



UNIVERSIDADE FEDERAL DA BAHIA
INSTITUTO DE MATEMÁTICA
DEPARTAMENTO DE CIÊNCIA DA COMPUTAÇÃO

Aline Duarte Bessa

PROVISÓRIO: Um estudo sobre *Opinion Mining*
PROVISÓRIO: 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 nonono nonono 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 nonono nonno. Nonono nonono nonono, nonono, nonono nonono nonono nononono nonno. Nonono nonono nonono, nonono, nonono nonono nonono nononono nonno. Nonono nonono nonono, nonono, nonono nonono nonono nononono nonno. Nonono nonono nonono, nonono, nonono nonono nonono nononono nonno.

Keywords: monograph, graduation, final project.

LISTA DE FIGURAS

LISTA DE ABREVIATURAS E SIGLAS

SUMÁRIO

1	Introdução	7
1.1	Motivação	7
1.2	Proposta	8
1.3	Estrutura da Monografia	8
2	Relação entre as características dos <i>datasets</i> e as metodologias utilizadas	9
2.1	Palavras utilizadas nos documentos	10
3	Conclusão	15
3.1	Dificuldades encontradas	15
3.2	Trabalhos futuros	15
	Apêndice A – Resultados experimentais	16
	Referências Bibliográficas	17

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;

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

3. **Classificação de pontos de vista ou perspectivas** - Dado um conjunto de documentos contendo perspectivas ou pontos de vista sobre um mesmo tema/conjunto de temas, classifique-os de acordo com essas perspectivas/pontos de vista;
4. **Reconhecimento de humor** - Dado um conjunto de textos com caráter emotivo/sentimental, como posts de blogs pessoais, identifique que tipos de humor permeiam os textos e/ou classifique-os de acordo com as diferentes emoções encontradas.

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

1.2 PROPOSTA

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

1.3 ESTRUTURA DA MONOGRAFIA

2 **RELAÇÃO ENTRE AS CARACTERÍSTICAS DOS DATASETS E AS METODOLOGIAS UTILIZADAS**

Os *datasets* estudados nesse projeto são oriundos de fontes diversas, incluindo *blogs* (??) (??), matérias jornalísticas (??) (??), artigos escritos por especialistas (??) (??), discussões *online* (??) (??) e debates políticos (??) (??). Os assuntos discutidos também são bastante variados, incluindo tópicos relativamente abstratos, como a discussão da pena de morte (??), e outros mais objetivos, como possíveis *designs* para um controle remoto (??) (??). As línguas empregadas nos documentos diferem bastante de um trabalho para outro, variando tanto na informalidade dos termos e construções empregadas quanto no teor opinativo das colocações **siglo citando?**. Outra característica importante, que distingue um estudo de outro, envolve a língua - ou línguas - nas quais os documentos se encontram. **Ler um pouco sobre isso para amadurecer este ponto** Por fim, o tamanho dos textos analisados, que varia de algumas sentenças a vários parágrafos, bem como o nível de engajamento de seus autores com as perspectivas defendidas, indica uma Web muito plural no que diz respeito aos tipos de conteúdo *online*.

Nos trabalhos estudados para este projeto, percebeu-se que as características inerentes a cada *dataset* pouco interferem na decisão dos métodos utilizados na mineração das perspectivas dos documentos. No decorrer deste capítulo, a forte relação que existe entre essas características e a escolha das metodologias será discutida, justificando parcialmente os resultados ruins encontrados em alguns artigos. Adicionalmente, através de experimentos em *datasets* referenciados nesses estudos, ou coletados *online*, este capítulo apontará possibilidades metodológicas que podem conduzir a melhorias nos resultados analisados. **Devo enfatizar a originalidade disso aqui? Acho q n, né? Fica na problematização.** O capítulo está estruturado da seguinte forma: **blablabla**. Por fim, na **Seção Y**, algumas combinações de características comuns em documentos da Web, como alto teor de linguagem opinativa em debates informais *online* (??),

serão analisadas conjuntamente.

2.1 PALAVRAS UTILIZADAS NOS DOCUMENTOS

Uma hipótese apresentada em (??), assumida por parte dos artigos estudados para este projeto, é de que a escolha de palavras em um documento reflete os pontos de vista e intenções de seu autor. O emprego de palavras semanticamente distintas para um mesmo propósito - como *Revolução* ou *Golpe* para o começo do Regime Militar Brasileiro em 1964 -, e também a frequência de seus usos, são elementos chave para a transmissão de posicionamentos diferentes sobre um determinado assunto. Essa hipótese encontra respaldo em (??), um estudo de Linguística de Corpus (??)(??) que indica que indivíduos defendendo perspectivas diferentes consolidam seus vocabulários através do uso de palavras específicas (*stigma words* e *banner words*), facilitando a identificação de adversários e aliados.

