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.

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.

Keywords: monograph, graduation, final project.

LISTA DE FIGURAS

LISTA DE ABREVIATURAS E SIGLAS

SUMÁRIO

1 INTRODUÇÃO

1.1 MOTIVAÇÃO

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

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

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

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

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

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

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

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

1.2 PROPOSTA

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

1.3 ESTRUTURA DA MONOGRAFIA

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 linguagens 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 **sigo 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. 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 padrão a um *dataset* de debates políticos *online* (**??**).

Se um classificador utiliza apenas as palavras de um documento, bem como suas frequências, para identificar sua perspectiva, é natural que ele erre muito se essas características não mudam muito de uma perspectiva para outra. Apesar disso, verificar a relação existente entre a uniformidade das palavras e suas frequências e o desempenho desses classificadores não é trivial. A depender do *dataset*, características como tamanho dos documentos e quantidade de palavras por documento contribuem de maneira mais decisiva para a taxa de acerto dos classificadores[1]. 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 desse tipo apresentará.

Para comprovar essa hipótese, foram executados três experimentos com um modelo de tópi-

[1] A relação entre essas características e os métodos de classificação é discutida nas seções BLAAAAAA E BLIII dessa monografia.

cos 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*[2] e retirada de *stop words*[3].

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

Dos artigos estudados para este projeto, relacionados diretamente à Mineração de Perspectiva, 3 associaram o mau desempenho de classificadores baseados em palavras e suas frequên-

[2] escrever sobre isso em algum lugar
[3] mesma coisa

cias à homogeneização do vocabulário contido no corpus: (**??**), (**??**) e (**??**). Não foi possível executar os experimentos descritos acima 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 corpora[4].

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 todos os seus *posts* 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% dos *posts* de Esquerda mencionam trechos de *posts* de Direita. Quanto aos *posts* de Direita, 77.5% deles mencionam *posts* de Esquerda. Essa forma de interação entre os *posts* é explorada em (**??**). Os autores criam um grafo de co-citação em que cada vértice representa um usuário e cada citação de um *post* a outro é indicada por uma aresta entre seus autores. A hipótese levantada é de que quão mais similares forem os padrões de citação de dois usuários, mais provavelmente eles defenderão uma mesma perspectiva política. Dada a matriz de adjacência desse grafo de co-citação, **M**, computa-se uma aproximação de baixo posto via Decomposição em Valores Singulares (SVD)[5], resultando na matriz **M'**. A aproximação da matriz por uma de posto menor via SVD é útil para extrair informações estruturais do grafo, como padrões de comunicação entre os vértices (**??**). Em seguida, calcula-se a distância entre os vértices em **M'** e agrupa-se os usuários de acordo com essa informação, através do algoritmo especificado em (**??**). Todos os *posts* referentes a cada grupo obtido são concatenados, gerando uma coleção menor de documentos. Um classificador Naïve Bayes, também baseado em palavras e suas frequências, é aplicado a essa nova coleção e os resultados obtidos são propagados para todos os usuários associados a cada grupo. Para usuários com mais de 500 palavras nos debates, a taxa de acerto dessa metodologia de classificação é de 73%.

Grafozinho do Taking sides

Padrões de citação entre documentos também foram investigados em (**??**). Neste artigo, os experimentos envolvem 2 *datasets*: o primeiro é composto de artigos políticos Estadunidenses de Direita ou Esquerda, coletados em *sites* e *blogs* políticos - explicitamente partidários ou não

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

[5]A técnica de aproximação para um posto menor via SVD tem que estar escrta em algum lugar!

-; o segundo é composto de textos retirados de *sites* sobre música, divididos entre as categorias *Alternative* e *Mainstream*. As aplicações de Naïve Bayes nestes corpora resultaram, respectivamente,**explicar melhor essa parte da coleta dos mainstream, q tá complexo.** Sejam **L** e **R** duas listas de URLs associadas a pontos de vista opostos sobre um tópico *T*. Para cada documento *d_i* do corpus, computa-se a probabilidade de *d_i* ser co-citado com algum elemento de **R** e, em seguida, com algum elemento de **L**. Se a razão entre estas probabilidades for maior que 1, o documento é classificado como portador da mesma perspectiva presente nas URLs de **R**. Caso esta razão seja menor que 1, a perspectiva associada ao documento será aquela relativa a **L**. Esta metodologia, aplicada ao primeiro *dataset*, resulta em uma taxa de acerto de 94.1%. No segundo *dataset*, a taxa obtida é de até 88.84%. Os resultados são muito bons, quando comparados com aqueles obtidos com Naïve Bayes e SVM, mas esta metodologia possui uma limitação importante: ela só é aplicável a *datasets* contendo duas perspectivas, diferentemente, por exemplo, de um Naïve Bayes, facilmente adaptável a mais pontos de vista[6]. Até mesmo o grafo de co-citação apresentado em (**??**) não apresenta, de antemão, uma restrição ao número de perspectivas presentes no *dataset*.

Concluir a seção.

[6]Indicar algo para ler sobre multiclass text classif. com naive bayes

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 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