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 CONCEITOS

Este trabalho relata o onon ono non ono non ono non ono non ono non on.

2.1 LISTAGENS

Na listagem 2.1 é mostrado o teste do método `Engine.initialize()`:

Listagem 2.1: Classe Factory2D

```
public class EngineTest
// JUnitDoclet begin extends_implements
extends TestCase
// JUnitDoclet end extends_implements
{
  //...
  public void testInitialize() throws Exception {
   // JUnitDoclet begin method initialize
   EngineState engineState = (EngineState) PrivateAccessor.
    getField(engine,"engineState");
   engine.initialize();
   assertEquals(engineState, new InitState());
   // JUnitDoclet end method initialize
  }
  ...
}
```

Figura 2.1: Brasão da UFBA (vetorial)

Figura 2.2: Brasão da UFBA (*bitmap*)

2.2 FIGURAS

É possível usar imagens vetoriais no formato PDF, como pode ser visto na figura 2.1, ou imagens *bitmap* no formato PNG, como a da figura 2.2.

3 *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.

3.1 O NÍVEL DE FORMALIDADE NA LINGUAGEM DOS DOCUMENTOS

MODO RASCUNHO AINDA

Explicar que uma diferença fundamental notada entre os datasets observados diz respeito à formalidade da linguagem empregada nos documentos. Enquanto alguns datasets utilizam documentos provenientes de meios onde a norma culta da linguagem impera, e uma revisão ortográfica é utilizada, outros são carregados de gírias, expressões e abreviações típicas da linguagem da Internet e contêm, eventualmente, grafias erradas para uma mesma palavra. <Dar exemplos textuais. Mostrar uma tabelinha, algo assim, indicando o volume de datasets com linguagem informal nos documentos estudados.>

A linguagem informal pode criar alguns desafios para a Mineração de Perspectiva [FECHAR O PROBLEMA EM: IDENTIFICAÇÃO DA PERSPECTIVA PRESENTE EM UM DOCUMENTO], especificamente no que diz respeito ao pré-processamento dos documentos e à escolha do método empregado. No pré-processamento, como indicado em (aaai-politics.pdf e 10.1.1.138.7160.pdf), a correção gramatical das palavras é bastante indicada. Com uma única versão de grafia para cada palavra (a correta), diminui-se a quantidade de ruído que grafias erradas podem causar na classificação.

Uma característica dos datasets que deveria ser considerada sempre antes de se escolher o método utilizado - mas não é prática entre as pessoas que estudam perspective mining - é a frequência das palavras nos documentos do dataset. Se o léxico empregado pelos autores dos textos muda sensivelmente a depender de sua ideologia/perspectiva defendida/ponto de vista, é possível resolver o problema utilizando classificadores que usam essa frequencia como feature. GASTE TEMPO DANDO ALGUNS EXEMPLOS. Em datasets informais, entretanto, como aponta Efron, a linguagem empregada por todos os lados da discussão pode ser basicamente a mesma: termos com forte carga de polaridade e gírias/jargões comuns na discussão [EXEMPLO]. Neste caso, é preciso utilizar métodos que utilizem mecanismos mais rebuscados do que a simples frequência/presença das palavras no texto. Alguns autores utilizam métodos mais gramaticais para lidar com datasets informais, como é o caso de , BLE e BLI. Os 2 primeiros criam o conceito de Opinion Frame (DEFINIR O CONCEITO). No primeiro caso, a ideia é associar corretamente opiniões a alvos, e associar alvos iguais utilizando técnicas de co-referência. Segundo Wiebe et al. (pegar a citação corretamente), ter as targets associadas, com as opiniões

próximas, é uma boa forma de entender o overall de opiniões de um documento. A estratégia é uma alternativa possível para documentos em que as perspectivas estão muito associadas à linguagem opinativa, e onde essa linguagem é comum para todos os lados. Infelizmente, os resultados alcançados indicam que a tarefa não é trivial: blablablablababla (resolver coreferencia pra linkar targets não é tão simples assim => e dê exemplo).

4 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

4.1 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

4.2 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