A hipótese, entretanto, não é comprovada por todos os *datasets* analisados. Em alguns deles, o conhecimento das palavras empregadas para cada perspectiva no *dataset*, bem como suas frequências, não é suficiente para inferir o perfil ideológico dos autores dos documentos. (??) prevê este comportamento, defendendo que o vocabulário usado em dois lados de uma discussão tende a ser basicamente o mesmo, o que contribui para o mau desempenho de classificadores baseados em frequências de palavras exclusivamente, como Naïve Bayes e SVM padrão. Esta ideia é explorada novamente em (??), a fim de justificar a taxa de acerto de apenas 63.59% obtida na aplicação de um classificador Naïve Bayes a um *dataset* de debates políticos *online* (??).

Se um classificador utiliza apenas a informação contida na presença e nas frequências das palavras do corpus, para identificar a perspectiva de seus documentos, é natural que sua taxa de acerto seja tão mais baixa quanto menos essas características mudam de uma perspectiva para outra. Apesar desse princípio, quantificar a relação entre as palavras e suas frequências e o desempenho desses classificadores não é trivial. A depender do *dataset*, características como tamanho dos documentos e quantidade de documentos para treinamento dos classificadores contribuem de maneira mais decisiva para suas taxas de acerto¹. Por este motivo, uma análise prévia das palavras presentes no corpus, a fim de identificar quais delas se associam mais fortemente a cada uma das perspectivas e quais co-ocorrem em perspectivas diferentes, não necessariamente evidencia o bom ou o mau desempenho que um classificador como Naïve Bayes ou SVM padrão

¹A relação entre essas características e os métodos de classificação é discutida nos capítulos A e B dessa monografia.

apresentará².

Para comprovar essa hipótese, foram executados três experimentos com um modelo de tópicos do tipo L-LDA, **apresentado na seção XYZ dessa monografia**. Para cada experimento, todos os documentos envolvidos foram marcados com um rótulo referente à perspectiva e outro genérico, idêntico para todos eles. Essa estratégia facilitou a identificação de quais palavras melhor se associam a cada perspectiva e quais colaboram para uniformizar o vocabulário do corpus. Intuitivamente, uma pré-análise possível envolveria a simples contabilização, e posterior comparação, das frequências das palavras no *dataset*, separadas por perspectiva. Essa metodologia, entretanto, não evidencia quais são as palavras que colaboram para a uniformização do corpus, algo que se obtém com o emprego de um tópico genérico em um L-LDA. Todos os experimentos utilizaram a implementação de L-LDA disponível em (??). Para medir a performance de um classificador baseado em frequências de palavras, foi utilizada a implementação de Naïve Bayes disponível em (??).

Para o primeiro experimento, foi utilizado o *dataset* do artigo (??), composto de artigos do *site* bitterlemons.org escritos sob uma perspectiva pró-Palestina ou pró-Israel. Para o L-LDA, cada documento estava associado a dois tópicos: sua perspectiva (pró-Palestina ou pró-Israel) e um genérico, idêntico para todos eles. Esses artigos foram escritos ou pelos editores do *site* ou por convidados, o que cria uma divisão natural corpus. Essa divisão foi utilizada em (??) para avaliar o desempenho de um Naïve Bayes: em um momento, os documentos de treino eram os escritos pelos editores e os de teste, aqueles escritos pelos convidados; em outro, tinha-se a situação inversa. Neste primeiro experimento, treinamos o Naïve Bayes com os documentos escritos pelos convidados e testamos o desempenho do classificador com os documentos escritos pelos editores. A situação inversa não precisou ser verificada para comprovar a hipótese desse experimento. É válido ressaltar que, em todos os experimentos, todas as palavras contidas nos documentos foram consideradas. Isto resultou em análises independentes de pré-processamentos comuns, como *stemming*³ e retirada de *stop words*⁴.

tabelas com palavras e desempenhos

No segundo experimento, foi utilizado um dos *datasets* estudados em (??), composto de colocações em um debate sobre *browsers* disponível no *site* convinceme.net. A discussão divide-se em apenas dois lados: pró-Firefox e pró-Internet Explorer. Os documentos foram rotulados de forma análoga aos artigos do primeiro *dataset* e o Naïve Bayes foi testado com

²Apesar de alguns artigos utilizarem outras metodologias e informações para classificar documentos de acordo com suas perspectivas, a maior parte deles desenvolve experimentos com Naïve Bayes e SVMs padrão.

³escrever sobre isso em algum lugar

⁴mesma coisa

uma **4-fold-cross-validation**. Este *dataset* é pequeno, contendo apenas 959 palavras diferentes.

