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. Em (**??**), por exemplo, foi observado que várias palavras, como *palestinian* e *israel*, são utilizadas tanto em documentos pró-Palestina quanto pró-Israel. Apesar disso, as frequências distintas no uso dessas palavras evidenciam os diferentes lados da discussão. Esta 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.

Essa 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* (**??**).

Dado que em boa parte dos artigos estudados neste projeto, como (**??**) e (**??**), atinge-se taxas de acerto superiores a 80% com classificadores baseados em frequência de palavras, conclui-se que a mineração de perspectivas em discussões, artigos opinativos e debates requer metodologias diferentes, a depender de como as palavras foram escolhidas pelos autores dos documentos. Nos debates estudados por (**??**), expressões de ataque e defesa são mais frequentes do que *stigma words* e, como o método empregado no artigo foi um SVM treinado com frequências de palavras, observou-se que a classificação obtida para os lados do debate não refletia as perspectivas *liberal* ou *conservadora* - mas sim os lados *oposição* (expressões de ataque) e *situação* (expressões de defesa). Estes estudos indicam a possibilidade de que, em debates e discussões nos quais há uma homogeneização do vocabulário empregado - o que

pode acontecer quando todos os lados utilizam, em proporções similares, tanto expressões de ataque quanto de defesa -, classificadores baseados exclusivamente nas palavras utilizadas e/ou em suas frequências apresentarão má performance.

A avaliação do desempenho desses classificadores[1] nos *datasets* estudados revela que artigos opinativos e notícias consolidaram perspectivas, através da escolha do vocabulário utilizado, melhor do que debates. Apesar disso, uma generalização neste sentido, restringindo o uso desses classificadores a artigos e notícias, não é recomendada por falta de indicativos linguísticos que comprovem essa tendência. Uma estratégia que pode ajudar na escolha ou descarte de um classificador desse tipo é uma análise das palavras que estão contidas nos documentos. O uso de um modelo de tópicos do tipo L-LDA, no qual cada documento é marcado com um rótulo referente à sua perspectiva e outro genérico, idêntico para todos os documentos, facilita a identificação de palavras específicas para cada perspectiva, bem como aquelas que co-ocorrem em lados distintos do corpus, uniformizando o vocabulário.

Para a análise das palavras contidas nos documentos, utilizou-se a implementação de L-LDA disponível em (**??**) aplicada a dois *datasets*. O primeiro, composto de artigos extraídos do site bitterlemons.org, foi classificado com um Naïve Bayes padrão no artigo (**??**). As taxas de acerto obtidas, com o uso do Naïve Bayes, variaram entre 84.85% e 93.46% para os experimentos elencados nesse artigo. O segundo, **definir dataset, espero que o politics.com**, também foi classificado com um Naïve Bayes padrão em (**??**) - mas as taxas de acerto foram bem mais baixas: **X%**. Para o experimento com o L-LDA, todas as palavras contidas nos documentos, incluindo *stop words* como *the*, foram consideradas, resultando em uma análise das palavras independente do pré-processamento executado nos *datasets* nos dois artigos citados.

No primeiro *dataset*, os artigos estão escritos sob uma das seguintes perspectivas: pró-Palestina ou pró-Israel. Por conta disso, o tópico *pal* foi atribuído a todos os documentos pró-Palestina e o tópico *isr*, a todos os pró-Israel. Estes tópicos indexam a agregação das palavras mais fortemente associadas a cada uma das duas perspectivas, de acordo com o vocabulário empregado nos artigos. Um terceiro tópico, *gen*, foi atribuído a todos os documentos, a fim de capturar as palavras que co-ocorrem neles independentemente de suas perspectivas. Após a execução do modelo, as 100 palavras mais fortemente associadas a cada um dos 3 tópicos foram coletadas. 35% das palavras mais fortemente associadas a *isr* não foram coletadas para *gen*. Para *pal*, 29% das palavras recolhidas não fazem parte do conjunto coletado para *gen*. Por fim, 32% das palavras mais fortemente associadas a *isr* não fazem parte do conjunto recolhido para *pal* e vice-versa. Essas percentagens indicam que os autores dos documentos pró-Israel

[1] **SVMs e Naive Bayes padrão; LSPM**

utilizam um vocabulário ligeiramente mais específico, na defesa de seus pontos de vista, do que os autores pró-Palestina. É importante ressaltar que nenhuma palavra foi filtrada na análise - ou seja, termos muito frequentes como *the*, *of* e *and*, comumente extraídos dos *datasets* antes da etapa de classificação, estavam presentes nos documentos processados pelo L-LDA.

Palavras associadas a *isr* que não foram associadas a *gen* ['arafat', 'be', 'some', 'roadmap', 'us', 'yet', 'out', 'sharon', 'ariel', 'support', 'three', 'bush', 'palestine', 'new', 'terrorism', 'leader', 'then', 'jewish', 'after', 'arab', 'leadership', 'plan', 'president', 'than', 'bank', 'prime', 'regarding', 'like', 'could', 'violence', 'against', 'while', 'time', 'american', 'first']

Palavras associadas a *pal* que não foram associadas a *gen* ['then', 'some', 'authority', 'against', 'occupied', 'negotiations', 'occupation', 'sharon', 'united', 'end', 'way', 'palestine', 'international', 'be', 'after', 'plan', 'president', 'law', 'those', 'prime', 'land', 'i', 'violence', 'us', 'q', 'while', 'time', 'situation', 'first']

Palavras associadas a *pal* que não foram associadas a *isr* ['because', 'people', 'authority', 'states', 'right', 'occupied', 'any', 'negotiations', 'occupation', 'what', 'united', 'end', 'also', 'been', 'their', 'other', 'way', 'international', 'law', 'do', 'which', 'government', 'very', 'they', 'now', 'those', 'about', 'land', 'these', 'q', 'i', 'situation']

Palavras associadas a *isr* que não foram associadas a *pal* ['arafat', 'into', 'settlements', 'years', 'yet', 'out', 'even', 'would', 'ariel', 'west', 'support', 'three', 'bush', 'gaza', 'new', 'terrorism', 'leader', 'we', 'jewish', 'arab', 'most', 'leadership', 'minister', 'roadmap', 'than', 'bank', 'both', 'regarding', 'like', 'could', 'war', 'american'] **As palavras estão ordenadas de acordo com a força da associação com cada um dos tópicos**

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