

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

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

RESUMO

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

Palavras-chave: monografia, graduação, projeto final.

ABSTRACT

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

Keywords: monograph, graduation, final project.

LISTA DE FIGURAS

3.1	Grafo extraído de (??). Os vértices são usuários e as arestas, citações entre eles.	
	Arestas mais escuras indicam frequência de citação mais alta	16

LISTA DE ABREVIATURAS E SIGLAS

SUMÁRIO

1	Introdução			
	1.1	Motivação	7	
	1.2	Proposta	8	
	1.3	Estrutura da Monografia	8	
2	Rela	ção entre as características dos <i>datasets</i> e as metodologias utilizadas	9	
3	Pala	vras utilizadas nos documentos (provisório)	11	
	3.1	Introdução	11	
	3.2	Experimentos com modelos de tópicos	12	
	3.3	Aumentando as taxas de acerto: técnicas empregadas na literatura	14	
	3.4	Conclusão	17	
4	Con	clusão	18	
	4.1	Dificuldades encontradas	18	
	4.2	Trabalhos futuros	18	
АĮ	oêndi	ce A – Resultados experimentais	19	
R	eferên	icias Bibliográficas	20	

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

- 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

PRECISA PASSAR POR REFORMULAÇÃO. TALVEZ ENTRE NA PROPOSTA.

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 amadure-cer 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.

3 PALAVRAS UTILIZADAS NOS DOCUMENTOS (PROVISÓRIO)

3.1 INTRODUÇÃO

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, 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* (??).

Este capítulo discute a relação entre o vocabulário de um corpus e o desempenho de classificadores baseados na presença ou frequencia de suas palavras. Experimentos com um modelo de tópicos do tipo L-LDA e um classificador Naïve Bayes padrão, descritos na seção **Experimentos com L-LDA e Naïve Bayes**, ilustram esta relação, indicando também o quanto é difícil quantificá-la. Na seção **Aumentando as taxas de acerto: outras técnicas**, é apresentada uma revisão, e posterior discussão, de artigos que tratam a questão do vocabulário dos corpora,

propondo técnicas para classificá-los melhor quando as palavras são utilizadas de modo muito uniforme. A seção **Conclusão** resume as discussões e ideias apresentadas nas seções anteriores.

3.2 EXPERIMENTOS COM MODELOS DE TÓPICOS

Se um classificador utiliza apenas a presença e a frequência das palavras do corpus para identificar perspectivas, é natural que sua taxa de acerto seja tão mais baixa quanto menos essas características mudam de uma perspectiva para outra. Nesta seção, serão descritos dois experimentos que evidenciam o vocabulário escolhido em cada *dataset*, as taxas de acerto obtidas na classificação dos documentos com um Naïve Bayes e a relação entre eles.

O primeiro *dataset* é composto de artigos pró-Israel e pró-Palestina, extraídos do *site* **bit-terlemons.com** e estudados pela primeira vez em (??). Cada documento recebeu um rótulo referente à sua perspectiva e outro genérico, idêntico para todos eles. Um modelo de tópicos do tipo L-LDA foi aplicado aos documentos assim anotados, agrupando palavras genéricas em torno do rótulo genérico, pró-Israel em torno do rótulo pró-Palestina em torno do rótulo pró-Palestina. As trinta palavras mais fortemente associadas a cada um dos rótulos, em ordem, podem ser conferidas na tabela abaixo:

Rótulo	Palavras
Genérico	the, of, and, to, a, in, that, is, it, israel , this, for, not, be, as,
	palestinian, on, are, israeli, have, with, but, palestinians,
	by, was, an, from, will, their, or
Israel	the, to, and, s, in, sharon, palestinian, for, his, a, with,
	he, arafat, is, peace, by, on, us, israeli, will, prime, bush,
	minister, american, process, violence, terrorism, presi-
	dent, new, security
Palestina	the, palestinian, to, and, is, that, israeli, of, in, palestini-
	ans, on, s, has, sharon, for, peace, will, occupation, this,
	international, political, by, united, he, us, people, his, vi-
	olence, american, process

Tabela 3.1: Palavras extraídas com um L-LDA. Em negrito: verbos, adjetivos e substantivos.

O uso de um rótulo 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.

