_d

$$t \sim Discrete(\theta_d) \tag{2.11}$$

e uma palavra w é gerada de acordo com ϕ_t

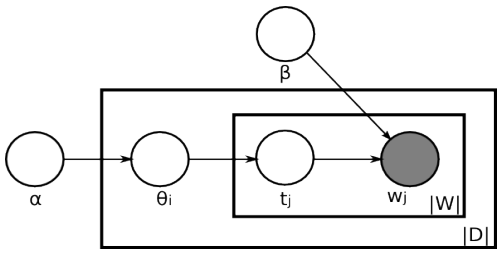

Figura 2.3: Modelo gráfico LDA.

$$w \sim Discrete(\phi_t) \tag{2.12}$$

A distribuição conjunta do modelo LDA é dada por

$$\prod_{i=1}^{|D|} \left\{ P(\theta_i|\alpha) \left[\prod_{j=1}^{|W|} P(t_j|\theta_i) P(w_j|t_j,\beta) \right] \right\} \tag{2.13}$$

$P(w_j|t_j,\beta)$ reflete o quanto a palavra w_j se relaciona com o tópico t_j; $P(t_j|\theta_i)$, por sua vez, funciona como uma medida do quanto o tópico t_j é importante no contexto do documento d_i (GRIFFITHS; STEYVERS, 2004). Na Figura 2.3, tem-se o modelo gráfico do LDA correspondente à distribuição conjunta 2.13.

A implementação do LDA utilizada para estes experimentos está disponível no repositório *online* de Alexandre Passos (**??**), e o número de iterações para amostragem de tópicos e palavras foi fixado em 100.

L-LDA

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

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

Experimentos com o L-LDA foram desenvolvidos ao longo deste projeto, associando cada tópico, por exemplo, a uma perspectiva a ser minerada. Após a execução do modelo, as palavras mais fortemente associadas a cada perspectiva ilustram como os assuntos discutidos são enfocados por elas. Quanto mais duas perspectivas se distanciam, mais diferentes são as palavras que se associam com destaque a cada uma delas.

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

A implementação de L-LDA utilizada neste projeto também está disponível no repositório *online* de Alexandre Passos (**??**). O número de iterações para amostragem de tópicos e palavras, em todos os experimentos, foi fixado em 100.

2.2 CLASSIFICADORES

A classificação de documentos de texto de acordo com suas perspectivas é um dos principais objetivos dos trabalhos revisados neste projeto. Grande parte deles utiliza os classificadores Naïve Bayes ou *Support Vector Machines* (SVMs), apresentados respectivamente nas seções 2.1.1 e 2.2.2, como parte de suas metodologias. O desempenho destes classificadores é comumente medido através das seguintes métricas: **taxa de acerto**, **precisão**, **rechamada** ou **métrica F1**.

A taxa de acerto é definida pela razão entre o número de documentos classificados corretamente e todos os documentos avaliados. A precisão, medida para uma classe c qualquer, é definida pela razão entre o número de documentos classificados corretamente como c e todos os documentos classificados como c. A rechamada, medida também para uma classe c qualquer, é definida pela razão entre o número de documentos classificados corretamente como c e a soma deste valor com o número de documentos classificados erroneamente para todas as demais classes. A métrica F1, também medida para uma classe c qualquer, é dada por

$$2 \times \frac{precisao \times rechamada}{precisao + rechamada} \tag{2.14}$$

Estas métricas revelam aspectos diferentes do desempenho de um classificador. Por este

motivo, é comum encontrar mais de uma delas sendo utilizada no mesmo contexo. 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).

2.2.1 NAÏVE BAYES

O modelo Naïve Bayes foi apresentado na seção 2.1.1. Nesta seção, será discutido como construir um classificador de documentos a partir dele. Esta seção trata de documentos de texto em particular, por se tratarem do objeto básico de estudo deste projeto. Ra Sabe-se que, em um Naïve Bayes, assume-se que as as características em um documento são condicionalmente independentes, o que equivale a afirmar, por exemplo, que a presença de uma palavra em um documento de texto não é informativa sobre a presença de nenhuma outra. Apesar desta hipótese simplificar bastante a estrutura linguística de um texto, classificadores construídos a partir do modelo Naïve Bayes, denominados classificadores Naïve Bayes, reportam um bom desempenho em várias tarefas de classificação baseadas em palavras (**??**) (**??**).