Dos artigos estudados para este projeto, relacionados diretamente à classificação de documentos de acordo com suas perspectivas, três associaram o mau desempenho dos classificadores Naïve Bayes e SVM padrão à homogeneização do vocabulário contido no corpus: (??), (??) e (??). Não foi possível executar experimentos com o L-LDA em nenhum dos *datasets* utilizados nesses trabalhos, pois eles não estão disponíveis na Web nem conseguiram ser obtidos mediante pedido, por *e-mail*, aos autores dos artigos. Por este motivo, esta seção se limitará a descrever as técnicas utilizadas nesses trabalhos para melhorar as taxas de acerto na classificação dos corpora⁵.

Em (??) e (??), o *dataset* estudado é o mesmo: um conjunto de debates políticos Estadunidenses extraídos do *site* www.politics.com. Os dois trabalhos visavam a classificar os 185 participantes da discussão de acordo com suas orientações políticas: Esquerda ou Direita. Cada participante era representado por um único documento, resultante da concatenação de todas as suas falas nos debates. Aplicando um Naïve Bayes na coleção de documentos, a taxa de acerto obtida em (??) foi de 60.37%; em (??), 63.59%. (??) analisa o *dataset* e conclui que 62.2% das falas de Esquerda mencionam trechos de falas de Direita. Quanto às falas de Direita, 77.5% delas mencionam falas de Esquerda. Essa forma de interação entre os participantes é explorada em (??). Neste artigo, 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. A ideia defendida é de que quão mais similares forem os padrões de citação de dois participantes, mais provável é a hipótese de que eles possuem uma mesma perspectiva política. Para agrupar participantes, a estratégia utilizada é a seguinte:

1. Dada a matriz de adjacência do grafo de co-citação, \mathbf{M} , computa-se uma aproximação com posto menor, \mathbf{M}' , via SVD⁶;
2. calcula-se a distância entre os vértices de \mathbf{M}' ;
3. agrupa-se os participantes da discussão de acordo com essa informação, através de algoritmo especificado em (??);
4. concatena-se todas as falas referentes a cada grupo obtido, gerando uma coleção menor de documentos;

⁵Técnicas de classificação desenvolvidas por artigos que não tratam da questão da uniformidade das palavras serão apresentadas em outras seções desta monografia.

⁶A técnica de aproximação para um posto menor via SVD tem que estar escrita em algum lugar! senão não escrevo aí.

5. aplica-se um Naïve Bayes a essa nova coleção de documentos;
6. os resultados obtidos com o Naïve Bayes são propagados para todos os participantes de cada grupo.

O uso da matriz \mathbf{M}' , em vez de \mathbf{M} , justifica-se por \mathbf{M}' possuir menos ruído e evidenciar informações estruturais do grafo, como padrões de comunicação entre os vértices, melhor do que $\mathbf{M}(??)$. Essa metodologia atinge resultados significativamente melhores que o simples uso de um Naïve Bayes: para participantes com mais de 500 palavras de fala, a taxa de acerto relatada é de 73%.

Grafozinho do Taking sides

Padrões de citação entre documentos também foram investigados em (??). Neste artigo, os experimentos envolvem dois *datasets*: o primeiro é composto de artigos políticos Estadunidenses de Direita ou Esquerda, coletados em *sites* e *blogs* políticos - explicitamente partidários ou não -; o segundo é composto de textos retirados de *sites* de artistas musicais, divididos entre as categorias Alternativo e Popular. As taxas de acerto relatadas para as aplicações de Naïve Bayes nestes corpora foram, respectivamente, 64.71% e 50.1%. Um SVM padrão foi aplicado apenas ao primeiro *dataset*, pois não havia recursos computacionais suficientes para a aplicação no segundo. A taxa de acerto verificada foi significativamente melhor: 72.96%. A fim de melhorar as taxas de acerto na classificação dos documentos, (??) estima a perspectiva de cada um deles avaliando a probabilidade de que eles sejam co-citados com documentos referência de cada perspectiva. Para o primeiro corpus, a metodologia relatada é a seguinte:

1. Duas listas de URLs são criadas: uma contendo *sites* referência de Esquerda; outra, de Direita;
2. para cada documento d_i , computa-se sua probabilidade de ser co-citado com alguma URL da lista de Direita e, em seguida, com alguma da lista de Esquerda;
3. se a razão entre estas probabilidades for maior que 1, o documento é classificado como de Direita; caso seja menor que 1, como de Esquerda.

As probabilidades foram computadas com o auxílio do buscador Altavista, que fornecia o número de documentos indexados que citavam, simultaneamente, d_i e alguma das URLs das listas de referência. Esta metodologia, aplicada ao primeiro *dataset*, resultou em uma taxa de acerto de 94.1%. No segundo *dataset*, o procedimento foi análogo: as listas de URLs foram divididas entre Alternativo e Popular. A taxa de acerto máxima obtida foi de 88.84%.

Concluir o capítulo.

3 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

3.1 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

3.2 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 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