Os artigos deste primeiro dataset foram escritos ou pelos editores do site ou por convidados,

divisão utilizada em (??) para avaliar o desempenho de um Naïve Bayes. No primeiro cenário, os documentos de treinamento eram os escritos pelos editores e os de teste, aqueles escritos pelos convidados; no segundo, tinha-se a situação inversa. Repetindo-se o experimento com a implementação de Naïve Bayes disponível em (??), as taxas de acerto obtidas foram de X e 98.98% para cada um dos cenários, respectivamente. **pensar se compara com a de lin et al.**

O segundo *dataset* é composto de colocações em debates da *House of Representatives*, órgão Estadunidense responsável por aprovar novas leis. Ele é parte de um *dataset* maior estudado pela primeira vez em (??), e está disponível em (??). A rotulação original dos documentos foi mantida neste *dataset*: cada um deles é marcado com R, caso represente a colocação de um Republicano; ou D, no caso Democrata. Os documentos também são marcados com Y, caso representem a colocação de alguém que votou pela aprovação de uma lei; ou N, em caso contrário. Esta segunda marcação não foi explorada nesta monografia, nem a combinação das duas. O que se analisou, portanto, foi o desempenho do Naïve Bayes, disponível em (??), na classificação dos documentos como Republicanos ou Democratas.

O L-LDA foi aplicado a este *dataset* de forma análoga ao primeiro experimento, aproveitando os rótulos R e D presentes nos documentos e utilizando um rótulo genérico. As trinta palavras mais fortemente associadas a cada um dos rótulos, em ordem, podem ser conferidas na tabela abaixo:

Rótulo	Palavras
Genérico	the, to, of, and, that, a, in, is, i, this, we, for, it, have, not,
	are, be, on, mr., from, as, with, they, speaker, will, would,
	has, do, but, or
Democrata	the, to, and, of, in, our, for, this, bill, s, i, will, by, on, that,
	security, their, legislation, a, states, chairman, country,
	act, billion, not, million, law, would, war, administration
Republicano	the, to, of, and, in, s, for, our, this, act, chairman, on, by,
	security, states, bill, that, a, as, legislation, 11, their, sup-
	port, 9, i, system, united, will, are, terrorists

Tabela 3.2: Palavras extraídas com um L-LDA. Em negrito: verbos, adjetivos e substantivos.

Comparando as palavras extraídas para cada *dataset*, observa-se que o segundo apresenta, associadas a cada perspectiva, mais palavras genéricas do que o primeiro. As distribuições observadas para o segundo *dataset* evidenciam menos particularidades de cada perspectiva, contendo, proporcionalmente, mais artigos, conjunções e pronomes do que o primeiro *dataset*. Como consequência, a taxa de acerto do Naïve Bayes, treinado com 80% do corpus e testado nos 20% restantes, foi de apenas 48.92%. Não é trivial quantificar a relação entre essa taxa de acerto e a homogeneização do vocabulário do corpus - mas, como o Naïve Bayes utiliza apenas

a distribuição das palavras para inferir a perspectiva dos documentos, é evidente que a escolha do vocabulário contribuiu para os erros na classificação.

Para todos os experimentos, o número de iterações do L-LDA foi fixado em 100; o número de iterações do Naïve Bayes, em 500. O primeiro *dataset* é composto de 594 documentos: 297 pró-Israel e 297 pró-Palestina. Ele contém 465.422 palavras, 14.500 delas únicas. O segundo *dataset* é composto de 983 documentos: 487 escritos por Democratas e 496 escritos por Republicanos. Ele contém 359.761 palavras, 13.025 delas únicas. A implementação de L-LDA utilizada nestes experimentos está disponível em (??); a de Naïve Bayes, em (??).

Os experimentos evidenciam a relação entre a linguagem presente nos corpora e a dificuldade de se classificar corretamente cada documento com o uso de um Naïve Bayes. É válido ressaltar que, a depender do *dataset*, outras questões podem ser 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 para se lidar com isso.

3.3 AUMENTANDO AS TAXAS DE ACERTO: TÉCNICAS EMPREGADAS NA LITERATURA

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 corpora¹.

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.

¹Técnicas de classificação desenvolvidas por artigos que não tratam da questão da uniformidade das palavras serão apresentadas em outras seções desta monografia.