Dados um documento d pertencente a um conjunto de documentos D, todas as palavras distintas de D, $F_1, ..., F_k$, uma variável aleatória c, que representa as possíveis classes de d, e um vetor v_d, em que cada posição corresponde a uma de suas n palavras, tem-se que

$$P(c|v_{d1}, ..., v_{dn}) = \frac{p(c) \times \prod_{i=1}^{n} p(v_{di}|c)}{p(F_1, ..., F_k)} \tag{2.15}$$

conforme discutido anteriormente na seção 2.1.1. Um classificador Naïve Bayes deve rotular o documento d com o valor de c que maximiza a equação 2.15. Como o denominador na equação 2.15 é o mesmo para todas as classes, ele pode ser ignorado nestes cálculos.

Normalmente, classificadores Naïve Bayes são utilizados de forma semi-supervisionada. Isto significa que eles são submetidos a uma etapa de treinamento, na qual aprendem as classes associadas a alguns documentos, e a uma etapa de classificação, na qual devem simular o processo gerador destes documentos e utilizar esta informação para classificar outros. Basicamente, as informações aprendidas na etapa de treinamento modelam as distribuições das palavras de D por classe, gerando parâmetros para as distribuições de probabilidade envolvidas na classificação de outros documentos.

Todos os experimentos com um classificador Naïve Bayes conduzidos neste projeto utilizam a implementação disponível no repositório *online* de Aline Bessa (**??**). O número de

iterações para a amostragem de documentos e classes, em todos os experimentos, foi fixado em 500.

2.2.2 SVMS

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

Dado um conjunto de n pontos $\{x_i, y_i\}$, onde x_i é a representação vetorial de um documento d em um espaço euclidiano $\mathbb{R}^M$ e y_i é sua respectiva classe, $y_i \in \{-1, 1\}$, um SVM deve decidir a classe y de um novo documento representado pelo vetor x. Para isso, assume-se que há pelo menos um hiperplano θ_0 que separa os pontos com $y_i = 1$ daqueles com $y_i = -1$. Um hiperplano θ_0 pode ser definido como o conjunto de pontos $\mathbf{x}$ que satisfazem

$$\mathbf{x} \cdot \mathbf{w} + b = 0 \tag{2.16}$$

$\mathbf{w}$ é a normal ao hiperplano e $|b|||\mathbf{w}||$ é sua distância perpendicular à origem (OGURI, 2006). A ideia é escolher os parâmetros $\mathbf{w}$ e b que maximizem a soma das distâncias dos hiperplanos θ_1 (vide equação 2.17) e θ_{-1} (vide equação 2.18) ao hiperplano θ_0.

$$\mathbf{x} \cdot \mathbf{w} + b = 1 \tag{2.17}$$

$$\mathbf{x} \cdot \mathbf{w} + b = -1 \tag{2.18}$$

θ_1 e θ_{-1} podem ser encontrados minimizando-se

$$\frac{1}{2}||\mathbf{w}||^2 \tag{2.19}$$

Para realizar esta otimização mais facilmente, o problema pode ser remodelado com multiplicadores de Lagrange $\{\alpha_i\}$, $1 \leq i \leq n$, levando à Equação 2.20 (OGURI, 2006). Busca-se,

então, a minimização desta equação com relação a $\mathbf{w}$ e b e maximização com relação a $\{\alpha_i\}$, com todo $\alpha_i \geq 0$.

$$L(\alpha, b, \mathbf{w}) = \frac{1}{2}||\mathbf{w}||^2 - \sum_{i=1}^{n} \alpha_i \left[y_i(x_i \cdot \mathbf{w} + b) - 1 \right] \tag{2.20}$$

Após a obtenção dos valores de $\{\alpha_i\}$ que maximizam 2.20, a obtenção da classe y de um documento representado por um vetor x é dada pelo sinal do somatório

$$y(x) = sinal\left(\sum_{i=1}^{n} \alpha_i y_i (x_i \cdot x) + b \right) \tag{2.21}$$