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:

- Dada a matriz de adjacência do grafo de co-citação, M, computa-se uma aproximação com posto menor, M', via SVD²;
- 2. calcula-se a distância entre os vértices de 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:
- 5. aplica-se um Naïve Bayes a essa nova coleção de documentos;
- os resultados obtidos com o Naïve Bayes são propagados para todos os participantes de cada grupo.

O uso da matriz **M'**, em vez de **M**, justifica-se por **M'** possuir menos ruído e evidenciar informações estruturais do grafo, como padrões de comunicação entre os vértices, melhor do que **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%.

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

 $^{^2}$ A técnica de aproximação para um posto menor via SVD tem que estar escrta em algum lugar! senão n escrevo aí.

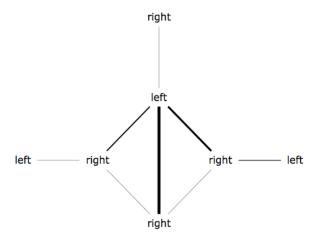


Figura 3.1: Grafo extraído de (??). Os vértices são usuários e as arestas, citações entre eles. Arestas mais escuras indicam frequência de citação mais alta.

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:

- 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%.

Todos os artigos que propuseram metodologias para melhorar a classificação em corpora com vocabulário muito uniforme consideraram algum esquema de citação. Se o corpus for composto de documentos pouco citados na Web, o esquema de (??) pode não trazer nenhuma

melhoria significativa ao problema. Se os autores dos documentos não mencionarem significativamente um ao outro, como aconteceu nos debates estudados por (??) e (??), a metodologia explorada por estes trabalhos também não pode ser empregada. Essa área de pesquisa, portanto, ainda tem alguns problemas em aberto - e, considerando que estes três artigos foram escritos há mais de quatro anos, ela não parece atrair muita atenção. A justificativa para isso, provavelmente, tem a ver com o fato de que uniformidade no vocabulário não é um problema tão comum em Mineração de Perspectiva. No **capítulo X**, técnicas envolvendo comparação entre documentos, que podem trazer benefícios à classificação de *datasets* difíceis, como aqueles em que o vocabulário é mais uniforme, serão discutidas.

3.4 CONCLUSÃO

Este capítulo apresentou a ideia de que pessoas costumam utilizar palavras específicas para defender perspectivas diferentes. O uso de classificadores baseados na presença ou frequência das palavras, portanto, costuma ser indicado para a identificação dessas perspectivas. Entretanto, em alguns casos, essa ideia não se confirma: as perspectivas defendidas se apóiam mais nas construções do discurso do que no uso de palavras específicas. Naturalmente, as taxas de acerto desses classificadores, nestes casos, são menores do que o desejado. Experimentos com o modelo de tópicos L-LDA foram conduzidos, conduzindo à visualização parcial de como palavras, em dois *datasets* diferentes, se relacionam com diferentes perspectivas. No primeiro *dataset*, as palavras listadas para cada perspectiva se relacionam, semanticamente, melhor com a perspectiva em si do que no segundo *dataset*. Neste último, palavras genéricas, como artigos e conjunções, se associam muito fortemente a todas as perspectivas, colaborando para uma uniformização do vocabulário. Isto explica, ainda que não seja trivial avaliar o quanto, as taxas de acerto mais altas obtidas com um Naïve Bayes no primeiro *dataset*.

Por fim, este capítulo também revisa artigos que discutem a questão do vocabulário do corpus. Estes artigos propõem metodologias para melhorar a classificação em *datasets* cujo vocabulário é considerado muito uniforme. Pouca pesquisa tem sido feita nesta direção - provavelmente porque o problema é pouco comum.

4 CONCLUSÃO

O trabalho onon ono non ono no

4.1 DIFICULDADES ENCONTRADAS

O trabalho onon ono non ono no

4.2 TRABALHOS FUTUROS

Pode-se indicar como trabalhos futuros:

n ono non ono non

ono non ono non ono non non ono non ono non ono non ono non controlador n ono non ono

ono non ono o non ono non on

APÊNDICE A – RESULTADOS EXPERIMENTAIS

No no nnononono no n ono o nn.

REFERÊNCIAS BIBLIOGRÁFICAS