Esta solução funciona em casos nos quais os pontos $\{x_i, y_i\}$ são linearmente separáveis - ou seja, obedecem à restrição

$$y_i(x_i \cdot \mathbf{w} + b - 1) \geq 0, \; i = 1, ..., n \tag{2.22}$$

Quando esses pontos não são linearmente separáveis, essa metodologia precisa ser ajustada, modelando a classificação errônea de documentos. Isto envolve a introdução de n variáveis frouxas ε_i[2], uma para cada ponto (x_i, y_i). $\varepsilon_i = 0$ se $y(x_i) = y_i$ e $\varepsilon_i = |y_i - y(x_i)|$ em caso contrário (BISHOP, 2006). O SVM deve, neste caso, minimizar

$$C \sum_{i=1}^{n} \varepsilon_i + \frac{1}{2}||\mathbf{w}||^2 \tag{2.23}$$

em que C é um parâmetro responsável por controlar o compromisso entre a penalidade das variáveis frouxas e a distância máxima dos hiperplanos θ_1 e θ_{-1} ao hiperplano θ_0. Na modelagem com multiplicadores de Lagrange, a Equação 2.20 deve ser otimizada de tal forma que todo α_i deve ser maximizado obedecendo à restrição $0 \leq \alpha_i \leq C$. Desta forma, também se obtém a Equação 2.21 para determinação da classe de um novo documento.

SVMs não foram utilizados em nenhum experimento deste projeto, mas fazem parte da metodologia de alguns dos trabalhos revisados.

[2]Do inglês *slack variables*.

3 PRINCIPAIS TRABALHOS E DATASETS *ESTUDADOS*

4 *MÉTODOS BASEADOS EM FREQUÊNCIAS DE PALAVRAS*

4.1 INTRODUÇÃO

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

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

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

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

4.2 TRABALHOS ANALISADOS

4.3 EXPERIMENTOS COM L-LDA E NAÏVE BAYES

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

4.4 CONCLUSÃO

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

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

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

5 METODOLOGIAS QUE USAM INFORMAÇÃO EXTRA-DOCUMENTO

5.1 CONCORDÂNCIA E DISCORDÂNCIA ENTRE DOCUMENTOS

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

5.2 META-INFORMAÇÕES SOBRE OS AUTORES

6 METODOLOGIAS QUE USAM RELAÇÕES INTRA-DOCUMENTO

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

7 ESTUDO DE CASO: ELEIÇÕES 2010

7.1 INTRODUÇÃO

Onde se motiva o estudo das eleições e fala pq tratará de colunas de jornalistas de notoriedade nacional. TUDO NESTE CAPÍTULO PRECISA SER BEM DIDÁTICO.

7.2 SELEÇÃO E PRÉ-PROCESSAMENTO DO CORPUS

Onde se fala dos critérios utilizados na escolha dos jornalistas (justifique com trechos e talz), download, tratamento de tags, período estudado, palavras-chave, número de caracteres. Idiossincrasias da coisa.

7.3 IDENTIFICANDO PERSPECTIVAS COM UM CLASSIFICADOR NAÏVE BAYES

Onde se explica detalhadamente as perspectivas do corpus, justifica-se o método, mostra-se o desempenho via 10-fold-cross-validation (fala pq escolheu essa também), discute-se questões de estilo e compara com o desempenho de outros artigos.

7.4 ILUSTRANDO A LINGUAGEM POR PERSPECTIVA

Onde se explica o uso do L-LDA, o overlap de palavras, mostra-se snippets e talz.

7.5 CONCLUSÕES

Onde se discute o que tudo isso quer dizer sobre nossa mídia, se fala um pouco da questão da linguística de corpus, como os jornais n necessariamente assumem um ponto de vista, mas contratam colunistas e talz. Amarrar questões futuras.

8 TRABALHOS RELACIONADOS

9 CONCLUSÃO

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

9.1 DIFICULDADES ENCONTRADAS

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

9.2 TRABALHOS FUTUROS

Pode-se indicar como trabalhos futuros:

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

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

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

APÊNDICE A – RESULTADOS EXPERIMENTAIS

No no nnononono no n ono o nn.

REFERÊNCIAS BIBLIOGRÁFICAS

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

CARPENTER, B. *Integrating out multinomial parameters in latent Dirichlet allocation and naive Bayes for collapsed Gibbs sampling*. [S.l.], 2010. Disponível em: <http://lingpipe-blog.com/2010/07/13/collapsed-gibbs-sampling-for-lda-bayesian-naive-bayes/>.

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

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

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

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

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

RESNIK, P.; HARDISTY, E. *Gibbs Sampling for the Uninitiated*. [S.l.], Abril 2010. Disponível em: <http://hdl.handle.net/1903/10058>.